UNIVERSIDAD DE MÁLAGA

Subsidia: Tools and Resources for Speech Sciences

José María Lahoz-Bengoechea
and Rubén Pérez Ramón (Eds.)

SUBSIDIA:
Tools and resources for speech sciences

José María Lahoz-Bengoechea
and Rubén Pérez Ramón (Eds.)

Universidad de Málaga

ISBN: 978-84-09-12881-5

Subsidia: Tools and Resources for Speech Sciences
Subsidia: Herramientas y Recursos para las Ciencias del Habla

Libro publicado por la Universidad de Málaga (UMA)

Subsidia: Tools and Resources for Speech Sciences is a scientific publication arising from the organisation of a congress with the same name, carried out in the city of Málaga (Spain) in June, 2017. Its main goal is to give voice to tools and resources developed with the aim of facilitating research in the field of speech sciences. This framework embraces subjects such as phonetics, experimental phonetics, phonology, discourse analysis or dialectology, among others. This book, outcome of the collaboration of expert researchers on their respective areas, aims to be an aid to the scientific community in the sense that it compiles and depicts a series of materials that, we hope, may result beneficial to keep moving forward in the research.

The papers collected in this volume are a selection of those submitted to the above-mentioned congress, and have undergone peer review.

Subsidia: Herramientas y Recursos para las Ciencias del Habla es una publicación científica resultante de la organización del congreso del mismo nombre desarrollado en la ciudad de Málaga (España) en junio del año 2017. Su objetivo principal es dar a conocer herramientas y recursos desarrollados con el objetivo de facilitar la investigación en el campo de las ciencias del habla. Dentro de este marco se engloban disciplinas tan variadas como la fonética experimental, la fonología, el análisis del discurso o la dialectología, entre otras. Este libro, resultado de la colaboración de investigadores expertos en sus respectivas áreas, pretende ser una ayuda a la comunidad científica en tanto en cuanto recopila y describe una serie de materiales que esperamos resulte provechoso para continuar avanzando en la investigación.

Los artículos recogidos en este volumen son una selección de los que se presentaron en dicho congreso y han pasado una evaluación por pares.

TECHNICAL EDITION / EDICIÓN TÉCNICA

Cover design / Diseño de la cubierta: Rubén Pérez Ramón & José María Lahoz-Bengoechea.

Subsidia: Tools and Resources for Speech Sciences

CREDITS

ARTICLES

Subsidia: Tools and Resources for Speech Sciences

1

Bringing together tools and resources for speech sciences

José María Lahoz-Bengoechea[1] and Rubén Pérez Ramón[2]

[1] Universidad Complutense de Madrid
[2] Universidad del País Vasco
e-mail: jmlahoz@ucm.es, rperez.ram@gmail.com

Citation / Cómo citar este artículo: Lahoz-Bengoechea, J. M. & Pérez Ramón, R. (2019). Bringing together tools and resources for speech sciences. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 1–3). Málaga: Universidad de Málaga.

ABSTRACT: Researchers in speech sciences develop materials and instruments to conduct their studies, to automatize analyses and tasks, and to provide utilities applicable to different fields. This paper and the book that it introduces present a panorama of the current interests in said fields, and seek to provide an interdisciplinary perspective in order to bootstrap the creation of more and better tools and resources relevant to those areas.

Keywords: speech sciences; speech technologies; interdisciplinary studies.

RESUMEN: Los investigadores de las diversas ciencias del habla desarrollan materiales e instrumentos para llevar a cabo sus estudios, para automatizar análisis y tareas, y para proporcionar aplicaciones de utilidad para sus distintos campos. Este capítulo y el libro al que introduce presentan un panorama de los intereses actuales en dichas disciplinas, y proporcionan una perspectiva interdisciplinar con la intención de incentivar la creación de más y mejores herramientas y recursos, que sean relevantes para esas áreas.

Palabras clave: ciencias del habla; tecnologías del habla; estudios interdisciplinares.

1. INTRODUCTION

One of the senses of the Latin word *subsidia* is "assistances, supports, aids" and, as a matter of fact, this book aims at raising awareness of some of the tools and resources (available or under development) that may assist professionals and scholars in speech-related areas in successfully conducting their research.

These resources are often intended for their use in a specific area of expertise but are equally suitable and practical in other related fields in which they might not be well-known. For this reason, the present book adopts an interdisciplinary approach and it embraces contributions on multiple aspects concerned with the study of speech from many of the different perspectives from which it is addressed.

Bringing together these tools and making them known to the scientific community is of paramount importance for various reasons. First of all, it keeps scholars from repeating work that is already done, so they can build on what exists in order to improve it, or simply devote themselves to filling other gaps. Secondly, making one's own materials accessible to others usually results in valuable feedback for further development. Moreover, creativity and new ideas are likely to be encouraged when exchanging viewpoints with researchers from related yet different fields (e.g. medicine, police, engineering, business, education, psychology, etc.).

It is unquestionable that tools are not an end, but a means. They are ancillary—or subsidiary, as the title of the book suggests—to the work of those professionals or other potential users and it depends on how they utilize such resources to answer their possible research questions and to achieve the most profitable results. That said, the current proliferation of instruments provides an excellent ground to develop studies and applications that used to be more difficult to pursue.

The following chapters deal with an array of tools and resources that represent a wide-scope sample of the current interests within speech sciences. Those tools can be understood as methods, websites, software, scripts, smartphone apps, etc. that perform different tasks or allow accurate calculations of some parameter. Resources encompass databases, corpora, or multiple kinds of interfaces that give access to data and information.

In chapter 2, Alku and his colleagues introduce an open-source software to obtain the glottal flow (or flow derivative) from an acoustic pressure waveform by means of a choice of two algorithms that estimate the inverse filter. The software is multiplatform (it runs on Windows and Mac OS) and includes a graphical interface that allows the user to select the best-looking resulting flow waveform. Glottal inverse filtering (GIF) has become a standard practice over the last

decades as an easy, non-invasive method to retrieve the glottal source. This field enjoys active research to look for new algorithms that introduce fewer artifacts in the signal. Furthermore, GIF facilitates modeling by means of a limited number of parameters, and this is promising for anyone interested in the study of voice, such as speech therapists, vocal coaches, forensic phoneticians, and developers of speech technologies for voice synthesis or for the recognition of emotions, to name just a few possibilities.

Chapter 3 also relates to the study of voice quality and its multiple applications (whether linguistic, paralinguistic, or extralinguistic), as mentioned above. The authors discuss the statistical link between perceptual judgments on voice quality, as provided by listeners trained in the Vocal Profile Analysis Scheme (VPAS), and a series of acoustic measurements, including f_0, intensity, harmonic-to-noise ratio, F1, and spectral slope (as captured by H1−A3). This study adds to the search of robust acoustic methods that may allow to assess voice quality without the need for trained listeners.

Madureira and Souza Fontes (chapter 4) extend the analysis of expressivity to facial gestures, and propose a method for jointly evaluating facial units, voice quality, and other acoustic parameters. This chapter presents some tools (protocols and software) to carry out the analysis of paralinguistic and nonverbal indices to emotions, whether voluntary or not, and implies attractive applications in fields such as marketing, safety, lie detection, pain detection, or education.

Chapter 5, by Garrido and colleagues, deals with a very transversal tool, which is in fact one of the keystones of many applications within phonetics and speech techologies, namely the transcriptor. This chapter discusses the typical steps and problems of the transcription process, and offers a free software tool that transcribes Spanish and Catalan, not only in their standard form but also in a good number of peninsular varieties. The user can switch the transcription output between IPA and SAMPA.

Chapter 6 (Cutugno et al.) presents an open source application that will prove invaluable for fieldwork, since it runs on Android. It is packed with a recorder, as well as an array of functions of acoustic analysis, including spectrogram visualization, and utilities to add different metadata.

Turning to chapter 7, we find a wordnet database for Italian, which includes information regarding different levels of linguistic analysis (phonological, morphosyntactic, and lexical) and allows declarative queries combining constraints on all those levels. This resource, by Origlia and colleagues, has interesting applications both for basic research on linguistics and for speech technologies.

Cebrian's chapter (number 8) reviews some of the different tasks that can be used to ascertain phonetic similarity between categories of two languages, and discusses the methodological issues that should be considered in each case. Studies on phonetic similarity contribute to our understanding of perception and learnability of L2 sounds and therefore may have a considerable role in developing new methods for language teaching, or in refining the existing ones.

Marrero and Fresno present in chapter 9 the prototype of an information-retrieving system that will draw from different academic repositories to obtain multimedia resources. The videos included so far are accompanied by high-quality subtitles, as well as phonetic and morphologic labels. A search engine enables queries by phonetic categories, by words (either in lemma or stem form), or by text, including the possibility to use regular expressions. Suggested applications focus on the teaching of Spanish as a second language, but also any kind of study targeting academic spoken Spanish.

The platform *Teach Yourself English Pronunciation* (chapter 10) is structured in eight sections dealing with segmental and prosodic issues of English phonetics. Having L1-Spanish learners of English in mind, each section covers the most typical pronunciation mistakes and includes tips, audios and exercises that provide immediate feedback, yet another advantage made possible by new technologies. Additionally, Estebas Vilaplana, drawing from both the British and the American traditions of intonational analysis, introduces TL_ToBI, a hybrid model to annotate intonation in a simpler way, more suitable for distance education and self-learning contexts.

In chapter 11, Carrera-Sabaté and her colleagues present an online platform that describes the sounds and the intonational patterns of Catalan. It includes a fair amount of phonetic terms, so it is best addressed to users familiarized with that metalanguage. The tool contains exercises and games, which makes it a good resource for learners of Catalan as a second language.

Schwab and Dellwo (chapter 12) compare two computer-assisted methods to teach Spanish lexical stress to native speakers of either French or German who have no prior knowledge of Spanish. The explicit method—including an instructional video as well as discrimination, identification, and repetition drills—proved more helpful for speakers of French, a language without contrasts based on stress position. Conversely, Germans benefited more from the implicit method, consisting on arbitrary relations between colored geometric shapes and each word from two minimal triplets, differing only in lexical stress location.

The proliferation of webs intended as tools to teach the correct pronunciation of a language to non-native speakers is also patent in chapter 13. Pons-Moll and Carrera-Sabaté offer an invaluable tool to support both teachers and students of Catalan, including several geographic varieties. The site provides different types of materials, such as IPA transcriptions, midsagittal views of the vocal tract, spectrograms, palatograms, MRI, and videos of the speakers' faces.

Moving from language teaching to forensic linguistics, Ramos et al. compare the Maximum Likelihood Ratio, a method widely used in voice comparisons,

Lahoz-Bengoechea, J. M. & Pérez Ramón, R. (2019). Bringing together tools and resources for speech sciences. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 1–3). Málaga: Universidad de Málaga.

with Bayesian calculations. In this chapter (number 14), the authors show that a fully Bayesian approach is preferrable, especially when there are few data to train the model.

The presentation of *EMULANDO*, a corpus that includes both genuine and feigned foreign-accented spoken Spanish, is the focus of chapter 15 (Lahoz-Bengoechea et al.). The corpus is an extensive resource to study a major kind of disguised speech often occurring in forensic contexts, namely that of subjects who pretend to be non-native speakers of a language in order to conceal their identity. The recordings comprise English-, French-, and Russian-accented Spanish, and may be a valuable material to compare the phonetics of those languages, as well as to study the degree of accentedness according to the proficiency of the speaker in the target language or accent.

In chapter 16, Gómez et al. describe an inverse filter-based method that takes the temporal evolution of the first two formants to compute the articulatory velocity of the joint structure formed by the jaw and the tongue. Basing on a case study, it is proposed that said velocity is a reliable parameter to evaluate disorders in general neuromotor activity that typically arise in patients with Parkinson's, Alzheimer's or Amyotrophic Lateral Sclerosis.

Pérez Ramón presents a package of tools developed for Praat, a software widely known in the phonetics community (chapter 17). This bundle of scripts allows to implement perception tests with a more refined design than is possible with the usual interface, e.g. enabling the participant to simultaneously choose several options, or to input an answer in text form.

Chapter 18 (Llisterri et al.) deals with VILE-P, a corpus designed to study within- and between-speakers variation of different prosodic aspects of Spanish. It is thoroughly annotated for 16 different levels, covering segmental, suprasegmental, and lexical informations. Its main application lies within the field of forensic studies (speaker identification), but it may also serve to train different kinds of automatic recognition systems, such as labelers and aligners.

ProDis (chapter 19) is a dialectometric tool developed by Fernández Planas and her colleagues that quantifies the prosodic distance between different geographic varieties of languages basing on their intonation. It has been tested with data from the fixed corpus of the AMPER project. Furthermore, the results can be output in several formats: as numeric data in a table, as a color-shaded matrix that maps either the similarities or the standard deviations of the samples, as a dendrogram, as a 2D or a 3D map, etc.

Finally, Elvira-García and Garrido (chapter 20) present FonetiToBI, a tool intended to analyze tonal events and annotate the intonational contours in large corpora, following the ToBI conventions for either Spanish or Catalan. It is packed with an ample tonal inventory and the evaluations carried out show a high level of accuracy, especially for nuclear pitch accents and for boundary tones, with a somewhat lower hit rate for prenuclear accents. The authors provide some thoughts on potential sources of error, and compare the machine's performance with that of human annotators, which results in a close coincidence.

As can be seen from the previous mosaic, the investigators' interests within the area of speech sciences are many and varied, ranging from basic to applied research. Much effort is being put into automatizing analyses and tasks, as well as into developing technologies and methods that recognize speech, the speaker's identity or even emotions. Another main direction is to build corpora of data annotated for different linguistic levels, as a resource for further, systematic studies. There is also a major trend for many of these studies to focus on prosodic aspects such as stress or intonation. Similarly, there is a growing interest in parameters that can properly describe voice quality.

All in all, these tools and resources have important applications in many fields such as clinical linguistics, forensic linguistics, language acquisition and teaching, psychology, business, engineering, safety, police, etc. Hopefully this book will help to make known some aspects of what is being done in each of these areas and will foster the dialog among them in such a way that they can enrich one another with new ideas and potentialities.

Lahoz-Bengoechea, J. M. & Pérez Ramón, R. (2019). Bringing together tools and resources for speech sciences. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 1–3). Málaga: Universidad de Málaga.

Subsidia: Tools and Resources for Speech Sciences

Aalto Aparat: A freely available tool for glottal inverse filtering and voice source parameterization

Paavo Alku[1], Hilla Pohjalainen[1] and Manu Airaksinen[1]

[1] Aalto University
e-mail: paavo.alku@aalto.fi

Citation / Cómo citar este artículo: Alku, P., Pohjalainen, H., & Airaksinen, M. (2019). Aalto Aparat: A freely available tool for glottal inverse filtering and voice source parameterization. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 5–10). Málaga: Universidad de Málaga.

ABSTRACT: A software tool, Aalto Aparat, is introduced for glottal inverse filtering analysis of human voice production. The tool enables using two inverse filtering methods (Iterative adaptive inverse filtering, Quasi closed phase analysis) to estimate the glottal flow from speech. The inverse filtering analysis can be conducted using a graphical interface either automatically or in a semiautomatic manner by allowing the user to select the best glottal flow estimate from a group of candidates. The resulting glottal flow is parameterized with a multitude of know parameterization methods. Aalto Aparat is easy to use and it calls for no programming skills by the user. This new software tool can be downloaded as a stand-alone package free of charge to be run on two operating systems (Windows and Mac OS).

Keywords: glottal inverse filtering; voice source; speech research tool.

RESUMEN: El artículo presenta Aalto Aparat, una herramienta informática que permite analizar la producción de la voz humana mediante el filtrado inverso de la onda glotal. El programa dispone de dos métodos de filtrado para estimar el flujo glotal a partir de la señal de habla, a saber, el filtrado inverso adaptativo iterativo, y el análisis de la fase cuasi cerrada. El análisis del filtrado inverso se puede realizar utilizando una interfaz gráfica tanto de forma automática como semiautomática, pues se permite al usuario seleccionar la mejor estimación del flujo glotal a partir de un grupo de candidatos. El flujo glotal resultante se parametriza siguiendo diversos métodos de parametrización conocidos. Aalto Aparat es fácil de manejar y no requiere conocimientos de programación por parte del usuario. La herramienta se puede descargar de forma gratuita como programa ya compilado para ejecutarse en dos sistemas operativos (Windows y Mac OS).

Palabras clave: filtrado inverso de la onda glotal; fuente de la voz; herramienta de investigación del habla.

1. INTRODUCTION

Voiced speech is excited by a quasiperiodic airflow pulse form which is generated at the vocal folds. This excitation waveform, referred to as the glottal volume velocity waveform (shortly glottal flow), is the source of some of the most important acoustical cues embedded in speech. The fluctuation speed of the vocal folds determines the cycle length of the glottal flow which in turn affects the sensation of pitch from speech signals. The human speech production mechanism is capable of varying not only the fluctuation *speed* of the vocal folds but also their fluctuation *mode* thereby generating glottal flow pulses whose shape varies from smooth (i.e. large spectral tilt) to more abruptly changing (i.e. smaller spectral tilt). The shape of the glottal pulse is known to signal acoustical cues which are used, for example, in vocal communication of emotions (Gobl & Ní Chasaide, 2003).

Direct non-invasive recording of the glottal flow is, unfortunately, not possible due to the position of the vocal folds in the larynx behind cartilages. Non-invasive analysis of the glottal flow is, however, enabled by using an alternative to direct acoustical measurements, the technique known as *glottal inverse filtering* (GIF) (Alku, 2011; Drugman, Alku, Alwan, & Yegnanarayana, 2014). This corresponds to using the idea of mathematical inversion: by recording the output of the speech production system, the pressure signal captured by microphone, a computational model is first built for those processes (i.e. vocal tract, lip radiation) that filter the glottal excitation. By feeding the recorded speech signal through the inverse models of the filtering processes, an estimate for the glottal flow is obtained. Analysis of speech production with GIF consists typically of two phases: (1) the estimation phase in which glottal flow signals are estimated from

speech utterances with a selected GIF method, and (2) the parameterization phase in which the obtained waveforms are expressed in a compressed form with selected glottal parameters.

Given the fact that digital GIF methods have been developed since the 1970's, there are plenty of known algorithms available today both for glottal flow estimation and parameterization. (For further details of GIF history, see recent reviews by Alku, 2011, and Drugman et al., 2014). It is delighting to observe that there is currently a growing interest among developers of GIF algorithms in open source practices and open repositories (Degottex, Kane, Drugman, Raitio, & Scherer, 2014; Drugman, n.d.; Kane, 2012, 2013). Inverse filtering and parameterization methods developed so far are, however, almost exclusively published in a manner which unfortunately hinders the utilization of these techniques by researchers who do not have programming skills. Therefore, the corresponding speech research methods can be fruitfully utilized only by those researchers who have engineering or computer science background while these open source tools (which are mostly made available today as MATLAB scripts) remain of limited practical value for individuals with non-technical background. While providing openly available MATLAB implementations in GIF helps, for example, in evaluating different GIF methods by the algorithm developers, we argue that it would be desirable to have GIF analysis available also for a wider speech research community. In other words, estimation and parameterization of the glottal flow should be made as easy as the Praat system (Boersma & Weenink, 2013) to researchers such as linguists, phoneticians, and physicians who typically do not have skills in programming languages such as MATLAB.

To the best of our knowledge, there are currently only two freely available GIF tools that do not call for any programming by the user to run the analysis. DeCap (Granqvist, Hertegård, Larsson, & Sundberg, 2003; Tolvan Data, n.d.) is a tool that enables voice source analysis in which the user adjusts each antiresonance of the vocal tract using the computer mouse by simultaneously monitoring the waveform of the GIF output on the computer screen. DeCap users typically define the optimal antiresonance setting as the one that results in the glottal flow pulse with the longest horizontal closed phase thereby utilizing a prevalent subjective inverse filtering criterion (Gauffin-Lindqvist, 1965; Lehto, Airas, Björkner, Sundberg, & Alku, 2007; Rothenberg, 1973). DeCap enables parameterizing the obtained glottal flow with, for example, H1−H2 (Titze & Sundberg, 1992) and NAQ (Alku, Bäckström, & Vilkman, 2002). TKK Aparat (Airas, 2008) is another user-friendly tool for glottal flow estimation and parameterization. (TKK stands for Teknillinen korkeakoulu, the former name of Aalto University.) Differently from DeCap, the user of TKK Aparat is given an option to select the best glottal flow signal from a set of candidates that have been computed from the input speech by varying two inverse filtering parameters (order of the vocal tract model, coefficient of the lip radiation). After the user has selected the best glottal flow candidate, the selected waveform can be parameterized in TKK Aparat by a rich set of parameterization methods. It is also worth noting that in addition to DeCap and TKK Aparat there are tools, such as VoiceSauce (Shue, Keating, Vicenik, & Yu, 2011; VoiceSauce, 2016), which have been developed for the parameterization of voice production based on quantifying the speech pressure signal or its spectrum with measures such as H1*−H2* (Kreiman et al., 2012). These tools, however, do not estimate the glottal flow as a time-domain signal and therefore they cannot be regarded as (true) GIF tools.

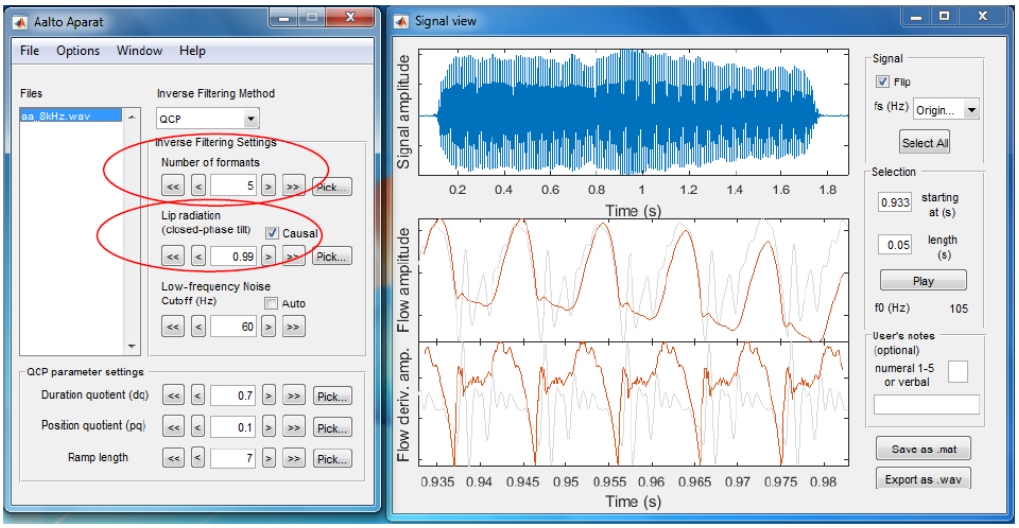
The current study introduces a new, updated version of TKK Aparat, named Aalto Aparat. Similarly to its predecessor described by Airas (2008), Aalto Aparat is a speech inverse filtering and parameterization software that enables analyzing the voice source using a user-friendly graphical interface. The interface enables the user to conduct GIF analysis and parameterization with no need to use a specific programming language or environment. The tool has been originally programmed in MATLAB but, importantly, it can be downloaded freely as a stand-alone package which can be used without access to MATLAB. Compared to its predecessor published by Airas (2008), Aalto Aparat includes three major improvements. First, the tool now supports a new GIF algorithm, Quasi closed phase analysis (QCP), which has been shown to be one of the most accurate, if not the most accurate, GIF method (Airaksinen, Raitio, Story, & Alku, 2014). Second, the user interface of Aalto Aparat has been improved, for example, by allowing the user to save the estimated flow waveforms as digital signals, not just their parameters. Third, the tool is now available (Aalto Aparat, 2016) as a stand-alone package that can be run in two operating systems (Microsoft's Windows, and Apple's Mac OS).

2. FEATURES OF AALTO APARAT IN A NUTSHELL

Aalto Aparat is a MATLAB-based tool designed for glottal inverse filtering studies of speech production. It supports the two phases (estimation and parameterization) that are typically needed in inverse filtering research. Given its user-friendly interface, the tool is well-suited particularly for studies in which large amounts speech signals need to be inverse filtered and parameterized. Inverse filtering in Aalto Aparat has been implemented in such a form that the user can fine-tune certain GIF settings thereby affecting the estimated glottal flow estimate if desired. The user is given a possibility to select the best glottal flow estimate from a group of candidates, hence enabling running GIF analysis that is not completely automatic (and therefore maybe more prone to errors) but allows feedback from the user.

Alku, P., Pohjalainen, H., & Airaksinen, M. (2019). Aalto Aparat: A freely available tool for glottal inverse filtering and voice source parameterization. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 5–10). Málaga: Universidad de Málaga.

Figure 1: Two windows of Aalto Aparat: control window (left) and signal view window (right). In control window, red circles show two settings (vocal tract filter order, lip radiation coefficient) that the user can vary if desired. In signal view window, the three panes show the input speech signal (top), the estimated glottal flow (middle), and the derivative of the estimated flow (bottom).



The input to Aalto Aparat is a speech pressure signal in the wav format. In the estimation phase, Aalto Aparat enables using two glottal inverse filtering algorithms, Iterative adaptive inverse filtering, IAIF (Alku, 1992) or Quasi closed phase analysis, QCP (Airaksinen et al., 2014), to estimate the glottal flow from the input speech. In IAIF, the user can select either conventional linear prediction, LP (Makhoul, 1975), discrete all-pole modeling, DAP (El-Jaroudi & Makhoul, 1991) or minimum variance distortionless response, MVDR (Wölfel & McDonough, 2005) as a vocal tract all-pole modelling method. In QCP, the user can fine-tune the parameters of the attenuated main excitation, AME (Alku, Pohjalainen, Vainio, Laukkanen, & Story, 2013; Airaksinen et al., 2014) weighting window. Once the user has selected the best estimate (see section 3.2), the obtained glottal flow is parameterized with several parameters both in the time domain, using for example ClQ (Timcke, von Leden, & Moore, 1958) and NAQ (Alku, Bäckström, & Vilkman, 2002), and in the frequency domain, using for example H1−H2 (Titze & Sundberg, 1992) and PSP (Alku, Strik, & Vilkman, 1997). In addition, it is possible to fit the Liljencrants-Fant (LF) waveform (Fant, Liljencrants & Lin, 1985) into the obtained glottal flow derivative. The parameterization procedures are equal to those in (Airas, 2008) where more details can be found.

3. DEMONSTRATION OF AALTO APARAT

The best way to describe Aalto Aparat is to study an example demonstrating the major parts that are needed in order to inverse filter and parameterize an input speech signal by this new tool. Given the space restriction in the current article, interested readers are referred to the manual of Aalto Aparat (Aalto Aparat, 2016) to get a more in-depth view on the system.

3.1. Step 1: Importing speech

When the Aalto Aparat tool is opened, the system displays two windows (Figure 1): control window (left) and signal view window (right). The former lists all the pre-recorded wav files (i.e. speech pressure signals) that the user wants to analyze. As a pre-processing step, the system enables removing ambient noise from the recorded signals with a liner phase high-pass filter whose cut-off frequency can be set automatically (according to the fundamental frequency of the input speech) or manually. In addition, the speech signal's sampling frequency can be changed and its polarity can be swapped if desired.

3.2. Step 2: GIF analysis

After the speech signal has been imported to the system, an analysis frame in which the GIF analysis is to be computed is set to a default duration (50 ms) and position (in the middle of the input signal). If desired, the user can, however, adjust both of these values. Next, the user selects the GIF method (either IAIF or QCP) after which the system automatically depicts the obtained glottal flow (Figure 1, right window, second pane from top) and its derivative (Figure 1, right window, bottom pane) on the computer screen. By pressing the corresponding buttons (Figure 1, left window, two red circles) the user can vary the value of two parameters of the selected GIF algorithm: the vocal tract filter order (Figure 1, upper red circle) or the lip radiation coefficient (Figure 1, lower red circle). After this, the system opens a new window which depicts a group of candidate glottal flow estimates that have been computed by varying the corresponding parameter (Figure 2 shows an example where the vocal tract order is varied). Once the user has screened the depicted waveforms, he/she can select the one that he/she considers best by clicking the waveform with

Figure 2: A group of candidate flow signals which have been obtained by varying the vocal tract filter order from 4 (top signal) to 16 (bottom signal).

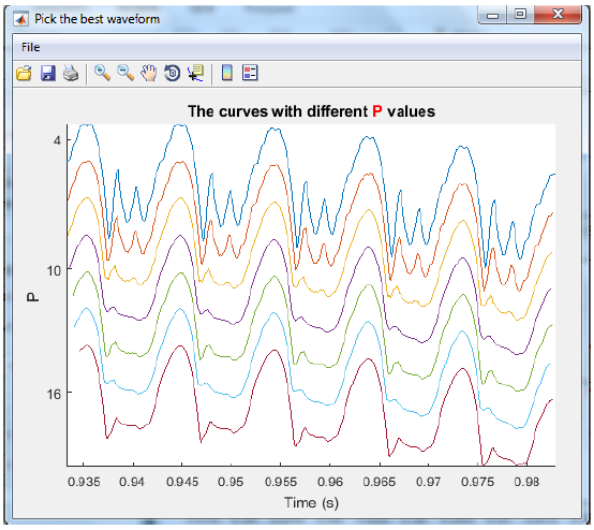
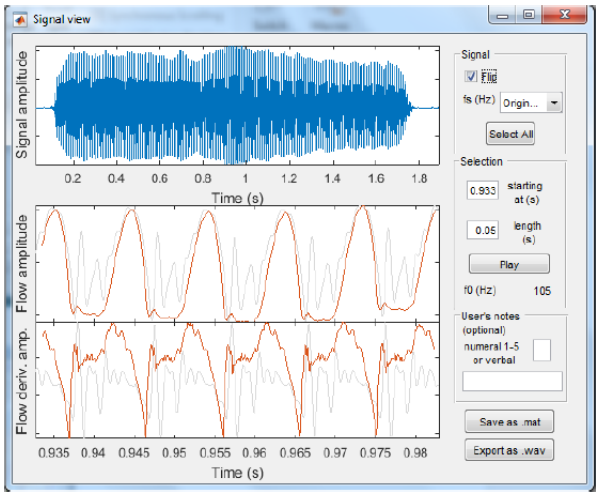


Figure 3: Signal view window after the user has made his/her selection for the best glottal flow estimate.



the mouse. Finally, the selected glottal flow and its derivative appear into signal view window (Figure 3).

The procedure described above is flexible because it enables running the inverse filtering analysis either in an automatic or a semi-automatic mode. In the former, no user feedback is required by Aalto Aparat (i.e. default parameter values are used for the corresponding GIF algorithm). In the latter, the tool allows utilizing subjective criteria in letting the user to take advantage of his/her expertise to select the waveform that is he/she considers to be the best estimate of the unknown true glottal flow.

3.3. Step 3: Parameterization

After inverse filtering, the obtained glottal flow is parameterized in a completely automatic manner using a multitude of parameters (for further details, see Airas, 2008). Parameterization is activated from the corresponding menu, after which a new window pops up indicating the obtained parameter values (Figure 4). By pressing the corresponding button (Figure 4, "LF-

Figure 4: Results of parameterizing the glottal flow shown in Figure 3. Parameters are organized into time-based, frequency-based and LF model-based.

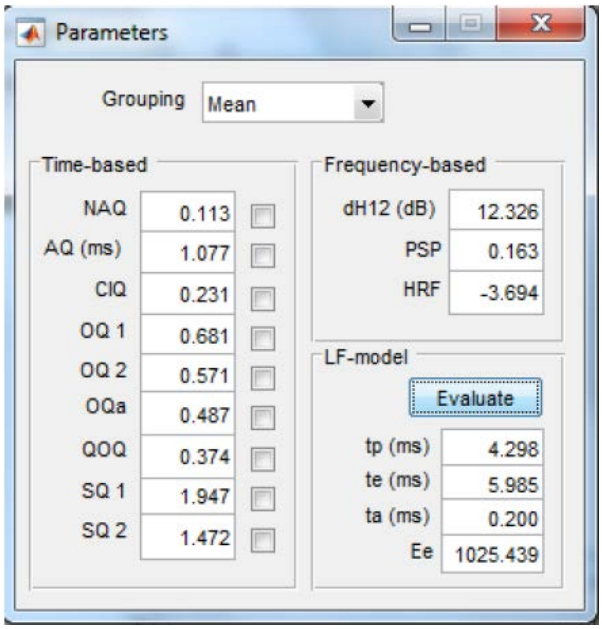
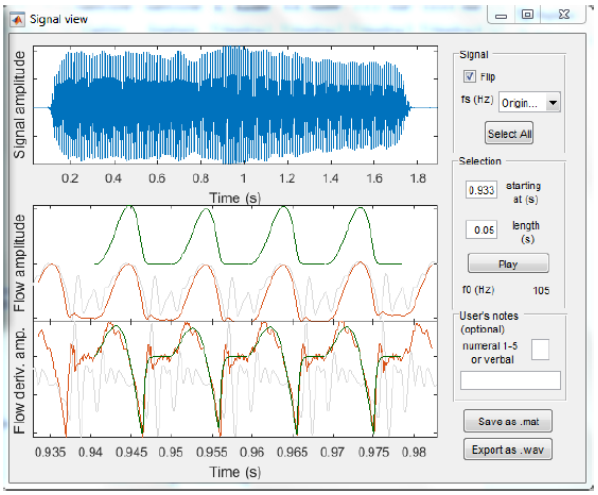


Figure 5: Signal view window after the user has selected the LF model based parameterization. Top pane shows the input speech signal. Middle pane shows the LF-synthesized flow (upper) and the estimated flow (lower). Bottom pane depicts two flow derivatives on top of each other: the one computed from the estimated flow (red) and the LF-modelled one (green).



model, Evaluate"), the system matches the obtained glottal flow derivative with the LF pulse form, and shows the obtained LF parameter values (Figure 4, right bottom corner). In addition, Aalto Aparat also depicts the output of the LF fitting by depicting both the synthetic flow and its derivative as time-domain waveforms (Figure 5).

3.4. Step 4: Exporting data

Aalto Aparat enables saving both the obtained parameter values as well as two signals (estimated glottal flow and input speech, both as time-domain signals spanning the frame that was selected in the GIF analysis). In a typical inverse filtering session, the user

has many input signals to be analyzed. Once all of these have been processed, one by one, the system enables combing the corresponding parameter data in a single array which can be later imported to, for example, Excel to be further processed (e.g. for statistical analysis and visualization).

4. CONCLUSIONS

A new glottal inverse filtering and voice source parameterization tool, Aalto Aparat, has been described in this article. Aalto Aparat is based on its predecessor, TKK Aparat, both offering a graphical interface. By using it, a user with no programming skills can conduct glottal inverse filtering analysis and parameterization of the estimated flow signals. The tool has been programmed in MATLAB but it can be downloaded as a stand-alone package which can be run without having access to MATLAB. In comparison to its predecessor, Aalto Aparat involves a few major changes, the most important one being an opportunity to use a recently proposed potential GIF method, QCP. In addition, the Aalto Aparat stand-alone package can be installed into two operating systems (Windows and Mac OS).

Usability of Aalto Aparat has not been formally evaluated. However, the tool's predecessor, TKK Aparat, went through a formal evaluation process in which the interface was developed into its current form by collecting user feedback in a usability test (Airas, 2008). As a conclusion, the usability test of TKK Aparat indicated that the system can be easily taken advantage of by anyone who have basic knowledge in glottal inverse filtering. Since the user interface of Aalto Aparat has been changed only slightly from that of TKK Aparat (e.g. by correcting minor bugs), we argue that also the Aalto Aparat software is easy to use by anyone who knows the basics of glottal inverse filtering.

Researchers interested in glottal inverse filtering and voice source parameterization are welcome to download the Aalto Aparat software free of charge from Aalto Aparat (2016).

5. REFERENCES

Aalto Aparat. (2016). Retrieved from http://research.spa.aalto.fi/projects/aparat/.

Airaksinen, M., Raitio, T., Story, B., & Alku, P. (2014). Quasi closed phase glottal inverse filtering analysis with weighted linear prediction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(3), 596–607.

Airas, M. (2008). TKK Aparat: An environment for voice inverse filtering and parameterization. *Logopedics, Phoniatrics and Vocology*, *33*(1), 49–64.

Alku, P. (1992). Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, *11*(2–3), 109–118.

Alku, P. (2011). Glottal inverse filtering analysis of human voice production: A review of estimation and parameterization methods of the glottal excitation and their applications. *Sadhana. Academy Proceedings in Engineering Sciences*, *36*(5), 623–650.

Alku, P., Bäckström, T., & Vilkman, E. (2002). Normalized amplitude quotient for parameterization of the glottal flow. *Journal of the Acoustical Society of America*, *112*(2), 701–710.

Alku, P., Pohjalainen, J., Vainio, M., Laukkanen, A.-M., & Story, B. (2013). Formant frequency estimation of high-pitched vowels using weighted linear prediction. *Journal of the Acoustical Society of America*, *134*(2), 1295–1313.

Alku, P., Strik, H., & Vilkman, E. (1997). Parabolic spectral parameter: A new method for quantification of the glottal flow. *Speech Communication*, *22*, 67–79.

Boersma, P., & Weenink, D. (2013). Praat: doing phonetics by computer. Retrieved from http://www.praat.org/.

Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, A. (2014). Covarep: A collaborative voice analysis repository for speech technologies. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 960–964).

Drugman, T. (n.d). Retrieved from http://tcts.fpms.ac.be/~drugman/Toolbox/.

Drugman, T., Alku, P., Alwan, A., & Yegnanarayana, B. (2014). Glottal source processing: from analysis to applications. *Computer, Speech and Language*, *28*(5), 1117–1138.

El-Jaroudi, A., & Makhoul, J. (1991). Discrete all-pole modeling. *IEEE Transactions on Signal Processing*, *39*, 411–423.

Fant, G., Liljencrants, J., & Lin, Q. (1985). A four-parameter model of glottal flow. *Speech Transmission Laboratory – Quarterly Progress and Status Report*, *26*(4), 1–13.

Gauffin-Lindqvist, J. (1965). Studies of the voice source by means of inverse filtering. *Speech Transmission Laboratory – Quarterly Progress and Status Report*, *6*(2), 8–13.

Gobl, C., & Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, *40*, 189–212.

Granqvist, S., Hertegård, S., Larsson, H., & Sundberg, J. (2003). Simultaneous analysis of vocal fold vibration and transglottal airflow: exploring a new experimental setup. *Journal of Voice*, *17*, 312–330.

Kane, J. (2012). Tools for analysing the voice: Developments in glottal source and voice quality analysis (Doctoral dissertation). Trinity College Dublin.

Kane, J. (2013). Retrieved from https://github.com/jckane/Voice_Analysis_Toolkit.

Kreiman, J., Shue, Y.-L., Chen, G., Iseli, M., Gerratt, B., Neubauer, J., & Alwan, A. (2012). Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation. *Journal of*

Alku, P., Pohjalainen, H., & Airaksinen, M. (2019). Aalto Aparat: A freely available tool for glottal inverse filtering and voice source parameterization. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 5–10). Málaga: Universidad de Málaga.

the Acoustical Society of America, *132*(4), 2625–2632.

Lehto, L., Airas, M., Björkner, E., Sundberg, J., & Alku, P. (2007). Comparison of two inverse filtering methods in parameterization of the glottal closing phase characteristics in different phonation types. *Journal of Voice*, *21*(2), 138–150.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, *63*(3), 561–580.

Rothenberg, M. (1973). A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. *Journal of the Acoustical Society of America*, *53*(6), 1632–1645.

Shue, Y.-L., Keating, P., Vicenik, C., & Yu, K. (2011). VoiceSauce: A program for voice analysis. In *Proceedings of the 17th International Congress on Phonetic Sciences* (pp. 1846–1849).

Timcke, R., von Leden, H., & Moore, P. (1958). Laryngeal vibrations: measurements of the glottic wave. *Archives of Otolaryngology*, *68*, 1–19.

Titze, I., & Sundberg, J. (1992). Vocal intensity in speakers and singers. *Journal of the Acoustical Society of America*, *91*(5), 2936–2946.

Tolvan Data. (n.d). Retrieved from http://www.tolvan.com/.

VoiceSauce. (2016). VoiceSauce: A program for voice analysis. Retrieved from http://www.seas.ucla.edu/spapl/voicesauce/.

Wölfel, M., & McDonough, J. (2005). Minimum variance distortionless response spectral estimation. *IEEE Signal Processing Magazine*, *22*(5), 117–126.

Alku, P., Pohjalainen, H., & Airaksinen, M. (2019). Aalto Aparat: A freely available tool for glottal inverse filtering and voice source parameterization. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 5–10). Málaga: Universidad de Málaga.

Subsidia: Tools and Resources for Speech Sciences

3

The phonetic approach of voice qualities: challenges in corresponding perceptual to acoustic descriptions

Zuleica Camargo[1], Sandra Madureira[1], Nathalia dos Reis[1] and Albert Rilliard[2, 3]

[1] Pontifical Catholic University of São Paulo
[2] LIMSI, CNRS, Université Paris-Saclay
[3] Federal University of Rio de Janeiro
e-mail: zcamargo@pucsp.br, fononana2@gmail.com, albert.rilliard@limsi.fr

Citation / Cómo citar este artículo: Camargo, Z., Madureira, S., dos Reis, N., & Rilliard, A. (2019). The phonetic approach of voice qualities: challenges in corresponding perceptual to acoustic descriptions. In J. M. Lahoz-Bengoechea, & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 11–17). Málaga: Universidad de Málaga.

ABSTRACT: This study introduces an innovative approach to the phonetic investigation of voice qualities, comprising the application of the Vocal Profile Analysis Scheme (VPAS) to describe perceived voice quality settings, the extraction of acoustic measures, and a statistical link between perception and acoustics, weighting the relative proximity of controlled factors. The corpus was perceptually annotated by the VPAS and data from 44 speakers were grouped in terms of the most frequent combinations in the VPAS system, generating two vocal profiles; the "Wide" and the "Short" vocal tract kinds of profiles. Acoustic measures (f_0, intensity, signal to noise ratio, spectral slope and the first formant) were extracted. Statistical analysis weighs the relative links between the voice quality profiles and the acoustic measures, compared to linguistic and gender constraints. f_0 measures were found to be the most relevant to establish perceptual and acoustic correlations. Some singularities of the correspondences detected are discussed.

Keywords: voice quality; auditory perception; speech acoustics; phonetics; statistical analysis.

RESUMEN: Este estudio presenta un enfoque innovador para la investigación de las cualidades de la voz. El método incluye la aplicación del Esquema de Análisis del Perfil Vocal (VPAS) para describir los ajustes de la cualidad de voz que se perciben, junto con la extracción de medidas acústicas y un análisis estadístico sobre la relación entre la percepción y la acústica, que permite calibrar la proxiimdad relativa de los factores controlados. Se anotó perceptivamente un corpus siguiendo el modelo VPAS y se agruparon los datos de 44 hablantes en función de las combinaciones más frecuentes del sistema VPAS, lo que dio lugar a dos grandes tipos de perfil vocal: el "Ancho" y el "Corto". Se extrajeron ciertas medidas acústicas (f_0, intensidad, ratio entre la señal y el ruido, declinación espectral y primer formante). El análisis estadístico mide el peso de la relación entre los perfiles de la cualidad vocal y las medidas acústicas, comparado con restricciones lingüísticas y de sexo. La f_0 resultó ser la medida más relevante a la hora de establecer correlaciones entre lo perceptivo y lo acústico. El artículo comenta algunos detalles de las correspondencias detectadas.

Palabras clave: cualidad de voz; percepción auditiva; acústica del habla; fonética; análisis estadístico.

1. INTRODUCTION

Voice quality descriptions (and / or evaluation scales) tend to focus on acoustic-perceptual correlations (Dejonckere et al., 1995; Hammarberg & Gauffin, 1995; Kreiman & Gerratt, 2000; Kreiman & Sidtis, 2011; Rabinov, Kreiman, Gerratt, & Bielamowicz, 1995). The relevant literature is also plenty of descriptions of perceptual labels and their acoustic and / or physiologic counterparts, especially for phonatory adjustments, i. e. the voice source events (d'Alessandro, 2006; d'Alessandro, Darsinos, & Yegnanarayana, 1998; Dejonckere et al., 1995; Garellek, 2014; Hammarberg & Gauffin, 1995; Rabinov et al., 1995; Sundberg & Gauffin, 1979). Some of them are based on voice quality settings described in the Vocal Profile Analysis Scheme (VPAS) system (Laver, Wirz, Mackenzie, & Hiller, 1981).

The concept of voice quality used in VPAS derives from the model of phonetic description of voice quality by Laver (1980). As a phonetically grounded model, voice quality is here considered as the result of phonatory and articulatory settings, that is, the result of specific adjustments of the vocal folds and of the articulators during speech.

To perform the voice quality evaluations based on the VPAS, judges need a phonetic background and experience on the use of the profile. The basic analytical unit is the Voice Quality Setting (VQS), a long-term muscular tendency in the vocal apparatus: supralaryngeal (articulators and resonators), laryngeal / phonatory (vocal folds vibrations) and muscular tension activity (laryngeal and supralaryngeal). The VQS are described as variations from a reference setting, the neutral one, in which no effect is found in longitudinal or transversal plans of the vocal tract, and there is no variation in terms of its muscular tension activity. For the neutral setting, the vibration of vocal folds must be periodic.

The VPAS is applied in two passes. The first pass comprises the identification of non-neutral VQS. The second pass involves the grading of non-neutral VQS in a scalar degree, generally from 1 to 6.

It is important to reinforce that the phonetic description of voice quality model (Laver, 1980) follows two principles: susceptibility and compatibility.

The susceptibility principle accounts for the fact that some speech segments are more susceptible to the effects of specific voice quality settings (Laver, 1980; Mackenzie-Beck, 1999, 2005). For example, oral speech segments are more susceptible to nasal settings than nasal segments and vice-versa; voiced sounds (vowels and some consonants) are susceptible to phonatory settings, like breathy and creaky voices. For the sake of describing phonetic voice quality settings, the corpus design must take into account the principle of susceptibility, making use of key speech segments and key sentences, containing the susceptible segments.

The compatibility principle states that some VQS can co-occur and others cannot (Laver, 1980; Mackenzie-Beck, 1999, 2005). Some settings can be easily combined, because they are physiologically compatible (lowered larynx and retracted body tongue; lowered larynx and pharyngeal expansion; raised larynx and pharyngeal constriction; raised larynx and laryngeal hyperfunction, for example). Other settings cannot be combined, since they are physiologically incompatible (lowered larynx and raised larynx; pharyngeal constriction and pharyngeal expansion).

So, the speakers' vocal profiles can be drawn in terms of one or more adjustments that can be combined during the time they are speaking, as a long-term composed VQS. The literature indicates recurrent tendencies of grouped voice quality events detected by different perceptual scales (d'Alessandro, 2006; d'Alessandro, Darsinos, & Yegnanarayana, 1998; Dejonckere et al., 1995; Hammarberg & Gauffin, 1995; Kreiman & Gerratt, 2000; Kreiman & Sidtis, 2011; Laver, 1980; Mackenzie-Beck, 2005; Mackenzie-Beck & Schaeffler, 2015; Rabinov, Kreiman, Gerratt, & Bielamowicz, 1995). These findings are also applicable to the VPAS system, taking into account the compatibility principle of voice quality settings (French, Harrison, Hughes, & Stevens, 2015; Laver, 1980; Mackenzie-Beck, 1999, 2005; Mackenzie-Beck & Schaeffler, 2015; Robieux & Meunier, 2015).

Furthermore, the discussion of the perceptual and acoustic counterparts of voice qualities is important in order to foster knowledge about the links between production and perception (Kreiman & Sidtis, 2011). Considering that voice quality is a prosodic element which has linguistic, paralinguistic and extralinguistic functions, the applications of voice quality analysis are multiple: language descriptions (cross-linguistic variations; Esling, 2000); expressivity investigations (voice expressivity; Barbosa, 2009; Fontes & Madureira, 2015); evaluation and rehabilitation of voice disorders (clinical procedures; Dejonckere et al., 1995; Gillespie, Dastolfo, Magid, & Gartner-Schmidt, 2014; Hammarberg & Gauffin, 1995; Maryn & Weenink, 2015), technological development (voice recognition and synthesis for many purposes, including Augmentative and Alternative Communication Systems) and forensic purposes (speaker recognition; French et al., 2015).

The answers to the following questions are pursued in this work: how can we investigate the perceptual and acoustic correspondences of voice qualities in a system offering so many parameters (VQS)? What about the relevance of some VQS (and their combinations) in the acoustic arena? How can we improve our voice quality descriptions systems?

This investigation aimed at addressing the correspondences between laryngeal and pharyngeal voice quality settings perceptually described by a phonetic grounded profile, the VPAS, and acoustic measures. The acoustic measures were chosen among a restricted set, so to address these phonetic and articulatory phenomena.

2. METHODS

2.1. Corpus description

The general voice quality database is composed of semi-spontaneous speech samples and repetitions of three key sentences (based on the susceptibility principle), read by 278 Brazilian Portuguese (BP) speakers. The voice quality database was perceptually annotated by means of the VPAS system.

Since susceptibility is an important issue for a phonetically grounded voice quality analysis, the use of key speech segments was proposed. For the present investigation, a sub-selection of [a] vowels was extracted from each of the three sentences of the corpus. Because of labeling costs, 44 subjects were selected (10 male and 34 female, ranging from 18 to 58 years old, with a mean age of 30), for a total of 826 vowel samples evaluated.

Four words were targeted from the three key sentences: *f<u>a</u>la*, *L<u>a</u>ra* and *cid<u>a</u>de* (with two occurrences for the latter: at the beginning of the key sentence and at the middle of the key sentence). Each sentence was

Camargo, Z., Madureira, S., dos Reis, N., & Rilliard, A. (2019). The phonetic approach of voice qualities: challenges in corresponding perceptual to acoustic descriptions. In J. M. Lahoz-Bengoechea, & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 11–17). Málaga: Universidad de Málaga.

repeated several times by each speaker; the number of repetition depends on the speaker.

As an open, lowered, backed and non-rounded vowel, [a] was found to be a susceptible segment for laryngeal and pharyngeal settings in perceptual evaluation (French et al., 2015; Laver, 1980; Mackenzie-Beck, 1999, 2005; Mackenzie-Beck & Schaeffler, 2015; Robieux & Meunier, 2015). For acoustic analysis, this vowel was chosen for its stability, and also because it is the vowel of choice for many investigations involving voice source analysis (Hanson, 1997), including intensity measures (Liénard & Barras, 2013) and periodic-aperiodic decomposition (d'Alessandro et al., 1998).

2.2. Perceptual and acoustic analysis

The perceptual (VPAS) parameters were estimated for each sentence in the dataset described by two expert raters and revised by one, with specific focus on the target vowels [a]. There were 37 parameters perceptually ranked on the profile (Figure 1), with degrees concentrated on a 0 to 4 range. Since VPAS judgments are componential, many zero scores (i. e. neutral voice *for that setting*) were generated in most vocal profiles.

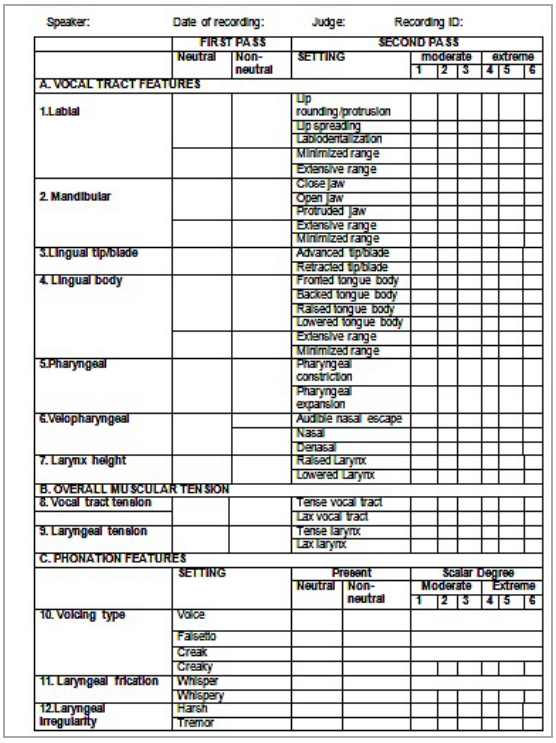
Acoustic measures were extracted from the [a] vowels (the same that were perceptually analyzed along with the VPAS system), by means of scripts run in Praat (Boersma & Weenink, 2016), and thanks to a Matlab implementation of a periodic-aperiodic decomposition algorithm (d'Alessandro et al., 1998):

- Fundamental frequency (f_0), expressed in semitones (with a reference frequency of 1 Hz), a measure notably linked to pitch and register (d'Alessandro, 2006).
- A-weighted *intensity*, expressed in dBA, linked to the perception of voice strength (Liénard & Barras, 2013; Traunmüller & Eriksson, 2000).
- Harmonic-to-Noise Ratio (HNR), expressed in dB, and measured after a periodic-aperiodic decomposition (d'Alessandro et al., 1998), taking into account both additional and structural noises in voiced segments.
- The first formant (F1), expressed in Hz, that is linked to the size of the back cavity (Apostol, Perrier, & Bailly, 2004) and to jaw opening (Erickson, Suemitsu, Shibuya, & Tiede, 2012).
- Amplitude difference between the first harmonic and the third formant (H1−A3), expressed in dB, that is linked to tension and voice strength (Hanson, 1997).

2.3. Integrating perceptual and acoustic data

A multiple factorial analysis was run on the two datasets (VPAS settings and acoustic measures), taking into account all the data. This analysis did not extract links between both datasets, notably because of the scarcity of most VPAS annotations: the presence of zeros as soon as a setting is absent did not mean the acoustic parameters won't change, mostly because other settings may have an effect on them. It is thus difficult to match both datasets. To bypass that limitation, analyses of the links between each acoustic measure and the VPAS parameters were run. In order to address the limitation introduced by scarcity, VPAS parameters that show frequent correlations were combined, in order to have a more robust estimation of an aggregated perceptual dimension.

Figure 1: The *Vocal Profile Analysis Scheme VPAS* 2007 version (Laver & Mackenzie-Beck, 2007).



Speaker: Date of recording: Judge: Recording ID:

	FIRST PASS		SECOND PASS						
	Neutral	Non-neutral	SETTING	moderate			extreme		
				1	2	3	4	5	6
A. VOCAL TRACT FEATURES									
1.Labial			Lip rounding/protrusion						
			Lip spreading						
			Labiodentalization						
			Minimized range						
			Extensive range						
2. Mandibular			Close jaw						
			Open jaw						
			Protruded jaw						
			Extensive range						
			Minimized range						
3.Lingual tip/blade			Advanced tip/blade						
			Retracted tip/blade						
4. Lingual body			Fronted tongue body						
			Backed tongue body						
			Raised tongue body						
			Lowered tongue body						
			Extensive range						
			Minimized range						
5.Pharyngeal			Pharyngeal constriction						
			Pharyngeal expansion						
6.Velopharyngeal			Audible nasal escape						
			Nasal						
			Denasal						
7. Larynx height			Raised Larynx						
			Lowered Larynx						
B. OVERALL MUSCULAR TENSION									
8. Vocal tract tension			Tense vocal tract						
			Lax vocal tract						
9. Laryngeal tension			Tense larynx						
			Lax larynx						
C. PHONATION FEATURES									
	SETTING		Present			Scalar Degree			
			Neutral	Non-neutral		Moderate			Extreme
					1	2	3	4	5	6
10. Voicing type	Voice									
	Falsetto									
	Creak									
	Creaky									
11. Laryngeal frication	Whisper									
	Whispery									
12.Laryngeal irregularity	Harsh									
	Tremor									

The voice quality settings detected by the VPAS were firstly categorized in four groups:

- Supralaryngeal (laryngeal height) VQS
- Supralaryngeal (pharyngeal) VQS
- Phonatory VQS
- (Supra)laryngeal tension VQS

In a second stage, for the sake of corresponding perceptual to acoustic descriptions, the compatibility principle was applied, generating the frequent vocal profiles, also based on some references (Camargo, Rusilo & Madureira, 2011; French et al., 2015; Laver, 1980; Mackenzie-Beck, 1999; Mackenzie-Beck & Schaeffler, 2015; Robieux & Meunier, 2015). The following vocal profiles were then generated from the aggregated VQS: "Short vocal tract", "Wide vocal tract". The vocal profile named "Short vocal tract" regroups VPAS parameters which cause vocal tract length and width reduction:

- Raised larynx VQS
- Laryngeal hyperfunction VQS
- Pharyngeal constriction VQS
- Closed jaw VQS
- Spread lips VQS

Camargo, Z., Madureira, S., dos Reis, N., & Rilliard, A. (2019). The phonetic approach of voice qualities: challenges in corresponding perceptual to acoustic descriptions. In J. M. Lahoz-Bengoechea, & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 11–17). Málaga: Universidad de Málaga.

The vocal profile named "Wide vocal tract" regroups VPAS parameters which cause the vocal tract length and width expansion:

- Lowered larynx VQS
- Laryngeal hypofunction VQS
- Pharyngeal expansion VQS
- Creaky voice VQS

The "Wide" and "Short" vocal profiles were attributed grades according to the number of VQS (VPAS parameters) perceived. If no VQS was perceived these vocal profiles were assigned "0". If just one VQS was perceived they were assigned "1" and if two or more VQS were perceived "+2" was assigned to the "Wide vocal profile" and "−2" to the "Short vocal profile".

The combination of the two vocal profiles (Wide *minus* Short) created a single parameter, named Size. These vocal profiles, being linked to the vocal tract length (and width) and to the muscular tension of the vocal apparatus, are supposed to be correlated to acoustic data, such as f_0, intensity, noise, first formant and spectral slope measures.

This project was approved by Ethics Committee (number 101/11).

2.4. Statistical analysis

Models of analysis of variance (ANOVA) were fitted to each acoustic parameter, so to explore which of the several factors controlled in the corpus and through the VPAS annotation do have an explanatory power for the measures' variation. The controlled factors are the speakers' *Gender*, the *Word* that contains the [a] vowel (4 levels), and the VPAS *Size* aggregate (5 levels), and their interactions. The alpha level was set at 5 %.

3. RESULTS

3.1. Fundamental frequency

The ANOVA model explains more than two thirds of the variance ($R^2 = 0.72$). The three main factors are significant, as well as the triple interaction. The factors that account for most of the explained variance are (in decreasing order): unsurprisingly the Gender (partial $\eta^2 = 0.52$), the Word (partial $\eta^2 = 0.31$) and the Size aggregate (partial $\eta^2 = 0.21$). The triple interaction accounts for only 3 % of the variance. Figure 2 shows the effect of both Gender and Size on f_0.

For female voices, f_0 changes are observed immediately with differences in Size—f_0 rising with smaller sizes. Meanwhile, these changes do reach a ceiling rapidly, while females continue to lower their pitch for Size above 1. A tendency for such changes may be also true for males, but differences are not significant.

3.2. First formant

The ANOVA model for F1 explains less than a third of the variance ($R^2 = 0.27$). All factors have a significant effect, and the factors that account for most of the explained variance are, in decreasing order, Word (partial $\eta^2 = 0.15$), Size (partial $\eta^2 = 0.10$), and Gender (partial $\eta^2 = 0.05$). Figure 3 shows the effect of both Gender and Size on F1.

Figure 2: Boxplot showing the distribution of f_0 estimations for the 5 levels of Size (from −2 to +2), for females (grey) and males (white).

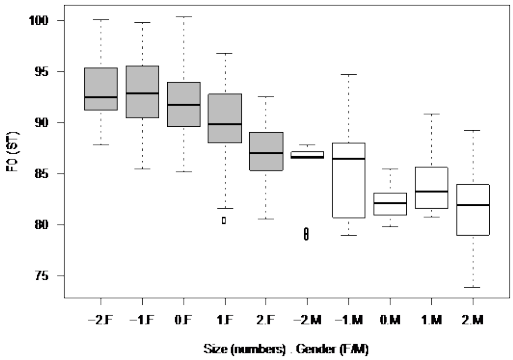
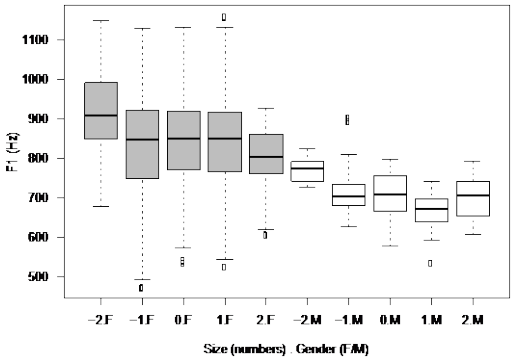


Figure 3: Boxplot showing the distribution of F1 estimation for the 5 levels of Size (from −2 to +2), for females (grey) and males (white).



The first formant is linked to the back cavity (Apostol et al., 2004) and to jaw opening (Erickson et al., 2012); the observed changes are mostly linked to female voices, as for f_0, but the effects are significant only for the higher levels of the Size factor (−2 or +2), respectively with a formant rise for "small" voice, and a formant fall for "wide" voices.

3.3. Intensity and spectral slope

The ANOVA model for intensity explains about half the variance ($R^2 = 0.43$); all the factors and their interactions are significant. The factors that account for most of the explained variance in intensity are, in decreasing order, Word (partial $\eta^2 = 0.35$), Size (partial $\eta^2 = 0.04$), and its interaction with Gender (partial $\eta^2 = 0.05$), while Gender accounts for about 1 % (partial $\eta^2 = 0.01$). Figure 4 reports the changes of intensity according to Size and gender. The ANOVA model for H1−A3 shows similar patterns, but the model is very messy ($R^2 = 0.27$). It is still linked mainly to the Word position (partial $\eta^2 = 0.15$), and then to Size (partial $\eta^2 = 0.10$) and Gender (partial $\eta^2 = 0.05$).

Changes in intensity are mainly explained by the position of word in the sentence, which is linked to the declination line. For changes that are linked to VPAS,

Figure 4: Boxplot showing the distribution of intensity for the 5 levels of Size (from −2 to +2), for females (grey) and males (white).

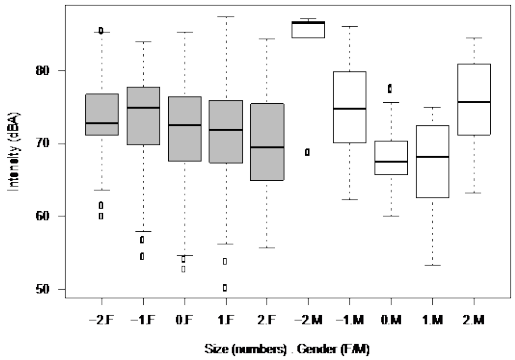
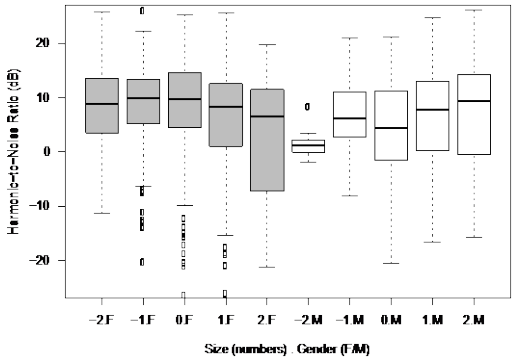


Figure 5: Boxplot showing the distribution of HNR estimation for the 5 levels of Size (from −2 to +2), for females (grey) and males (white).



they show an interaction with Gender: it seems that mostly males do vary intensity in correlation with the Size factor, increasing it when their voices depart from the neutral setting. A slight declination is observed in female voices characterized by lower pitches and wider tracts or laxer voice qualities and thus may be explained by changes in pitch (Titze & Sundberg, 1992).

3.4. Harmonic-to-Noise Ratio

The ANOVA model for HNR explains only a small part of the variance ($R^2 = 0.16$), if all factors have a significant effect. The factors that account for a part of the explained variance are, in decreasing order, Word (partial $\eta^2 = 0.07$) and Size (partial $\eta^2 = 0.02$). Figure 5 shows the effect of both Gender and Size on HNR.

One may observe on the graphs that only the most extreme value of the size factor do show more aperiodicities—typically for the lowest pitch voice (which are also at a low intensity) in female speakers, that could be related to creaky phenomenon.

This aggregate is not mostly related to this measure of noise, and one may rather observe potential relation with other possible aggregates.

4. DISCUSSION

To face the challenge of relating perceptual and acoustic data from voice qualities, speaker-standardized and non-standardized acoustic measures were initially taken into account. Non-standardized measures provided more interesting results than standardized measures, as soon as the speaker's gender was considered in the statistical models. This is linked to the fact that standardization, when removing speaker-specific changes, also removes the specificities of voice quality that characterize the speaker's voice.

Among the various acoustic measures extracted from the corpus, f_0 was found to be the most relevant measure to relate the "Short" and "Wide" VPAS aggregates (cf. Figure 2). A point to consider in this discussion is the set of VPAS parameters that contribute to this general "size" vocal profile. Despite the fact that we found similar effects on f_0 measures, it is important to consider that many possible VQS combinations could be implemented in the speakers' vocal tract, leading to increasing (or decreasing) f_0 in distinct proportions. Some of these combinations of settings include notably a dimension of noise, that would probably be better expressed by measures such as harmonic-to-noise ratio (d'Alessandro, 2006; d'Alessandro et al., 1998).

Among the findings of this study is an effect of gender on the acoustic parameters which are relied on this "Short" / "Wide" vocal profiles. In this corpus, females clearly go for pitch as a primary cue, and then for F1; on the contrary males seems to rely on intensity rather than pitch.

This result is to be taken with caution, because of the relatively small set of male speakers included in the corpus (10 speakers, compared to the 34 female speakers). Meanwhile, it could be related to a social habit in BP for males to use a lower voice register, and / or for females to use a comparatively higher voice register. This could be related and explained in a similar way to the difference in pitch described between Japanese and Dutch women, that is linked by Van Bezooijen (1995) to the representations of gender in these two societies, and expected to be found in males also.

Yet, some factors influencing voice quality patterns have not been addressed in studies focusing acoustic-perceptual correlations. Some of them are related to speakers age and gender normalization, and intra-speaker variations. Other challenges are related to the overlapping of voice quality events and the degree of influence of the setting in the general vocal profile.

To address these limitations, we proposed to explore the vocal profiles, i. e. the combinations of voice quality settings that tend to be productive in daily communications, and even, in the voice disorder arena. In future explorations, the relevance of each VQS for the final vocal profile definition must also be addressed. The difficulty to focus on long-term events in relating them to intermittent or shot-term occurrences may also interfere in the study of acoustic correlates of VQS, like voice breaks and sudden voice quality changes. They were not frequent in the *corpus* analyzed, because this corpus was annotated in VPAS and vowel samples were revised to search for some

specific events in the key-speech segment (the vowel [a]).

To face all these limitations in describing voice qualities, a statistical model has been proposed and is meant to be improved in later works.

5. CONCLUSIONS

In this investigation, we depart from the phonetic approach of voice qualities, taking into account the susceptibility and compatibility theoretical principles. To fit the principles of the model by Laver et al. (1981), we had to consider the combinations of laryngeal and pharyngeal VQS and the inherent phonetic characteristics of the speech segments.

The methodological procedures made it possible to identify the relevance of f_0 measures as a main cue, and F1, intensity and H1−A3 as secondary cues, to describe perceptual data related to pharyngeal and laryngeal adjustments, generating the "Wide" and "Short" vocal tract kinds of profiles.

The findings provide evidence in favor of the relevance of the phonetic description of voice qualities.

6. REFERENCES

Apostol, L., Perrier, P., & Bailly, G. (2004). A model of acoustic interspeaker variability based on the concept of formant–cavity affiliation. *The Journal of the Acoustical Society of America*, *115*(1), 337–351.

Barbosa, P. A. (2009). Detecting changes in speech expressiveness in participants of a radio program. In *Proceedings of Interspeech* (pp. 2155–2158).

Boersma, P. & Weenink, D. (2016). Praat: doing phonetics by computer [Computer program]. Version 6.0.20, retrieved 3 September 2016 from http://www.praat.org/

Camargo, Z., Rusilo, L. C., Madureira, S. (2011). Evaluating speech samples designed for the Voice Profile Analysis Scheme for Brazilian Portuguese. Presented at the Fourth ISCA Tutorial and Research Workshop on Experimental Linguistics, Paris, France.

d'Alessandro, C. (2006). Voice source parameters and prosodic analysis. In S. Sudhoff, D. Lenertova, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, & J. Schlieer (Eds.), *Language, context, and cognition: Methods in empirical prosody research* (pp. 63–87). Berlin: Walter de Gruyter.

d'Alessandro, C., Darsinos, V., & Yegnanarayana, B. (1998). Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources. *IEEE Transactions on Speech and Audio processing*, *6*(1), 12–23.

Dejonckere, P. H., Remacle, M., Fresnel-Elbaz, E., Woisard, V., Crevier-Buchman, L., & Millet, B. (1995). Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements. *Revue de laryngologie-otologie-rhinologie*, *117*(3), 219–224.

Erickson, D., Suemitsu, A., Shibuya, Y. & Tiede, M. (2012). Metrical structure and production of English rhythm. *Phonetica*, *69*, 180–190.

Esling, J. H. (2000). Crosslinguistic aspects of voice quality. In R. D. Kent, & M. J. Ball (Eds.), *Voice quality measurement*. San Diego: Singular Publishing Group.

Fontes, M. A. S., Madureira, S. (2015). Gestural prosody and the expression of emotions: a perceptual and acoustic experiment. Presented at the 18th International Congress of Phonetic Sciences, Glasgow, Scotland.

French, P., Harrison, P., Hughes, V., & Stevens, L. (2015). The vocal tract as a biometric: output measures, interrelationships, and efficacy. Presented at the 18th International Congress of Phonetic Sciences, Glasgow, Scotland.

Garellek, M. (2014). Voice quality strengthening and glottalization. *Journal of Phonetics*, *45*, 106–113.

Gillespie, A. I., Dastolfo, C., Magid, N., & Gartner-Schmidt, J. (2014). Acoustic analysis of four common voice diagnoses: moving toward disorder-specific assessment. *Journal of Voice*, *28*(5), 582–588.

Hammarberg, B., & Gauffin, J. (1995). Perceptual and acoustic characteristics of quality differences in pathological voices as related to physiological aspects. In O. Fujimura, & M. Hirano (Eds.), *Vocal Fold Physiology: Voice Quality Control* (pp. 283–303). San Diego: Singular Publishing Group.

Hanson, H. M. (1997). Glottal characteristics of female speakers: Acoustic correlates. *The Journal of the Acoustical Society of America*, *101*(1), 466–481.

Kreiman, J., Gerratt B. (2000). Measuring vocal quality. In R. D. Kent & M. J. Ball (Eds.), *Voice quality measurement* (pp. 73–101). San Diego: Singular Publishing Group.

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. Oxford: Wiley-Blackwell.

Laver, J. (1980). *The phonetic description of voice quality*. Cambridge: Cambridge University Press.

Laver, J., & Mackenzie-Beck, J. (2007). Vocal Profile Analysis Scheme-VPAS. *Queen Margaret University College-QMUC, Speech Science Research Centre, Edinburgh*.

Laver, J., Wirz, S., Mackenzie, J., & Hiller, S. (1981). A perceptual protocol for the analysis of vocal profiles. *Edinburgh University Department of Linguistics Work in Progress*, *14*, 139–155.

Liénard, J.-S. & Barras, C. (2013). Fine-grain voice strength estimation from vowel spectral cues. In *Proceedings of Interspeech* (pp. 128–132).

Mackenzie-Beck J. (1999). Organic variation of the vocal apparatus. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The handbook of phonetic sciences* (pp. 256–297). Malden, MA: Blackwell.

Mackenzie-Beck, J. (2005). Perceptual analysis of voice quality: the place of vocal profile analysis. In

W. J. Hardcastle, & J. Mackenzie-Beck (Eds.), *A figure of speech: a festschrift for John Laver* (pp. 285–322). Mahwah: Lawrence Erlbrum Associates.

Mackenzie-Beck, J., & Schaeffler, F. (2015). Voice quality variation in Scottish adolescents: gender versus geography. Presented at the 18th International Congress of Phonetic Sciences, Glasgow, Scotland.

Maryn, Y., & Weenink, D. (2015). Objective dysphonia measures in the program Praat: smoothed cepstral peak prominence and acoustic voice quality index. *Journal of Voice*, *29*(1), 35–43.

Rabinov, C. R., Kreiman, J., Gerratt, B. R., & Bielamowicz, S. (1995). Comparing reliability of perceptual ratings of roughness and acoustic measures of jitter. *Journal of Speech, Language, and Hearing Research*, *38*(1), 26–32.

Robieux, C., Meunier, C. (2015). *Phonetic considerations in vocal effort assessment.* Presented at the 11th Pan-European Voice Conference, Firenze, Italy.

Sundberg, J., & Gauffin, J. (1979). Waveform and spectrum of the glottal voice source. *Frontiers of speech communication research*, 301–322.

Titze, I. R. & Sundberg, J. (1992). Vocal intensity in speakers and singers. *The Journal of the Acoustical Society of America*, *91*(5), 2936–2946.

Traunmüller, H. & Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *The Journal of the Acoustical Society of America*, *107*(6), 3438–3451.

Van Bezooijen, R. (1995). Sociocultural aspects of pitch differences between Japanese and Dutch women. *Language and Speech*, *38*(3), 253–265.

Camargo, Z., Madureira, S., dos Reis, N., & Rilliard, A. (2019). The phonetic approach of voice qualities: challenges in corresponding perceptual to acoustic descriptions. In J. M. Lahoz-Bengoechea, & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 11–17). Málaga: Universidad de Málaga.

Subsidia: Tools and Resources for Speech Sciences

4

The analysis of facial and speech expressivity: tools and methods

Sandra Madureira[1] and Mario Augusto de Souza Fontes[1]

[1] Pontifical Catholic University of São Paulo
e-mail: madusali@pucsp.br, fonts@pucsp.br

Citation / Cómo citar este artículo: Madureira, S., Fontes, M. A. S. (2019). The analysis of facial and speech expressivity: tools and methods. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 19–26). Málaga: Universidad de Málaga.

ABSTRACT: The objective of this paper is to present an approach to investigate speech expressivity based on perceptual and acoustic analysis. To illustrate this approach, a video excerpt of the reciting of a poem by a professional actor is used. The research methodology proposed comprises the use of tools and perceptual analysis protocols for facial, voice quality and voice dynamics analysis, a semantic componential analysis questionnaire to evaluate speech expressive usages as well as a script for performing acoustic analysis. For the analysis of the facial gestures, Affectiva SDK is used as a tool and the analytical units are the facial action units. For the analysis of the vocal quality settings and the vocal dynamic features, the VPAS is used and for the acoustic analysis, the ExpressionEvaluator script. Correlations among facial, vocal, acoustic and semantic features are discussed and processed across multiple levels by means of multidimensional statistical analysis.

Keywords: facial expression; vocal quality settings; acoustic measures; perceptual analysis; speech expressivity.

RESUMEN: El objetivo de este artículo es presentar un método de investigación de la expresividad del habla basado en el análisis perceptivo y acústico. A modo de ilustración, se utiliza un fragmento de un vídeo de un actor profesional recitando un poema. La metodología propuesta incluye el uso de herramientas y de protocolos de análisis perceptivo para la expresión facial, la cualidad de voz y la dinámica de la voz, así como un cuestionario de análisis de componentes semánticos para evaluar los usos expresivos del habla y un script para realizar los análisis acústicos. Para el análisis de los gestos faciales, se utiliza la herramienta Affectiva SDK y las unidades de análisis son las unidades de acción facial. Para el análisis de los ajustes de la cualidad de voz y los rasgos de la dinámica vocal, se utiliza el VPAS y, para el análisis acústico, el script ExpressionEvaluator. Se comentan las correlaciones entre los rasgos faciales, vocales, acústicos y semánticos, y estos se procesan en múltiples niveles por medio de un análisis estadístico multidimensional.

Palabras clave: expresión facial; cualidad de voz; medidas acústicas; análisis perceptivo; expresividad del habla.

1. INTRODUCTION

Facial and vocal gestures are interpreted by participants in face-to-face spoken interactions and are used to convey linguistic, pragmatic and extralinguistic meanings.

On the one hand, the interactants rely on visual codes, such as the facial signal system that is used to express and recognize emotions and, on the other hand, on the sound symbolic codes which comprise the frequency code (Bolinger, 1986; Hinton, Nichols & Ohala, 1994; Morton, 1994; Ohala, 1997) as well as the effort, the respiratory and the sirenic codes (Gussenhoven, 2002, 2004, 2016) in order to perceive and produce meanings.

Facial and vocal gestures are highly communicative and they are fully integrated with vocal gestures in the expression of attitudes, modality and emotions in the oral discourse. In fact the facial nerve (VII cranial nerve), which innervates the facial muscles, is also connected to the brain motor areas responsible for speech production (Sanders, 2010).

Facial expressivity in relation to emotion expression has been largely studied since the seminal work by Ekman & Friesen (1971). Theoretical issues on emotion expression are discussed by Cornelius (1996), Fontes (2016), and Scherer (2005).

From the investigation of facial expressions of emotions, a perceptual protocol called the Facial Action Coding System (FACS) was derived. Facial analysis has been applied to marketing, clinical, educational and safety purposes. Understanding consumers' and learners' behaviors and detecting lies and pain cues are some of the issues which have been explored in these applications.

Speech Expressivity is a research topic which has received increasing attention in the phonetic literature in the last years: Barbosa (2009); Beller (2009); Madureira & Camargo (2010); Silva, Barbosa & Abelin (2016), to mention a few. This recognition has to do with the fact that speech is used not only to express meanings but also to impress listeners (Bolinger, 1986) and those two aspects are socially and communicatively relevant.

More recently, the investigation of the association between facial and vocal gestures has been focused in the phonetic literature (Hönemann, Mixdorff & Rilliard, 2014; Lu, Aubergé, Audibert, & Rilliard, 2014; Madureira, 2016; Rilliard, Erickson, Shochi, & Moraes, 2013).

This is not an easy enterprise since both facial and vocal gestures are complex to be analyzed and pose both methodological and theoretical issues to be investigated. The understanding of how linguistic prominence, modality and meaningful effects are conveyed, how speakers use vocal and facial gestures to express meanings, attitudes and emotions and how listeners attribute physical, affective, social, educational, regional, attitudinal, emotional and other characteristics based on vocal and facial gestures are highly relevant from both theoretical and applied perspectives.

The same facial movement or vocal gesture may be involved in the expression of prominence, modality, emotion and attitude.

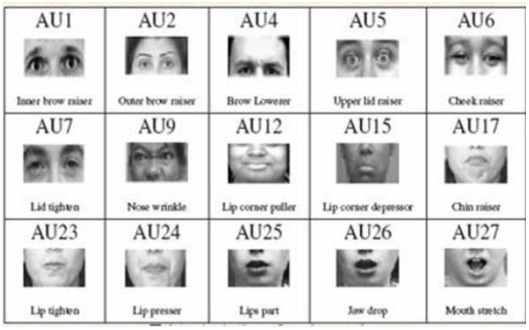
In order to tackle these kinds of issues related to the complexity of the expressive uses of face and vocal gestures, this paper aims to present a methodological approach to investigate speech expressivity based on perceptual and acoustic analysis. The kind of methodological procedures proposed in this paper can be applied to acted, spontaneous and semi-spontaneous speech. As tools, software for facial and speech analysis, perceptual protocols for facial and voice quality and voice dynamics analysis, a script for automatic extracting acoustic measures and a semantic questionnaire are considered.

2. FACIAL AND VOCAL GESTURES AND SPEECH EXPRESSIVITY

As facial, voice quality and vocal dynamics play a very important role in speech expressivity (Beller, Obin, & Rodet, 2008; Campbell & Mokhtari, 2003; Ekman & Friesen, 1971; Ekman, Friesen, & Ellsworth, 1972; Fonagy, 1983, 2001; Fontes & Madureira, 2015; Gobl & Ní Chasaide, 2003; Madureira, 2011; Scherer, 1984, 1986, 2005), protocols which provide means of analyzing them are worth resorting to.

In the following subsections two protocols are considered and compared. One of them is meant to analyze facial gestures and the other vocal quality gestures.

Figure 1: List of 17 AUs and related facial pictures. Extracted from Kanade, Cohn, & Tian (2000).



2.1. The facial gestures and FACS

For the analysis of facial gestures, as introduced earlier in this paper, the Facial Action Coding System, known as FACS, is a perceptual protocol which enables the description of facial movements and micro expressions which are very brief facial expressions lasting fractions of second. Facial expressions can be controlled or uncontrolled, but micro expressions are uncontrolled.

In interpreting emotions, not only the action unit movements and micro expressions are important to analyze but also aspects related to the assimetry of facial organs.

The analytic unit of FACS is the action unit (AU). In its last reviewed version (Ekman, Friesen, & Hager, 2002) 64 AUs are presented. In Figure 1, 15 of these AUs are presented.

The AUs are further categorized according to their intensity and their laterality (left or right). In interpreting emotions, not only the AU movements are important to analyze but also aspects related to the assimetry of facial organs.

The categorization of the AUs is made with reference to a neutral face image of the person analyzed. The neutral face is thought not to convey an emotion expression and not to present overt muscle movement related to any emotion expression. This concept of neutral face is however argued against by Adams Jr., Nelson, Soto, Hess, & Kleek (2012, p. 3) who defend that “most faces cannot aptly be described as emotionally neutral”.

Identifying AUs is a time-consuming task and expertise is hard to achieve. There are several automatic systems for facial and emotion analysis available to research purposes nowadays. The automatic systems perform real time analyses, while non-automatic analyses take 2 hours to analyze one-minute videos. They are helpful in identifying dynamic changes in the expressions of emotions over time.

Automatic analysis can be applied to large databases but recordings must attend some requirements in order to undergo such an analysis. Automatic analysis, for instance, fails in identifying facial expressions in non-frontal facial expressions or in bad lighting conditions. However, manual analysis of facial movements also fails in these conditions.

Madureira, S. & de Souza Fontes, M. A. (2019). The analysis of facial and speech expressivity: tools and methods. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 19–26). Málaga: Universidad de Málaga.

Lighting levels in Affedex vary from an RGB range of 0 (pitch black) to 255 (very bright). The threshold for accurate classifier performance in a RGB range from dark to very bright is 30 according to the research team working on Affedex, a face recognition system that analyzes face movements and correlates them to emotional and cognitive states.

2.2. The vocal gestures and VPAS

Among the protocols for voice quality analysis, the Voice Profile Analysis Scheme (VPAS) developed by Laver & Mackenzie-Beck (2007) has the advantage of being based on a componential descriptive phonetic model of voice quality analysis devised by Laver (1980).

Being componential, it allows the identification of shared and non-shared features in comparing speakers' vocal profiles. Being phonetically based it is comprehensive since articulatory and phonatory maneuvers can be mapped.

Perceived quality, anatomical and physiological factors, and acoustic measurement are theoretically linked in Laver's model of voice quality description (Laver, 2000; Mackenzie-Beck, 2005).

The analytical unit of Laver's phonetically based model is the setting, a long-term muscular adjustment of the vocal apparatus.

In the phonetic descriptive model of voice quality proposed by Laver (1980), 53 types of vocal quality settings and two principles governing them were introduced.

The *principle of susceptibility* accounts for the fact that some speech segments are more susceptible to the effects of some voice quality settings than others. Unrounded sounds, for example, are more susceptible to the spreading lips setting of voice quality than rounded sounds are. The *principle of compatibility* holds that some voice quality settings can co-occur while others can't. For example, a rounded setting cannot co-occur with a spreading lip setting.

The VPAS includes vocal tract features, overall muscular tension features, phonatory features, prosodic features, and temporal organization. The fact that labels in VPAS are phonetically grounded may be viewed as an advantage to identify vocal quality settings, provided that the analyst has solid phonetic knowledge.

To identify the vocal tract settings, it is necessary to have knowledge of the inherent phonetic characteristics of the speech segments of a given language so that the effects of articulatory and phonatory settings on segment production can be evaluated. The effects of the settings on certain kinds of segments, namely *key segments*, are more salient. Key segments play an important role in the description of voice quality settings because they are more susceptible to the effects produced by a given setting.

The identification of the settings is made in reference to a *neutral setting*. The neutral setting is characterized by a balanced mode of vocal fold vibration in terms of adduction forces and longitudinal tension, with no audible phonatory noise source, no constricted or expanded laryngeal and supralaryngeal vocal tract cavities, no shortening or lengthening adjustments of the vocal tract unless demanded by segment production restrictions, and moderate laryngeal and supralaryngeal tenseness.

Intra and inter-rater agreement tests are necessary to validate perceptual judgements.

2.3. Comparison between the FACS and the VPAS protocols

The FACS and the VPAS were created and revised about the same time. The former was created in 1976 and revised in 2002 and the latter was created in 1980 and revised in 2007.

Both protocols refer to a neutral element: the neutral face in FACS and the neutral voice quality setting in the VPAS.

Both protocols use scales. In the case of FACS, the intensity of the facial movements are classified with degrees varying from A to E and in the case of the VPAS the intensity of the setting is from 1 (weak) to 6 (very strong). However, FACS and VPAS differ in the way intensity of their analytical units are viewed, since AUs are thought to change their intensity levels from onset to peak to offset but that does not happen in relation to the vocal quality settings.

Both FACS and VPAS are componential and some of these components, facial movements in FACS and voice quality settings in VPAS, can be combined. One can infer from that possibility of component combination that both protocols are ruled by the same principle of compatibility.

Both protocols are also concerned with the characteristics of the speech segments since interference from their effect on judging the facial or the vocal quality settings is pinpointed. This is understood as the principle of susceptibility.

The two protocols share other features as well. They are perceptually oriented and theoretically based. FACS follows a Darwinian approach and VPAS a phonetic approach.

The similarities between the two protocols are thought to play a facilitation role in analyzing the association between facial and vocal gestures in speech expressivity studies.

3. METHOD

To illustrate the kind of methodological procedures proposed, an example analysis of a video excerpt is considered. Three kinds of expressive uses of facial and vocal gestures were considered in the analysis of that video excerpt: the production of semantically positive and negative adjectives; the expression of basic emotions; the expression of emotional primitives.

Madureira, S. & de Souza Fontes, M. A. (2019). The analysis of facial and speech expressivity: tools and methods. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 19–26). Málaga: Universidad de Málaga.

3.1. Corpus

The chosen corpus is a video excerpt of the reciting of the poem "A Valsa" (The Waltz) by a professional actor. This video can be watched on YouTube at the following address:
https://www.youtube.com/watch?v=le7UnUaxv90.

An audio recording is also available in a commercial CD entitled *Quatro Séculos de Poesia Brasileira*, which was released by Luz da Cidade Productions in 2002.

The poem has twelve stanzas characterized by a tertiary rhythm and was written by Casimiro de Abreu, a Brazilian poet, playwright and novelist whose works belong to the romantic movement of the nineteenth century. The poem narrates a couple's performance of a waltz. The narrator expresses his feelings (love, anger, jealousy, sadness, admiration) towards the girl he loves and her partner while watching them dancing a waltz.

An analysis of the rhythmic characteristics of the poem is found Moraes (1989) and the analysis of uses of sound symbolism in Madureira & Camargo (2010).

The choice of the corpus was motivated by two factors, one related to the semantic and pragmatic content of the poem and the other to the nature of the oral interpretation. Several emotive states are taken into account in the poem "The Waltz". There are reports of pragmatic situations involving anger, jealousy, sadness, contempt, love, deception, admiration, and joy. The actor is very skilled and known for the excellence of his interpretation. Interpreting, as he mentioned in an interview once, is about text meaning production and based on that vocal and body gestures come naturally. His interpretation of the poem is extremely moving.

3.2. Methodological procedures

A research methodology is proposed comprising the use of tools and perceptual analysis protocols for facial, voice quality and voice dynamics analysis and a script for the acoustic analysis. For the analysis of the facial gestures, the Affectiva SDK is used as a tool and the analytical units are the facial action units.

For the analysis of the vocal qualities and the vocal dynamic features the VPAS is used and for the acoustic analysis the ExpressionEvaluator script. Correlations among facial, vocal, acoustic and semantic features are discussed and processed across multiple levels by means of multidimensional statistical analysis.

3.2.1. The perceptual analysis of facial action

The analysis of the facial action units was performed automatically with the use of the Affectiva, a software from Affedex.

The voice quality settings were described by a phonetician with fifteen years of experience using the VPAS protocol.

In order to synchronize the data, the software Elan from Max Planck Institute was used.

3.2.2. Acoustic Analysis of Speech Expressive Data

The acoustic measures were automatically extracted by the ExpressionEvaluator script developed by Barbosa (2009) for Praat.

The script extracts 12 acoustic measures related to fundamental frequency, intensity, spectral tilt and Long Term Average Spectrum (LTAS). Measures involving f_0 include the median, inter-quartile semi-amplitude, skewness, and 0.995 quantile; additionally, several measures refer to the f_0 derivative: mean, standard deviation, and skewness. As for intensity, the skewness was measured. The spectral tilt measures were instantiated in mean, standard deviation, and skewness. Finally, the LTAS standard deviation was measured.

3.2.3. The componential analysis questionnaire

The componential analysis questionnaire contains semantic descriptors which are judged in scalar degrees varying usually from 1 to 5 or 1 to 7.

The descriptors can refer to psychological, social, economic, instructional, physical, or physiological characteristics, among other possibilities.

When expressive uses of facial and vocal gestures are analyzed, basic emotions such as anger and joy or emotional primitives can be considered. The emotional primitives are: valence (positive / negative); activation (excited / not excited) and dominance (strong / weak).

The componential analysis questionnaire can be applied to a group of judges. For the sake of statistics, a group of more than 30 is recommendable. The GTrace (McKeown, Valstar, Cowie, Pantic, & Schröder, 2012) is a helpful tool to be used in perceptual tests to present data to the judges since it allows the dynamic analysis of emotions over time based on video frames.

When facial expression of emotions is the focus of a research, one can resort to automatic analysis. The automatic analysis systems designed to analyze facial expressions, provide information about AUs, micro facial expressions, basic emotions, emotional primitives and other features over time such as smirk, smiley, relaxed and kissing.

3.2.4. Statistical procedures

In order to correlate the quantitative and qualitative variables, the statistical factorial method called Multiple Factor Analysis (MFA) is an adequate method. It can be applied using the FactorMInerR (Husson, Josse, Lê, & Mazet, 2013). Quantitative variables should be normalized by z-scores.

The use of MFA involves three steps: finding a common structure among the group variables; describing the specificity of each group of variables by means of correlation analysis; and comparing the resulting values by means of the individual analyses of the variables.

Figure 2: Pictures displaying facial gestures related to the productions of a semantically positive and a semantically negative adjective.

The Multiple Factor Analysis was used to study similarities among stimuli relative to the 29 research variables structured in the groups studied in this work: Gc1 (emotions), Gc2 Expression Evaluator measures and Gc3 VPAS vocal quality settings. All measures were normalized by z-scores. In order to verify similarity among the groups of variables the Pearson Lg coefficient was used.

4. RESULTS

The analysis of the productions of the semantically positive adjectives "contente" (content), serena (calm), bela (beautiful), formosa (pretty) and risonho (smiley)" showed that AU 12 (lip corners are pulled obliquely) and lip spreading vocal quality setting were involved. The production of the adjective "serena" involved lax voice quality setting and absence of facial tensing movements as well.

The production of the negative adjective "triste" involved AUs 1 (inner brower raiser), 4 (brow lowerer), 15 (lip corner depressor) and 43 (eyes closed). The production of the other semantically negative adjectives "louco" (crazy), "falsa" (false), "perjuros" (deceitful), involved some kind of lateral or downward movement of the head, face or one of its parts.

The downward head, face or eyes movements are usually related to the expression of negative emotions. The analogous downward movement in voice is low pitch and it is usually related to strength or negative emotions. The downward movement indicates symbolic uses of facial and vocal gestures.

Figure 2 shows two pictures: on the left side, his production of the semantically positive adjective "contente" and on the right side the production of the semantically negative adjective "triste". The two facial expressions are used accordingly to the semantic features of the lexical items, since "contente" refers to a positive quality and "triste" to a negative one.

The automatic analysis of the expression of discrete emotions by Affectiva SDK showed that contempt and disgust were predominant but other emotions were also identified: surprise, anger, sadness and fear. These findings are according to the situational context described in the poetic text.

Figure 3: Graphic from Affectiva SDK describing emotions based on facial expressions.

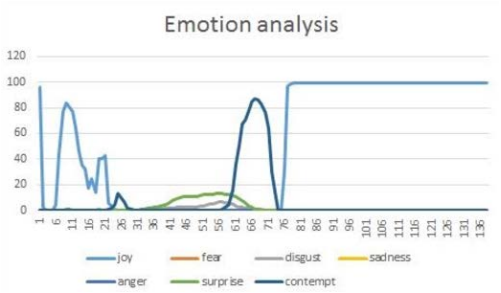
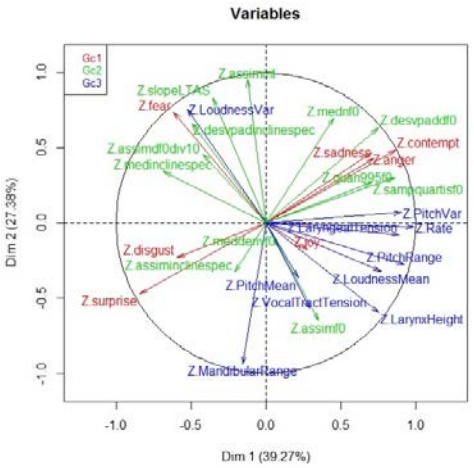


Figure 4: Acoustic, emotion and vocal quality variables.



In Figure 3, a graphic derived from Affectiva SDK is shown. It refers to the analysis of the speaker´s facial expression of emotions between a verse in which he praises the smile of the woman he loves and the following verse in which he complains she is smiling at someone else. In the vertical axis, the intensity of the emotion is plotted and in the horizontal axis the number of frames analysed by the system. The emotions are indicated by different colors. Happiness and contempt were detected in higher intensity levels. Surprise and disgust were detected but their intensities are very low.

The representativeness of the acoustic variables can be seen in the graphic displayed in Figure 4. The Z refers to the z-scores applied to the extracted measures.

In Table 1 the variables which were found to be significant are displayed. From that table, in dimension 1, it can be observed that there is correlation between the VPAS variables and contempt. The emotion "surprise" is inversely proportional to the variables contempt and the 0,995 quantile (quant995f0). In dimension 2, LTAS slope and intensity skewness (assimint) are correlated but they are inversely proportional to mandibular vocal quality setting.

Figure 5 displays the projection of the three groups of variables. The more projected in the vector space, the more representative is the variable. The most representative variable in the graphic of Figure 4 is the Gc2, that is the group of the acoustic measures extracted with the ExpressionEvaluator.

Table 1: Significant variables in the dimensions 1 and 2.

Dimension 1		
Variable	**Correlation**	**p-value**
Z.Rate	0.988	0.0002
Z.PitchRange	0.926	0.0081
Z.PitchVar	0.905	0.0132
Z.LaryngealTension	0.890	0.0175
Z.contempt	0.864	0.0265
Z.quan995f0	0.861	0.0276
Z.surprise	−0.841	0.0360
Dimension 2		
Variable	**Correlation**	**p-value**
Z.assimint	0.956	0.0029
Z.slopeLTAS	0.837	0.0377
Z.MandibularRange	−0.940	0.0052

Table 2: Pearson coefficient values.

Lg	**Gc1**	**Gc2**	**Gc3**	**MFA**
Gc1	1.3268	0.9477	0.8519	1.3058
Gc2	0.9477	1.893	1.0192	1.6121
Gc3	0.8519	1.0192	1.2067	1.2855
MFA	1.3058	1.6121	1.2855	1.7556

In Table 2, the Lg Pearson coefficients are provided. The higher the Lg coefficients the more related the variables.

The automatic analysis of the expression of emotional primitives in the poem by Affectiva showed that 33% of the utterances were positive and 67% negative. This finding is congruent with the semantic content of the poem, since jealousy, anger, sadness, contempt, deception are predominant.

5. CONCLUSION

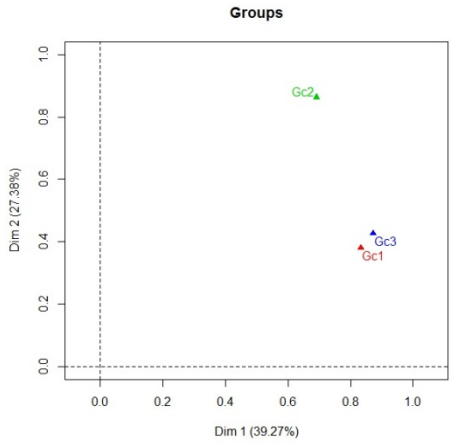
The methodological procedures associating tools, protocols, qualitative and quantitative data and statistical procedures introduced in this paper allows considering the analysis of visual and phonetic details in relation to linguistic, paralinguistic and extralinguistic variables. They are thought to enlighten the complexity of speech expressivity which is the core of oral communication.

The tools (ExpressionEvaluator, Affectiva SDK and Elan) as well as the perceptual protocols (FACS and VPAS) and the multidimensional statistical analysis (MFA) used to analyse the data in this paper proved to be extremely useful.

As far as the perceptual protocols are concerned, the fact that they share structural organizational features and principles is advantageous in terms of making the process of application of the protocols easier and correlating their analytical units, which are the vocal quality settings in the VPAS protocol and the action units in the FACS protocol.

Notwithstanding the value of automatic analysis of emotions, perceptual tests applied to judges to evaluate

Figure 5: Projection of the groups of variables.



emotions should be used either for contrastive or validation purposes. There are several tools to apply these tests, among them the GTrace or the perceptual scales we have mentioned earlier.

Getting to know visual and phonetic details is crucial to achieve a better understanding of how humans use codes to express and attribute meanings in speech production and perception or, in other words, to enhance knowledge of speech expressivity.

The non-verbal aspects, both visual and phonetic, involved in the process of oral communication can reinforce the semantic content, contradict the semantic content or add extra information.

These three kinds of occurrences were identified in the video analyzed in this paper. This interplay among visual, phonetic and semantic aspects can be traced by associating acoustic measures, perceptual descriptions and meaning effects.

In order to investigate speech expressivity both qualitative and quantitative variables must be taken into account. It is a time-consuming task even with the help of automatic tools, but it opens a path which is worth exploring.

Facial movements and the speech organ movements are gestures and gestures are interpreted as indices by listeners and used by speakers to convey linguistic, paralinguistic and extralinguistic meanings.

6. REFERENCES

Adams, Jr. R. B., Nelson, A. J., Soto, J. A., Hess, U., & Kleck, R. E. (2012). Emotion in the neutral face: A mechanism for impression formation?, *Cognition and Emotion, 26*(3): 431–441.

Barbosa, P. A. (2009). Detecting changes in speech expressiveness in participants of a radio program. *Proceedings of Interspeech* (pp. 2155–2158).

Beller, G. (2009). Transformation of expressivity in speech. In P. Lang (Ed.), *The role of prosody in the expression of emotions in English and in French*. Peter Lang Publishing Group.
Acessed on February, 8, 2016, retrieved from http://articles.ircam.fr/textes/Beller09c/index.pdf

Beller, G., Obin, N., & Rodet, X. (2008). Articulation degree as a prosodic dimension of expressive speech. In P. Barbosa, S. Madureira, & C. Reis, (Eds.), *Proceedings of the Fourth International Conference on Speech Prosody*.

Bolinger, D. (1986). *Intonation and its parts: Melody in spoken English*. London: Edward Arnold.

Campbell, N. & Mokhtari, P. (2003). Voice quality: the 4th prosodic dimension. *Proceedings of the XVth International Congress of Phonetic Sciences* (pp. 2417–2420).

Cornelius, R. R. (1996). *The science of emotion. Research and tradition in the psychology of emotion*. Upper Saddle River, NJ: Prentice-Hall.

Ekman, P. & Friesen, W. V. (1971). Constants across culture in the face and emotion. *Journal of Personality and Social Psychology, 17*, 124–129.

Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). *Emotion in the human face: Guidelines for research and an integration of findings*. New York: Pergamon Press.

Ekman, P., Friesen, W. V., & Hager, J. C. (Eds.) (2002). *Facial Action Coding System* (2nd ed.). Salt Lake City, UT: Research Nexus eBook.

Fónagy, I. (1983). *La vive voix: Essais de psycho-phonétique*. Paris: Payot.

Fónagy, I. (2001). *Languages within Language: an evolutive approach*. Amsterdam: John Benjamins.

Fontes, M. A. S. (2016). Os papéis das prosódias vocal e visual na expressão de emoções na fala. In S. Madureira (Org.), *Sonoridades—Sonorities*. São Paulo: Edição da Pontifícia Universidade Católica de São Paulo.

Fontes, M. A. S. & Madureira, S. (2015). Gestural prosody and the expression of emotions: A perceptual and acoustic experiment. In *Proceedings of the 18th International Congress of Phonetic Sciences* (paper number 0390). Retrieved from http://www.icphs2015.info/pdfs/Papers/ICPHS0390.pdf

Gobl, C. & Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication, 40*(1–2), 189–212.

Gussenhoven, C. (2002). Intonation and interpretation: phonetics and phonology. *Proceedings of the 1st International Conference on Speech Prosody* (pp. 47–57).

Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge: Cambridge University Press.

Gussenhoven, C. (2016). Foundations of intonational meaning: Anatomical and physiological factors. *Topics in Cognitive Science, 8*, 425–434.

Hinton, L., Nichols, J., & Ohala, J. J. (Eds.) (1994). *Sound symbolism.* Cambridge: Cambridge University Press.

Hönemann, A., Mixdorff, H., & Rilliard, A. (2014). Social attitudes: Recordings and evaluation of an audio-visual corpus in German. *Proceedings of the 7th Forum Acusticum*.

Husson, F., Josse, J., Lê, S., & Mazet, J. (2013). *FactoMineR: Multivariate exploratory data analysis and data mining with R* (version 1.25) [R package].

Kanade, T., Cohn, J. F., & Tian, Y. (2000). Comprehensive database for facial expression analysis. *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 46–53).

Laver, J. (1980). *The phonetic description of voice quality*. Cambridge: Cambridge University Press.

Laver, J. (2000). Phonetic evaluation of voice quality. In R. D. Kent & M. J. Ball (Eds.), *Voice quality measurement* (pp. 37–48). San Diego: Singular Thomson Learning.

Laver, J. & Mackenzie-Beck, J. (2007). *Vocal Profile Analysis Scheme—VPAS*. Edinburgh: Queen Margareth University College.

Lu, Y., Aubergé, V., Audibert, N., & Rilliard, A. (2014). Audiovisual perception of expressions of Mandarin Chinese social affects by French L2 learners. *Proceedings of the 7th International Conference on Speech Prosody* (pp. 169–173).

Mackenzie-Beck, J. (2005). Perceptual analysis of voice quality: the place of the Vocal Profile Analysis. In W. J. Hardcastle J. & Mackenzie-Beck (Eds.), *A figure of speech: A festschrift for John Laver* (pp. 285–322). Mahwah, NJ: Lawrence Erlbaum.

Madureira, S. (2011). The investigation of speech expressivity. In H. Mello, A. Panunzi, & T. Raso (Eds.), *Illocution, modality, attitude, information patterning and speech annotation* (pp. 101–118). Firenze: Firenze University Press.

Madureira, S. (Org.) (2016). *Sonoridades—Sonorities*. São Paulo: Edição da Pontifícia Universidade Católica de São Paulo.

Madureira, S. & Camargo, Z. A. (2010). Exploring sound symbolism in the investigation of speech expressivity. *Proceedings of ISCA* (pp. 105–108).

McKeown, G., Valstar, M., Cowie, R., Pantic, M. & Schröder, M. (2012). The SEMAINE database: annotated multimodal records of emotionally coloured conversations between a person and a limited agent. *IEEE Transactions of Affective Computing, 3*, 165–183.

Moraes, J. A. de (1989). Sobre o ritmo da poesia em língua portuguesa. Análise acústica das marcas prosódicas de fim de verso. In R. Lorenzo (Org.), *Separata das Actas do XIX Congreso Internacional de Lingüística e Filoloxia Românica* (pp. 1017–1026).

Morton, E. S. (1994). Sound symbolism and its role in non-human vertebrates. In L. Hinton & J. J. Ohala (Eds.), *Sound symbolism* (pp. 348–365), Cambridge: Cambridge University Press.

Ohala, J. J. (1997). Sound symbolism. *Proceedings of the 4th Seoul International Conference on Linguistics (SICOL)* (pp. 98–103).

Madureira, S. & de Souza Fontes, M. A. (2019). The analysis of facial and speech expressivity: tools and methods. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 19–26). Málaga: Universidad de Málaga.

Rilliard, A., Erickson, D., Shochi, T., & Moraes, J. A. de (2013). Social face to face communication: American English attitudinal prosody. *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH)* (pp. 1648–1652).

Sanders, R. D. (2010). The trigeminal (V) and facial (VII) cranial nerves: Head and face sensation and movement. *Psychiatry (Edgmont), 7*(1), 13–16.

Scherer, K. R. (1984). On the nature and function of emotion: a component process approach. In K. R. Scherer & P. Ekman (Eds.), *Approaches to Emotion* (pp. 293–318). Hillsdale, NJ: Lawrence Erlbaum.

Scherer, K. R. (1986). Vocal affect expression: a review and a model for future research. *Psychological Bulletin, 99*(2), 143–165.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information, 44*(4), 693–727.

Silva, W. da, Barbosa, P. A., & Abelin, A. (2016). Cross-cultural and cross-linguistic perception of authentic emotions through speech: An acoustic-phonetic study with Brazilian and Swedish listeners. *DELTA: Documentação de Estudos em Linguística Teórica e Aplicada, 32*, 449–480.

Subsidia: Tools and Resources for Speech Sciences

TransText, un transcriptor fonético automático de libre distribución para español y catalán

Juan María Garrido[1], Marta Codina[2] y Kimber Fodge[2]

[1] Universitat de Barcelona
[2] Universitat Pompeu Fabra
e-mail: juanmaria.garrido@ub.edu

Citation / Cómo citar esta publicación: Garrido, J.M., Codina, M. y Fodge, K. (2019). TransText, un transcriptor fonético automático de libre distribución para español y catalán. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 27–31). Málaga: Universidad de Málaga.

RESUMEN: En este trabajo se describe TransText, una herramienta de libre distribución para la transcripción fonética automática en español y catalán. La herramienta permite transcribir textos tanto en la variedad estándar de estas dos lenguas como en una serie de variedades habladas en diferentes áreas de España: Andalucía occidental, Andalucía oriental, Extremadura norte, Extremadura sur, Canarias, Castilla-La Mancha, Madrid y Murcia, en el caso del español; y Ribagorza, Pallars, Tortosa, zona occidental central, Valencia septentrional, Valencia central, Valencia meridional y Alicante, en el caso del catalán. Se describen los principales módulos de la herramientas y se explica brevemente el desarrollo de las reglas de transcripción fonética para las diferentes variantes. También se presentan los resultados de dos evaluaciones, una para el catalán y otra para el español, llevadas a cabo para comprobar la robustez de las reglas desarrolladas.

Palabras clave: transcripción fonética automática; español; catalán; dialectos**.**

ABSTRACT: This paper presents TransText, a free software tool for the automatic phonetic transcription of Spanish and Catalan. The tool allows users to transcribe both the standard pronunciations of both languages, and reproduces the most prototypical pronunciations of different areas of Spain: Western Andalucía, Eastern Andalucía, Northern Extremadura, Southern Extremadura, the Canary Islands, Castilla-La Mancha, Madrid and Murcia, in the case of Spanish; and Ribagorza, Pallars, Tortosa, Central Western area, Northern Valencia, Central Valencia, Southern Valencia and Alicante, in the case of Catalan. The main modules of the tool are described, and the development process of the phonetic transcription tools for the different variants is outlined. The results of two evaluations, one for Catalan and one for Spanish, that were carried out to check the robustness of the developed rules, are presented.

Keywords: automatic phonetic transcription; Spanish; Catalan; dialects.

1. INTRODUCCIÓN

La transcripción fonética automática de textos es una tarea con múltiples aplicaciones, a caballo entre el procesamiento del texto y del habla. Todos los sistemas de conversión texto-habla, por ejemplo, incluyen un transcriptor fonético, encargado de determinar la cadena de sonidos que deberá sintetizarse a partir del texto de entrada. También puede tener aplicaciones en reconocimiento de habla, analítica del habla o enseñanza de lenguas, entre otros campos.

Los primeros transcriptores fonéticos desarrollados se diseñaron para transcribir en la variedad estándar de la lengua en cuestión. Un transcriptor fonético que permitiera reproducir la pronunciación estándar de una lengua era suficiente para determinadas aplicaciones, como la conversión texto-habla, pero no para otras que requieren tener en cuenta las diferentes posibilidades de pronunciación de una palabra o sonido, como es el caso del reconocimiento automático del habla. Se hizo necesario, por tanto, incorporar esta variación al proceso de transcripción fonética.

En la mayoría de los casos, los transcriptores fonéticos existentes para el español y el catalán han sido desarrollados específicamente para un sistema o aplicación concreta, normalmente comercial, y no están disponibles para uso público (Bonaventura, Giuliani, Garrido y Ortín, 1998, por ejemplo). La mayoría de los transcriptores no comerciales que pueden encontrarse para estos dos idiomas ofrecen limitaciones a su uso (por ejemplo, porque no permiten la transcripción de elementos que no sean palabras, como números o símbolos, solo manejan un alfabeto fonético —AFI o SAMPA— o solo permiten la transcripción a través de

una página web), su base lingüística está poco contrastada, ni tampoco permiten la transcripción en diferentes dialectos (López, 2004; Molino de Ideas, 2012). En el caso del español, Saga es una de las pocas herramientas de transcripción fonética de libre distribución sin restricciones de uso, que permite la transcripción fonética de texto en español estándar y en diferentes variedades del español de España y de Hispanoamérica y cuya transcripción fonética es el resultado de un trabajo previo conjunto entre lingüistas e ingenieros (Llisterri y Mariño, 1993; Moreno y Mariño, 1998). Para el catalán, Segre también es una herramienta sin limitaciones de uso que permite la transcripción en la variedad estándar y en cuatro variedades geográficas distintas (catalán oriental, catalán occidental, balear y valenciano), y que se desarrolló de forma conjunta por un equipo de ingenieros y lingüistas (Pachès *et al.*, 2000).

En este trabajo se presenta TransText, un transcriptor fonético automático de textos para castellano y catalán de libre distribución, que permite tanto la transcripción en español y catalán estándar como en distintas variedades geográficas habladas en España. Ha sido desarrollado a partir de TexAFon (Garrido, Laplaza, Marquina, Schoenfelder y Rustullet, 2012; Garrido, Laplaza, Kolz y Cornudella, 2014), un sistema completo de procesado lingüístico de texto basado en reglas para conversión texto-habla, que incluye también un módulo de transcripción fonética, al que se le han realizado diferentes mejoras para permitir la transcripción fonética en variedades distintas de la estándar. En los siguientes apartados se presenta su estructura general y sus principales funcionalidades, y se explica de forma breve la estrategia seguida para implementar en la herramienta la información sobre las diferentes variedades.

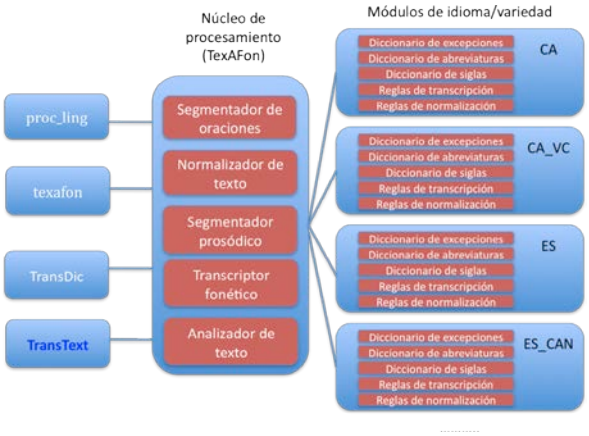
2. DESCRIPCIÓN GENERAL

TransText es una herramienta multiplataforma, que puede utilizarse tanto en Linux como en Mac OS o Windows, para lo cual es necesario descargar la versión adecuada en cada caso. Se ejecuta desde línea de comandos, y requiere especificar una serie de argumentos en el momento de la ejecución (por ejemplo, la lengua o variedad en que se quiere realizar la transcripción o el alfabeto en que se quiere obtener la transcripción). Este modo de funcionamiento permite utilizar la herramienta tanto de forma aislada como integrada en otras herramientas o procesos que impliquen llevar a cabo la transcripción fonética de textos.

La estructura interna de TransText, tal como se observa en la Figura 1, incluye tres capas o niveles:

- La aplicación en sí misma, TransText.
- El núcleo de procesamiento, común a otras aplicaciones (como TransDic, para la generación de diccionarios fonetizados, o textafon, orientada al procesado de texto en conversión texto-habla), que incluye, además del módulo de transcripción fonética, otros módulos de procesamiento del texto, algunos de los cuales, como el de preprocesamiento del texto, son utilizados también por TransText.
- Los módulos de idioma o dialecto, que incluyen los diccionarios y reglas específicos de cada variedad en que se puede transcribir. De acuerdo con esta estructura, cada variedad, incluida la estándar, es considerada como un 'idioma' autónomo, con sus propias reglas y diccionarios. La incorporación de una nueva variedad implica la creación de un nuevo módulo, que habitualmente se construye a partir de otro ya existente (el estándar normalmente).

Figura 1: Estructura de TransText.



El proceso de transcripción en TransText implica cuatro fases fundamentales:
(1) Segmentación del texto en oraciones
(2) Preproceso del texto
(3) Predicción del acento léxico
(4) Transcripción fonética

2.1. Entrada

La entrada de TransText es un fichero de texto sin formato, codificado en UTF-8, que contenga el texto que se quiera transcribir. También debe especificarse la variedad en que se quiere realizar la transcripción fonética. Actualmente TransText permite transcribir en las siguientes variedades del español y el catalán:

- Catalán: estándar (ca), Ribagorza (ca_ri), Pallars (ca_pa), Tortosa (ca_to), occidental central (ca_ac), Valencia septentrional (ca_vs), Valencia central (ca_vc), Valencia meridional (ca_vm), Alicante (ca_al).
- Español: estándar peninsular (es), Andalucía occidental (es_aoc), Andalucía oriental (es_aor), Extremadura norte (es_exn), Extremadura sur (es_exs), Canarias (es_can), Castilla-La Mancha (es_clm), Madrid (es_mad) y Murcia (es_mur).

2.2. Segmentación en oraciones

El módulo encargado de esta tarea es independiente del idioma, y básicamente se encarga de detectar en el texto signos de puntuación que indiquen final de oración. Los casos ambiguos, como los de los puntos que aparecen al final de abreviaturas o siglas, o los de las direcciones

Garrido, J.M., Codina, M. y Fodge, K. (2019). TransText, un transcriptor fonético automático de libre distribución para español y catalán. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 27–31). Málaga: Universidad de Málaga.

electrónicas, no son tratados como marcas de final de oración. La salida de este módulo es una lista de oraciones, que serán procesadas por separado por los siguientes módulos.

2.3. Preproceso

El módulo de preproceso (también llamado normalización) contiene reglas específicas para cada idioma, y se encarga de convertir todos los elementos del texto que no constituyen propiamente palabras (direcciones de correo electrónico, cifras, símbolos, etc.) en su equivalente en palabras. Esta tarea se lleva a cabo en dos fases:

- **Etiquetado**: se asigna a cada elemento (cadena de caracteres entre dos espacios en blanco) del texto una etiqueta de categoría (‘palabra’, ‘fecha’, ‘hora’, ‘dirección electrónica’, ‘número entero’, etc.). Así, por ejemplo, el elemento ‘11/06/2012’ recibiría la etiqueta ‘fecha’, ‘1-1’ la etiqueta ‘cifras y signos’, y ‘22@’ la etiqueta ‘cadena con símbolos’.
- **Expansión**: los elementos con una etiqueta distinta de ‘palabra’ se sustituyen por su equivalente en palabras, aplicando las reglas de expansión correspondientes a la etiqueta asignada. En el caso de los ejemplos anteriores, ‘11/06/2012’ se sustituiría por ‘once de junio de dos mil doce’, ‘1-1’ por ‘uno uno’, y ‘22@’ por ‘veintidós arroba’.

La aplicación de este módulo antes de llevar a cabo la transcripción fonética propiamente dicha asegura que cualquier elemento del texto de entrada aparecerá correctamente transcrito a la salida.

En la versión actual de TransText, las reglas de preproceso son comunes para todas las variantes de un mismo idioma (español o catalán), aunque la arquitectura del sistema permitiría, si fuera necesario, la inclusión de reglas específicas para una variante determinada.

2.4. Predicción de acento léxico, silabificación y transcripción fonética

Las tareas de predicción del acento léxico, silabificación y transcripción fonética son las últimas de todo el proceso. Para llevarlas a cabo se utilizan un diccionario de excepciones de pronunciación y dos conjuntos de reglas fonéticas, uno para la transcripción de palabras aisladas y otra para la transcripción de los procesos que se producen entre palabras. El proceso se realiza en dos fases:

- **Transcripción palabra a palabra:**
  Se comprueba en primer lugar si la palabra procesada está en el diccionario de excepciones de pronunciación. Si es así, se toma directamente del diccionario la transcripción fonética (que incluye ya la marca de acento léxico, si es una palabra tónica). En caso contrario, se aplican las reglas de predicción de acento y de transcripción fonética para generar la transcripción de la palabra. Finalmente, con independencia de cómo se haya obtenido la transcripción (diccionario o reglas), se aplican las reglas de silabificación para introducir en la transcripción las marcas de sílaba.
- **Transcripción en el ámbito del grupo fónico:**
  La salida de la etapa anterior (una serie de cadenas de palabras transcritas y silabificadas, y con marcas de acento léxico, cada una correspondiente a un grupo fónico) se procesan para detectar y tratar adecuadamente los casos en que la transcripción fonética en los límites de palabra deba modificarse debido al contexto. Para ello, se aplica un segundo conjunto de reglas, denominadas de ‘fonética sintáctica’. En el caso del español estándar, estas reglas cubren fenómenos como la asimilación del punto de articulación de las consonantes nasales finales de palabra (‘en Madrid’, [emmaˈðɾið]), o la realización como aproximante de las oclusivas iniciales de palabra que aparecen en contexto intervocálico (‘a Barcelona’, [aβaɾθeˈlona]).

Las primeras versiones de TexAFon (Garrido *et al.*, 2012; Garrido *et al.*, 2014) incluían solo módulos de idioma solo para español y catalán estándar. TransText, que sí permite la transcripción fonética en diferentes variedades de estos dos idiomas, ha podido desarrollarse gracias a diferentes mejoras realizadas en el núcleo de procesamiento de TexAFon, al que se han añadido nuevos módulos de idioma, tantos como variantes consideradas. Cada módulo de idioma contiene el diccionario de excepciones de pronunciación y las reglas de transcripción fonética (palabra a palabra y de fonética sintáctica) específicas de la variante correspondiente. La descripción detallada del proceso de desarrollo de estas reglas excede el ámbito de este trabajo, pero puede resumirse, para cada una de las variantes consideradas, en los siguientes pasos:

(1) Definición de los fenómenos fonéticos que, de acuerdo con la bibliografía existente, están lo suficientemente extendidos en el área geográfica donde se habla la variante que se está desarrollando como para considerarse generales. Por ejemplo, el caso del canario, el seseo, la aspiración de las <s> final de sílaba y la aspiración del sonido [x] del castellano son tres fenómenos que pueden considerarse generales en las islas, por lo que se incluyeron para su implementación.
(2) Implementación de las reglas fonéticas que cubriesen los fenómenos seleccionados, e inclusión de las excepciones de pronunciación necesarias en diccionario correspondiente. Las reglas están directamente implementadas en Python, y tienen el formato de reglas dependientes del contexto, tal como se observa en la figura. Un ejemplo de regla para el español puede encontrarse en la Figura 2.
(3) Evaluación de la transcripción obtenida y corrección de errores en caso necesario. Este proceso se realizó tantas veces como fue necesario.

Una descripción detallada del proceso de desarrollo de los módulos para las distintas variantes del español y del catalán consideradas en TransText puede encontrarse en Fodge (2014) y Codina (2016), respectivamente.

Garrido, J.M., Codina, M. y Fodge, K. (2019). TransText, un transcriptor fonético automático de libre distribución para español y catalán. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 27–31). Málaga: Universidad de Málaga.

Figura 2: Ejemplo de implementación en Python de una regla de transcripción fonética para el andaluz del conjunto de reglas palabra a palabra (Fodge, 2014). El fenómeno implementado es la aspiración de <s> en final de palabra.

```
        # UPDATE Aspiration of word-final /s/ followed by C; lohh gatos
        # s-> hh (rule finished in fonetica_sintactica)
        if ch=="s" and nch=="NIL":
                salida.append(["S",0,False])
```

2.5. Salida

La salida del programa es otro fichero de texto que contiene la transcripción del texto de entrada en alfabeto IPA o SAMPA, según se haya especificado en los parámetros de entrada. También puede especificarse mediante el correspondiente parámetro si se desea que la transcripción de salida esté silabificada o no, o con los límites de palabra marcados. La Tabla 1 y la Tabla 2 presentan dos ejemplos de salida en español y catalán, tanto utilizando el alfabeto IPA como el SAMPA.

3. EVALUACIÓN

La evaluación de los módulos desarrollados para las variantes del español, descrita en Fodge (2014), se llevó a cabo utilizando como material de base un listado de 307 palabras aisladas, representativas de los fenómenos implementados. También se diseñó un párrafo específico (579 palabras) para la evaluación de las reglas de fonética sintáctica y del funcionamiento general de la herramienta con texto. La lista y el párrafo fue procesado con cada conjunto de reglas para conseguir la correspondiente transcripción de salida, que fue revisada manualmente para detectar los errores de transcripción. Los resultados obtenidos mostraron que las reglas implementadas transcribían correctamente todas las palabras contenidas en la lista en las distintas variantes, y solo en el caso de las reglas de fonética sintáctica se detectó algún error aislado. Pero el resultado global cabe calificarlo de bueno.

El funcionamiento de las reglas para las diferentes variantes del catalán fue evaluado utilizando dos listas diferentes de palabras aisladas (99 palabras en total) y dos párrafos (511 palabras en total) escritos en catalán estándar (Codina, 2016). De nuevo, el corpus de evaluación fue procesado con cada conjunto de reglas para conseguir la correspondiente transcripción de salida, que fue revisada manualmente para detectar los errores de transcripción. A partir del número de errores para cada variante se calculó un coeficiente de error, aplicando el procedimiento descrito en Van Bael, Boves, Van den Heuvel y Strik (2007). La Tabla 3 presenta los resultados obtenidos, en los que se observa que en todos los casos se obtuvieron coeficientes de error bastante similares, con la excepción del ribagorzano, pero en general bajos, lo que indica un buen funcionamiento en general de las reglas de transcripción.

Tabla 1: Ejemplo de transcripción obtenida para un texto de muestra en español.

Texto de entrada	Gemma Baltasar Roura se licenció en Filología Hispánica por la Universidad de Barcelona en 1998. Tras especializarse como profesora de E/LE en International House, trabajó en diferentes academias privadas de Barcelona.
Transcripción IPA	x ˈe . m a # B a l . t a . s ˈaɾ # r ˈo w . ɾ a # s e # l i . θ e n . θ j ˈo # e n # f i . l o . l o . x ˈi . a # i s . p ˈa . n i . k a # p oɾ # l a # u . n i . β eɾ . s i . ð ˈa d # ð e # b aɾ . θ e . l ˈo . n a # e n # m ˈi l # n o . β e . θ j ˈe n . t o s # n o . β e n . t a # i # ˈo . ʧ o ‖ t ɾ a s # e s . p e . θ j a . l i . θ ˈaɾ . s e # k ˈo . m o # p ɾ o . f e . s ˈo . ɾ a # ð e # e	ˈe . l e # e # e n # i n . teɾ . n a . t j o . n ˈa l # x ˈa w s	t ɾ a . β a . x ˈo # e n # ð i . f e . ɾ ˈe n . t e s # a . k a . ð ˈe . m j a s # p ɾ i . β ˈa . ð a s # ð e # b aɾ . θ e . l ˈo . n a ‖
Transcripción SAMPA	x "e . m a # B a l . t a . s "a r # rr "o w . r a # s e # l i . T e n . T j "o # e n # f i . l o . l o . x "i . a # i s . p "a . n i . k a # p o r # l a # u . n i . B e r . s i . D "a d # D e # B a r . T e . l "o . n a # e n # m "i l # n o . B e . T j "e n . t o s # n o . B e n . t a # i # "o . tS o ... t r a s # e s . p e . T j a . l i . T "a r . s e # k "o . m o # p r o . f e . s "o . r a # D e # e ... "e . l e # e # e n # i n . t e r . n a . t j o . n "a l # x "a w s ... t r a . B a . x "o # e n # D i . f e . r "e n . t e s # a . k a . D "e . m j a s # p r i . B "a . D a s # D e # B a r . T e . l "o . n a ...

Tabla 2: Ejemplo de transcripción obtenida para un texto de muestra en catalán.

Texto de entrada	Per raons tècniques, aquest programa no es va emetre ahir, com s'havia anunciat. En el seu lloc es va tornar a emetre el programa anterior. Els demanem disculpes.
Transcripción IPA	p ə ɾ # r ə . ˈo n s # t ˈɛ k . n i . k ə s | ə . k ˈɛ t # p ɾ u . ɣ ɾ ˈa . m ə # n ˈo # ə s # p ˈa # ə . m ˈɛ . t ɾ ə # ə ˈi | k ˈɔ m # s ə . β ˈi . ə # ə . n u n . s i . ˈa t ‖ ə n # ə l # s ˈe w # ʎ ˈɔ k # ə s # p ˈa # t u ɾ . n ˈa # ə # ə . m ˈɛ . t ɾ ə # ə l # p ɾ u . ɣ ɾ ˈa . m ə # ə n . t ə . ɾ i . ˈo ɾ ‖ ə l s # t ə . m ə . n ˈɛ m # d i s . k ˈu l . p ə s ‖
Transcripción SAMPA	p @ r # rr @ . "o n s # t "E k . n i . k @ s ... @ . k "E t # p r u . G r "a . m @ # n "o # @ s # p "a # @ . m "E . t r @ # @ "i ... k "O m # s @ . B "i . @ # @ . n u n . s i . "a t ... @ n # @ l # s "e w # L "O k # @ s # p "a # t u r . n "a # @ # @ . m "E . t r @ # @ l # p r u . G r "a . m @ # @ n . t @ . r i . "o r ... @ l s # t @ . m @ . n "E m # d i s . k "u l . p @ s ...

Tabla 3: Coeficientes medios de error obtenidos en la evaluación de los módulos de transcripción fonética de las variantes del catalán implementadas.

Dialecto	Resultado
Ribagorzano	6,9
Pallarés	3,8
Tortosino	2,9
Área central	2,8
Valenciano central	3,3
Valenciano meridional	3,6
Alicantino	3,8

Cabe señalar, de todas formas, que ambas evaluaciones no se llevaron a cabo utilizando TransText, sino otra herramienta que emplea igualmente TexAFon como núcleo de procesamiento. La evaluación sistemática de TransText y la corrección de posibles errores de funcionamiento específicos y con un corpus de textos más amplio es todavía una tarea pendiente. De todas formas, los resultados presentados aquí parecen

indicar que la herramienta ofrece un resultado de una calidad aceptable.

4. CONCLUSIONES Y MEJORAS FUTURAS

En este trabajo se ha presentado TransText, un transcriptor fonético para el español y el catalán que permite transcribir, además de en la variante estándar, en diferentes variantes de estas dos lenguas habladas en España. Desarrollado para ser distribuido gratuitamente, puede descargarse desde la URL https://sites.google.com/site/juanmariagarrido/research/resources/tools/transtext.

TransText está concebido como una herramienta para la fonetización de textos de uso general, que puede utilizarse en aplicaciones de tecnología del habla, pero también en otros campos, como la enseñanza de idiomas o la transcripción fonética de corpus. Para la fonetización de diccionarios, está disponible otra herramienta, TransDic (https://sites.google.com/site/juanmariagarrido/research/resources/tools/transdic), que permite la generación automática de diccionarios fonetizados, y permite la generación de variantes de pronunciación para una misma palabra, una funcionalidad de especial utilidad para el desarrollo de diccionarios para el reconocimiento de habla. El principal interés de TransText y TransDic reside en que se trata de herramientas de libre distribución, que puede instalarse en un ordenador personal para su uso sin restricciones, frente a la mayoría de soluciones disponibles (en páginas web, normalmente), que imponen restricciones a su uso.

Frente a otras herramientas de libre distribución similares, como Saga y Segre, TransText presenta también una serie de ventajas. Por ejemplo, TransText permite la transcripción en variedades del español y el catalán que Saga y Segre no contemplan (Saga, por ejemplo, está pensado solo para la transcripción de variantes de América, e incluye algunos fenómenos fonéticos relevantes tanto en América como en España que Saga no tiene en cuenta). Además, permite especificar mediante un solo argumento la variedad en que se quiere realizar la transcripción, en lugar de especificar por medio de diferentes argumentos, como se hace en Saga, los fenómenos fonéticos no estándar que se tendrán en cuenta para la transcripción del texto de entrada. De esta manera un usuario no experto puede transcribir un texto sin conocer específicamente cuáles son los fenómenos fonéticos relevantes en esa variante. Finalmente, TransText permite seleccionar también el alfabeto fonético en que se obtendrá la transcripción (AFI o SAMPA).

Sin duda, la principal tarea pendiente es llevar a cabo una evaluación sistemática de la herramienta, para comprobar su robustez y su funcionamiento con grandes cantidades de texto real. También sería conveniente que dialectólogos expertos en las variantes implementadas evaluaran la transcripción que se genera actualmente, para determinar hasta qué punto es la más adecuada.

5. REFERENCIAS

Bonaventura, P., Giuliani, F. Garrido, J. M. y Ortín, I. (1998). Grapheme-to-phoneme transcription rules for Spanish, with application to automatic speech recognition and synthesis. En *Proceedings of the Workshop 'Partially Automated Techniques Transcribing Naturally Occurring Continuous Speech', 16th August 1998, Université de Montréal, Montreal, Quebec, Canada, Coling-ACL'98*, pp. 33–39.

Codina, M, (2016). *Automatic Phonetic Transcription of dialectal variance in Catalan* (Trabajo de fin de máster). Barcelona: Universitat Pompeu Fabra.

Fodge, K, (2014). *Introducing Spanish dialects in a linguistic processing module for improved ASR and novel speech synthesis capabilities* (Trabajo de fin de máster). Barcelona: Universitat Pompeu Fabra.

Garrido, J. M., Laplaza, Y., Marquina, M., Schoenfelder, C. y Rustullet, S. (2012). TexAFon: a multilingual text processing tool for text-to- speech applications. En *Proceedings of IberSpeech 2012, Madrid, Spain, November 21-23, 2012*, pp. 281–289.

Garrido, J. M., Laplaza, Y., Kolz, B. y Cornudella, M. (2014). TexAFon 2.0: A text processing tool for the generation of expressive speech in TTS applications. En *Proceedings of LREC 2014, Ninth International Conference on Language Resources and Evaluation, Reykjavik (Iceland), 26-31 May 2014*.

Llisterri, J. y Mariño, J. B. (1993). Spanish adaptation of SAMPA and automatic phonetic transcription, Esprit Project 6819. Report SAM-A/UPC/001/V1.

López, X. (2004). Transcriptor fonético automático del español. http://www.aucel.com/pln/transbase.html.

Molino de Ideas (2012). Transcriptor fonético. http://www.fonemolabs.com/transcriptor.html.

Moreno, A. y Mariño, J. B. (1998). Spanish dialects: phonetic transcription. En *Proceedings of the International Conference on Spoken Language Processing, Sydney, Australia (ICSLP'98)*, pp. 189–192.

Pachès, P., de la Mota, C., Riera, M., Perea, M. P., Febrer, A., Estruch, M., Garrido, J. M., Machuca, M. J., Ríos, A., Llisterri, J., Esquerra, I., Hernando, J, Padrell, J., y Nadeu, C. (2000). Segre: An automatic tool for grapheme-to-allophone transcription in Catalan. En *Proceedings of the Workshop on Developing Language Resources for Minority Languages: Reusability and Strategic Priorities, LREC*, pp. 52–61.

Van Bael, C., Boves, L., Van den Heuvel, H., Strik, H. (2007). Automatic phonetic transcription of large speech corpora. *Computer Speech & Language, 21*, 652–668.

Garrido, J.M., Codina, M. y Fodge, K. (2019). TransText, un transcriptor fonético automático de libre distribución para español y catalán. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 27–31). Málaga: Universidad de Málaga.

Subsidia: Tools and Resources for Speech Sciences

6

dVoice: doing phonetics by smartphones

Francesco Cutugno[1,2], Enrico Leone[1], Antonio Origlia[1,2] and Renata Savy[3]

[1] Federico II University
[2] Consiglio Nazionale delle Ricerche
[3] Università degli Studi di Salerno
e-mail: cutugno@unina.it

Citation / Cómo citar esta publicación: Cutugno, F., Leone, E., Origlia, A. & Savy, R. (2019). dVoice: doing phonetics by smartphones. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 33–36). Málaga: Universidad de Málaga.

ABSTRACT: We propose dVoice, an Android application for on-the-field speech recordings and simple acoustic phonetic analysis on smartphones and tablet. The product we present is a completely open source project and presents all typical features expected in a speech recording and analysis tool with specific intervention in the interface design in favor of the use with touchscreens and limitations in the graphic capacities given by the presence of small screens. dVoice includes most of the functionalities normally available in software for mobile devices and, in particular, it consents cloud connection and facilities to synchronize local database with copies in the home PC. dVoice allows labelling, metadata entry and local database function for users, recording sessions and file descriptors.

Keywords: fieldword; Android applications; recording; phonetic analysis; labelling.

RESUMEN: Presentamos dVoice, una aplicación de Android para trabajos de campo que impliquen obtener grabaciones de habla o realizar análisis acústicos sencillos en teléfonos móviles y tabletas. El producto que se presenta es completamente de código abierto y dispone de todas las funciones esperables en una herramienta de grabación y análisis del habla, con una interfaz especialmente diseñada para la interacción táctil, al tiempo que se intentan solventar las limitaciones gráficas derivadas del escaso tamaño de las pantallas. dVoice incluye además la mayoría de las funciones de una aplicación de móvil; en particular, permite conectarse a la nube y sincronizar las bases de datos locales con copias en el ordenador de casa. dVoice permite etiquetar, añadir metadatos y utilizar bases de datos locales para los usuarios, sesiones de grabación y descriptores de los archivos.

Keywords: trabajo de campo; aplicaciones de Android; grabación; análisis fonético; etiquetado.

1. INTRODUCTION

On-field phonetics is still alive. Speech sciences need portable and easy-to-use recording tools and most of the current technologies to acquire sounds are small, friendly and powerful. Present technology offers devices that can be extremely miniaturized, with very high performances, at risible costs and having a wide variety of different possible uses. All the old forms of analogic audio data storing are gone and solid state sound recorders are presently dominating the scene of on-field sound acquisition. However, if phoneticians need to add metadata to their files or if they need to check out measurability of any given speech feature, or to annotate some part of the recordings, then they choose to use a laptop to make recordings or to couple the PC with a portable solid state recorder. In the smartphone era, you no longer need to use a laptop if the proper app is installed on your mobile phone and, in addition, but not mandatorily, you have a good quality external microphone connected to the audio jack on the device. In this paper, we present dVoice (sounding like [di'vois] recalling the word 'device'), an app for Android devices thought to be used as a mini phonetic lab for on-field recordings. The title of our contribution recalls the very famous motto "... doing Phonetics by Computers", that characterized Praat (Boersma y Weenink, 2016) since its birth. Praat constitutes a milestone for Speech Sciences, its arrival has made possible to observe speech signals and "doing Phonetics" with a high quality and with no costs and the research community is nowadays exponentially grown also thanks to it. dVoice has not such an ambition; however, the project is conceived as a work in progress, totally open to the communities of both users and developers that will decide to contribute to its growth. In its design, particular attention is given to stability, usability, modularity to favour integrations and expansions. Obviously, these choices make the final product something that is always subject to changes and improvements. Some parts of what we are going to

Figure 1: The activity diagram of the dVoice software.

describe here will receive upgrading and changes, but what we claim with this paper is the announcement of a new way of doing on-field phonetics, which could imply some interesting falls in the next future.

2. dVoice MAIN CHARACTERISTICS

dVoice is an Android app devoted to recording and analyzing speech signals, which allows the user to freely operate in mobility, to be independent from electrical power connection, to make good quality recordings and, at the same time, to use features such as metadata management, local database access, and simple annotation on signals or on some portion thereof. Online recording level monitoring and, in this preliminary release, offline speech processing algorithms can also be executed and their output shown on the smartphone screen. The project will be perpetually considered as an open source and communitarian work-in-progress initiative, so features listed here will constantly be updated (provided that a community of developers and skilled users will give their support).

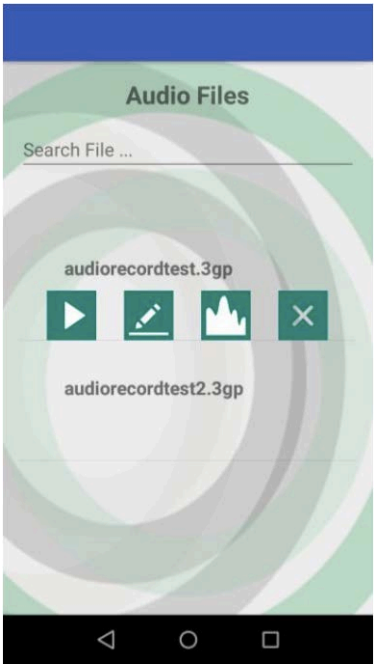
2.1. dVoice MAIN CHARACTERISTICS

Figure 1 shows the basic architecture of the present release of dVoice. Basic instances of recording and saving interfaces have been considered as a starting configuration able to recall typical operations that are normally conducted during speech recording procedures. What is not said in this activity diagram is that (1) usability is guaranteed recurring at the application of the main Human-Computer Interaction guidelines (Donald, 2002; Shneiderman, 1997); and (2) the full power of wireless internet connection is an improving facility for this system as main cloud archives (Dropbox, Drive, OneDrive in principle) can be directly invoked in the application to facilitate the transfer of single files or entire sessions including database partial dump where related metadata are conserved. Then all the produced material can be moved to other system when returning in the lab.

2.2. Actions in the app

Before starting recordings, users can start the metadata

Figure 2: File selection interface.



entry process. This is made filling up a form that requires to enter data concerning the general description of the recording sessions, operator id, environmental condition, etcetera. These data constitute records of a more complex database that will be described in section 3. This database includes descriptions of users, of recording sessions and of audio file features.

We will skip here the description of the recording module and related interfaces, as it does not present any substantial difference (but the graphic layer) from any similar audio-recording application, and we will go into a more accurate description of what can be done on the files that are available in the system at a given time. The file list is available as a scrolling graphic object and, on each file, users can choose to perform operations like (1) listening, (2) waveform, pitch and spectrogram visualization, (3) file-related metadata entering or editing, or (4) light labelling process.

File-related metadata are a further component of the previously cited database (see relevant discussion in section 3). The light labelling process is also a novelty (we will give some hints on it in section 4).

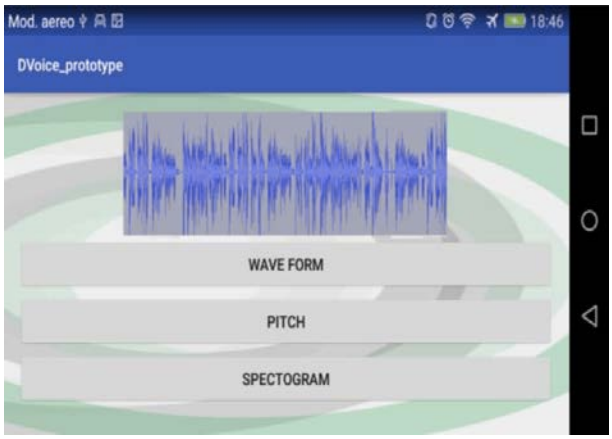
Signal waveform and other temporal patterns visualization will remain, in the first beta versions of the app, quite traditional, and limited, as in this phase we decided that these functions are to be used just for fast controls and coherence checking, while accurate analysis and fine measurements are demanded to more consolidated tools the researchers already use in their daily activity. Figure 2 shows the screenshot of a file-list and icons to launch listening, metadata editing and processing actions.

Figure 3 (a) and (b) show screenshots of the preliminary aspect of these views that are mandatorily seen only in landscape orientation.

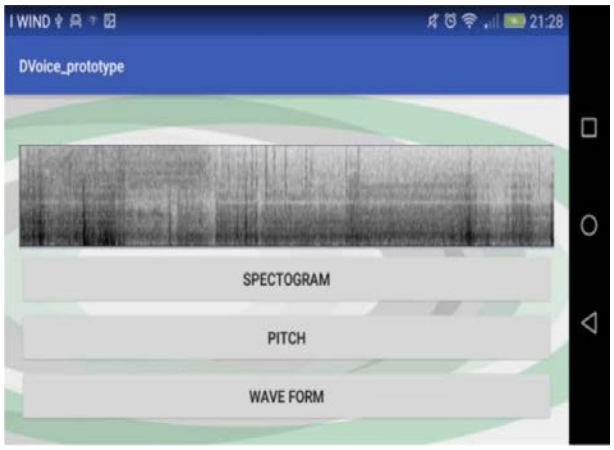
The play (listen) function is, as usual, implemented to listen either the entire audio file or selected parts.

Figure 3: Mockups of waveform (a) and spectrogram (b) views.

(a)



(b)



However, the process of selecting a portion to listen in isolation is not finely tuneable using touches on the screen in alternation to pich-out or pitch-in to zoom. Solutions to model this kind of interaction aiming at increasing the sensibility and to perform then a finer selection are under study. We will turn back to this issue in section 4.

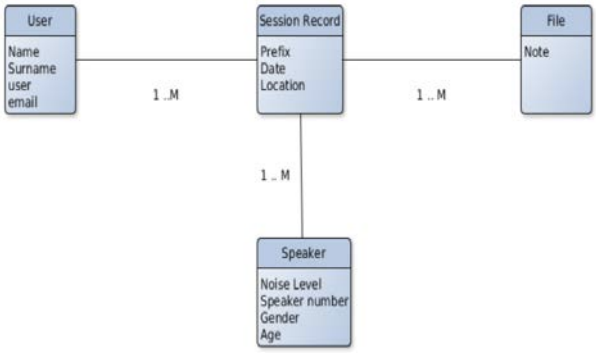
When a portion to listen is reaching the right limit of the screen an animation scrolls left the waveform or the spectrogram.

3. DATABASES AND METADATA

Figure 4 illustrates the database structure at its actual state of evolution. The tool we used to implement this service in the app is SQlite, a small and powerful Data Base Management System largely used in mobile application development.

The extremely simple database shown here introduces four entities presently formalized in the data structure: (1) the user, (2) the recording session, (3) the speaker(s) involved in the session, (4) the files stored in the session. The database approach is new if compared with Praat and, in our vision, it will help to manage recording procedures and further processing activities done when the sessions are downloaded from the smartphone to a PC. In the next future, specific

Figure 4: Entity/relationships diagram for the dVoice local database.



interfaces will be designed to navigate and filter these data. When the users decide to transfer all their work on the cloud or on another generic platform to process their recordings with other tools, the data related to the exported session can be formatted as a SQL relational dump, an xml or a JSON file.

4. LABELLING

Given the actual status of available technologies and the related constraints we think that the system we propose cannot be used to "professionally" perform multi-layered annotations and labelling. Graphic resolution (sufficiently high, but on too small screens) and finger pointing accuracy during temporal selection on signals and its processed versions, would slow and overload the process of fine labelling, time scrolling and layer switch. There remains the challenge to try and make the most of the potentialities presently offered by the Android touch interaction system in order to be able to add a light version of an annotation facility.

In this phase, we decided not to pursue the annotation layer concept any further. Layers are the result of an accurate design concerning corpus linguistics or research purposes where annotation layers constitute the skeleton of the data structure that leads toward the experimental verifications of the linguistic theoretical hypothesis. Nothing of that would necessarily be set up already during the recording phase; no linguistic abstraction and specific labelling procedure should be done on a smartphone or a tablet.

Annotation and labelling on dVoice is thought as a complement or a support to further offline work to be made in a lab. More in general, during on-field recordings it would be useful to annotate some specific phenomena that the operators want to bookmark before they forget and lose it.

4.1. Annotation structure

The alternative to layers is the self-descriptive paradigm. XML, JSON are an example of this approach. Each single label is atomic and unrelated with all the remaining. Type of labelling, name of label, temporal values and label value are repeated at any annotation instantiation producing a virtuous redundancy. Each

Table 1: Labelling structure.

TagName	*Value*
LabelName (LN)	string
LabelValue (LV)	string
TypeOfLabel (TOL)	[instant | interval]
TimeLable (TL)	time(instants) | time+offset(interval)

label is then, in its basic form, a quadruple (Table 1).

However, self-description paradigm allows to dynamically change this structure if needed, to add fields, to omit some, to change order in the structure increasing in this way the expressiveness power of the data structure. Examples of labels are:

{"LN": "word", "LV": "hallo", "TOL": "interval", "TL": {{"start": "(*start_time*)"}, {"offset": (*offset*)}}}

{"LN": "noise", "LV": "cough", "TOL": "instant", "TL": "(*start_time*)"}

Self-descriptive data are expandable, tags can be recursive, considerable freedom is left to the user in the design of even complex annotation system structures. The totality of modern programming languages can manage these data with specific and easy-to-use parsers. In dVoice each audio file can be enriched by a collection of self-descriptive labels, user can decide if these data must be saved in a specific database table or, after some selection processing, in a Praat TextGrid file.

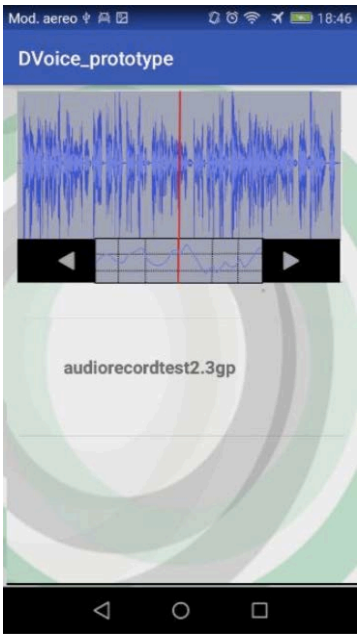
4.2. Producing labels in dVoice

The simplest idea is to use instant flags marking a point, and not an interval, on the signal. This choice can appear convincing if bookmarks are sporadic and unrelated each other. The user select the labelling function when the speech signal view is active, taps once in a point on the waveform (or on other signal views) and fills in the form that generates the label data as seen before. Interval annotations can also be introduced but, once again, it is not conceivable to perform a complex task of interval labelling on smartphones and small tablets, as, on these devices, a fine temporal selection process requires an effort that is not compatible with the context in which the system is used.

We decided to approach the problem of fine temporal interval selection recurring to the slow down magnifier glass metaphor. If users want to add an interval annotation on a signal view, they must double tap around the area where presumably the interval starts, a guided procedure asks them if a fine tuning of the start time is required. The metaphor we used requires users to scroll the signal using left and right arrow while watching the same signal in two windows having different time scales (slow down term derives by the different speeds of scrolling in the two windows under a single command).

Users must put the required interval start time in the centre of the window with the finest time scale. If required, the same procedure can then be repeated for the end point. In Figure 5 you may find a schematic mockup of the proposed procedure.

Figure 5: A mockup of slow down magnifier glass metaphor



5. CONCLUSIONS

dVoice is a project subject to a continuous upgrading. Its status in the days we are proposing this paper is as follows: all functions, interfaces, database, cloud connection and other features have been designed and documented according to software engineering good practices. The main software structure, signal acquisition, signal processing and internal file systems have been implemented. The development process is fluent and the first release on the Google Play store is expected in spring 2017. In the same period, we will launch a campaign of common creation process, all source code will be posted on the main code-sharing platforms, hoping to raise interest in the phonetics research community to continue developing the app and receiving help and suggestions.

6. REFERENCES

Boersma, P. & Weenink, D. (2016). *Praat: doing phonetics by computer* [Computer program]. Version 6.0.22, retrieved 15 November 2016 from http://www.praat.org/

Donald A.N. (2002). *The Design of Everyday Things*. New York, NY: Basic Books, Inc.

Shneiderman, B. (1997). *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (3rd ed.). Boston, MA: Addison-Wesley Longman Publishing Co.

Subsidia: Tools and Resources for Speech Sciences

7

MWN-E: a graph database to merge morpho-syntactic and phonological data for Italian

Antonio Origlia[1], Giulio Paci[2] and Francesco Cutugno[3]

[1] Università degli Studi di Padova
[2] Consiglio Nazionale delle Ricerche
[3] Università degli Studi di Napoli 'Federico II'
e-mail: antonio.origlia@unipd.it, giulio.paci@pd.istc.cnr.it, cutugno@unina.it

Citation / Cómo citar este artículo: Origlia, A., Paci, G., & Cutugno, F. (2019). MWN-E: a graph database to merge morpho-syntactic and phonological data for Italian. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 37–45). Málaga: Universidad de Málaga.

ABSTRACT: In this paper, we present a wordnet database mainly derived from the Italian subset of MultiWordNet. First of all, we report on the conversion of the original SQL MultiWordNet dataset in a modern framework based on graphs. We then describe the process of importing data that was not originally found in MultiWordNet like transcriptions, Part-Of-Speech details and derived word forms. After this first dataset expansion, we further increase the number of relationships among words by computing, for each word, its phonological neighbourhood. The final result is a wordnet that merges syntactic, lexical and phonological information in a format that can be queried in a declarative way, thus making it easier to setup cross-domain constraints to extract interesting data. In order to demonstrate this, we present a number of examples to take advantage of this kind of knowledge. Applications span linguistic studies and technological systems design.

Keywords: wordnets; Italian; graph databases; neo4j; linguistic resources.

RESUMEN: En este artículo se presenta una base de datos de redes léxicas derivada, fundamentalmente, del subcorpus del italiano de MultiWordNet. En primer lugar, se detalla la adaptación de los datos originales de MultiWordNet en SQL a un modelo más moderno basado en gráficos. A continuación se describe el proceso de importación de aquellos tipos de datos que no se encontraban originalmente en MultiWordNet, como las transcripciones, la información categorial sobre las distintas partes de la oración, o las formas léxicas derivadas. Tras esta primera expansión del conjunto de datos, se incrementó el número de relaciones entre palabras computando, para cada una de ellas, su vecindario fonológico. El resultado final es una red léxica que aúna información sintáctica, léxica y fonológica en un formato que permite búsquedas mediante un lenguaje declarativo, lo que permite establecer fácilmente restricciones a través de distintos dominios para extraer los datos de interés. Para demostrar la utilidad de la herramienta, se presentan varios casos en los que se puede hacer uso de tal tipo de conocimiento. Las aplicaciones se extienden tanto a los estudios lingüísticos como al diseño de sistemas tecnológicos.

Palabras clave: redes léxicas; lengua italiana; bases de datos gráficas; neo4j; recursos lingüísticos.

1. INTRODUCTION

Human language is a very complex topic that has been studied from many different perspectives. Morphology, phonology, phonetics, semantics and syntax are all wide fields of research that have generated a large number of resources, over the years, to support researchers working on each aspect. While large amounts of data have been collected, however, the different goals each resource had with respect to its possible applications led to a fragmented scenario in which different data concerning words and their relationships are distributed among unrelated resources. As a consequence, it is difficult to efficiently support cross-domain research that may help to investigate unsolved problems posed by language. A resource designed to allow this kind of investigation may also be useful to provide technological products with the information necessary to handle spoken natural language on all its levels as part of an interpretation process leading to interactive responses, such as in the case of dialogue systems. As an example, the Opendial toolkit (Lison & Kennington, 2016) is designed to support and manage Automatic Speech Recognition (ASR) uncertainties by using confidence scores. After the utterance has been transcribed, Opendial supports dependency trees extraction by

using MaltParser (Nivre *et al.*, 2007) to allow the interpretation of user input and react accordingly. Opendial has, at the same time, the need to handle transcriptions coming from an ASR engine, which is obviously influenced by how similar different words sound, and to interpret these transcriptions to produce a response. While the two problems can be treated separately, considering the syntactic role of words and their pronunciation at the same time can help, for example, to correct the transcription coming from the ASR engine and to avoid missing an utterance. Phonological neighbourhoods are of particular interest in this sense as they connect words depending on how similar they sound and can be easily introduced in a network of items and relationships among items.

Other than information fragmentation, the technology used to store linguistic data is also becoming obsolete and should be updated to take advantage of the most recent findings in the field of representation and access. While large and well-known databases like WordNet (Miller, 1995) have been distributed as Linked Open Data using the Resource Description Framework (RDF) standard (Manola & Miller, 2004), many language-specific resources are still distributed in table-based formats such as Comma-Separated Value and SQL databases. While such formats are convenient and easy to manage, they offer less opportunities to explore, discover and re-use the same data by exploiting their meaning as a whole system rather than simply as collections of items. In this work, we present an ongoing effort to merge a number of available resources describing different aspects of Italian words into a single, graph-oriented database. Such a database can be queried using hybrid constraints that explicitly rely on data representation using sets of items, relationships among items and variable length paths connecting items to each other.

The paper is organised as follows: first of all, we summarise the specific technology used to obtain the presented database in its current form. Then, we describe the resources from which we obtained the original data, we present how these were merged and how new information, namely phonological neighbourhoods, were computed to extend the amount of represented knowledge. We also present a number of possible queries to better illustrate the applications of this approach.

2. NEO4J

In this section we present the concept of graph databases and compare them to RDF. We then present the specific database manager used in this work and the characteristics of its query language, Cypher.

2.1. Corpus description

With the advent of the Big Data and, in particular, with the increasing availability of Linked Open Data, the need to establish a representation format suitable for dynamic, rapidly changing and interconnected objects arose. RDF represents the most widely used solution to this need and has been adopted to implement the most widely known repositories of linked knowledge available today. RDF is based on the concept of triple stores: every statement about a specific domain is represented by a subject, a predicate and an object. From a conceptual point of view, an RDF statement represents a link, whose semantic value is described by the predicate, between two objects. The RDF standard has been developed in the last decade mainly to support semantic web applications and, because of this, it is heavily focused on the concept of Uniform Resource Identifier (URI), later generalised to International Resource Identifier (IRI). In RDF, every concept is represented by an IRI and there is no explicit difference between items and relationships, which are only distinguished by their syntactic role in the statement. Although RDF is agnostic about the semantic meaning of IRIs, these can be unequivocally interpreted if a source vocabulary is specified. For example the IRI 'http://dbpedia.org/resource/Name' specifies that the item Name is supposed to be interpreted in the way specified by DBpedia. Whenever this property is used while specifying the DBpedia vocabulary, all clients accessing the item are informed about its meaning. While the set of RDF triple stores defines a graph of relationships, the representation is edge-centred in the sense that there is no way to specify item or relationship properties in any other way than by establishing new edges.

Graph databases, on the other hand, are node-centred in the sense that the main data container are the graph nodes, with relationships being separated objects describing how these nodes are related to each other. This conceptual separation between nodes and relationships is one of the main differences between RDF and graph databases. Graph databases also explicitly distinguish properties from relationships. While in RDF an item is linked to its properties by the use of relationships, graph databases explicitly represent properties as part of either nodes or relationships. Another fundamental difference between RDF and graph databases is that, in RDF, it is not possible to specify properties for a specific relationship between two nodes.

From the data access point of view, graph databases and RDF both have their strengths and weaknesses so choosing one over the other mainly depends on the intended use. Graph databases are designed to be very efficient in graph traversing and path finding. RDF has, instead, powerful support to relationship inference through the application of logical rules.

2.2. Neo4J and Cypher

Neo4J (Webber, 2012) is an open source graph database manager that has been developed over the last sixteen years and has been applied to a high number of tasks related to data representation (Dietze *et al.*, 2016), exploration (Drakopoulos, Kanavos, Makris, & Megalooikonomou, 2015) and visualisation (Jiménez, Diez, & Ordieres-Mere, 2016). It can be deployed in

Origlia, A., Paci, G., & Cutugno, F. (2019). MWN-E: a graph database to merge morpho-syntactic and phonological data for Italian. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 37–45). Málaga: Universidad de Málaga.

server mode and be queried over a specific port using HTTP or Bolt protocols. It can also be embedded in Java applications through dedicated APIs. In Neo4J, nodes and relationships may be assigned labels, which describe the type of the object they are associated to. In this work, labels are mainly used to represent morpho-syntactic characteristics of words and the nature of the relationships among nodes. Nodes and relationships may have properties, which are used here to store the details of each single node or relationship. Labels and properties are the main way used by Neo4J to filter data and retrieve answers to user queries.

Neo4J is characterised by high scalability, ease of use and by its proprietary query language: Cypher. Cypher is designed to be a declarative language that highlights patterns structure by using an SQL inspired ascii-art syntax. A brief overview of the syntactic elements of Cypher queries is given here to help to understand the example queries presented in this paper. The reader is referred to the online Cypher manual[1] for a more detailed presentation of Cypher. As in graphical representations of graphs nodes are usually represented by circles, in Cypher nodes are represented by round brackets. For example, the query "MATCH (n:VERB) RETURN n" returns all the nodes of the graph labelled as verbs. In the same way, since relationships are usually represented by labelled arrows in graph schemas, relationships between nodes are described by using ASCII arrows, too. The query "MATCH (m)-[:DERIVES_FROM] -> (:VERB word: 'essere') RETURN m" returns all the nodes that contain a term that derives from the essere verb. The SQL-like WHERE clause may also be used to filter results using boolean logic. It is also possible to use Cypher to detect paths of specified length between nodes by using the * operator in the relationship definition.

3. RESOURCES

In this section we summarise the characteristics of the databases used to populate MultiWordNet-Extended (henceforth MWN-E). For each database, we report on the amount of data it contains, about its contents typology and about the way it was collected.

3.1. MultiWordNet

WordNet is a lexical database for the English language, grouping words into sets of synonyms called synsets. It provides short definitions and usage examples for each synset, and records a number of relations among these synsets. A few attempts have been made to extend WordNet to other languages, either by creating synsets in the new language that are transpositions of the synsets in the original WordNet or by creating independent synsets in the new language and trying to associate them to the originalWordNet. Starting from WordNet 1.6, MultiWordNet followed the first approach (Pianta, Bentivogli, & Girardi, 2002) and includes a freely available Italian WordNet (32,673 synsets; 57,934 meanings; 41,491 entry words). EuroWordNet followed the second approach (Vossen, 1998) and includes an Italian WordNet (50,308 synsets) that has also been aligned with WordNet 3.0 (Toral, Bracale, Monachini, & Soria, 2010). The methodologies are mostly independent on the specific Italian WordNet used, so we decided to use MultiWordNet because it is freely available.

3.2. Morph-it

Morph-it is a morphological lexicon for the Italian language. It has been developed using a semi automatic procedure, consisting in word forms extraction from a 25 million tokens corpus automatically annotated with Part-Of-Speech (POS) tags, automatic candidate lemmas identification and inflected forms generation, comparison with word forms extracted from the corpus and manual correction of errors (Zanchetta & Baroni, 2005). The resource includes 506,827 word forms and 31,955 lemmas and for each word form, the associated lemma and the POS tag are reported.

3.3. ISTC pronunciation dictionary

The ISTC pronunciation dictionary includes standard Italian phonetic transcriptions of 3,177,286 distinct word forms in SAMPA format. The lexicon has been originally developed for the Italian module of the Festival Text-To-Speech system (Cosi, Tesser, Gretter, Avesani, & Macon, 2001), but has been later expanded to the current size by several contributors. POS tag information (using Tanl POS tag set) has been added to each pronunciation for the purpose of pronunciation disambiguation; for this reason this information is reliable for all those words with multiple possible pronunciations, but many admissible tags are still missing, especially for word forms with unambiguous pronunciations. A large number of word forms has been added by automatically generating some inflected forms starting from the existing ones and by identifying the most frequent words in Wikipedia whose pronunciation where missing. Pronunciations have been generated using automatic grapheme to phoneme conversion, followed by manual correction.

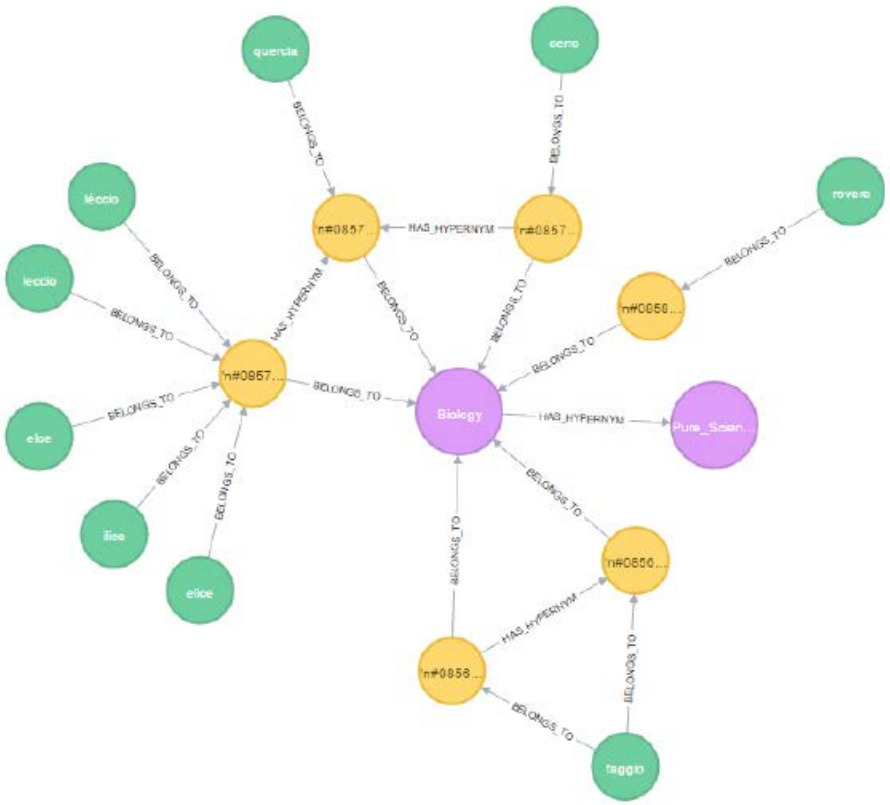
4. MULTIWORDNET-EXTENDED

We now present the generation process of the MWN-E database. First of all, the set of Italian lemmas found in MultiWordNet was imported. In the graph, words are represented as nodes and are assigned the label NOUN, VERB, ADVERB of ADJECTIVE depending on their senses. Then, the Italian synsets were loaded as SYNSET nodes and words were linked to them by BELONGS_TO relationships. In order to keep the alignment with Wordnet, synset IDs are stored in the graph as node properties. Synsets were subsequently linked among each other using the set of common and Italian-specific semantic relationships (e.g. HAS_HYP-ERNYM). As a last step, semantic fields were loaded

[1] https://neo4j.com/developer/cypher-query-language

Origlia, A., Paci, G., & Cutugno, F. (2019). MWN-E: a graph database to merge morpho-syntactic and phonological data for Italian. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 37–45). Málaga: Universidad de Málaga.

Figure 1: A subgraph resulting from the MultiWordNet import step. Semantic fields are shown in purple, synsets are shown in yellow, and nouns are represented in green.



as nodes labelled SEMANTIC_FIELD and synsets were linked to these by BELONGS_TO relationships. The semantic fields hierarchy was also imported by setting up relationships among SEMANTIC_FIELD nodes. At the end of this procedure, the network of semantic relationships depicted in MultiWordNet for the Italian language is represented as a graph. An excerpt of it is shown in Figure 1. As relationships, in Neo4J, are always oriented, relationships of type HAS_HYPONYM were not imported as they are symmetric to the HAS_HYPERNYM relationship and can therefore be easily described in Cypher without being explicitly represented in the database.

MultiWordNet constitutes the fundamental structure on which the database we present is built on. In order to extend words coverage beyond lemmas, derived forms found in the ISTC lexicon were imported. To keep consistency, the final database contains only word forms that are or derive from MultiWordNet lemmas. When the considered ISTC word is a MultiWordNet lemma, its properties must simply be updated. For every other word form, if this was found in the Morph-it database and the corresponding lemma was present in MultiWordNet, a new node was created for each POS type associated with the word. This was linked to the node representing its lemma by a DERIVES_FROM relationship. POS-tagging data were imported from Morph-it when the word was found in this database. If the considered word was not found in Morph-it, a further attempt to find its lemma was made by using Treetagger. If the lemma provided by Treetagger was present in MultiWordNet, the new node was created and its properties were setup using POS data collected from the ISTC lexicon. In all cases, SAMPA transcriptions were also imported from the ISTC lexicon. The full set of properties that are setup for each node is summarised in Table 1. An excerpt of the graph resulting from the MultiWordNet expansion step is shown in Figure 2.

Starting from the extended MultiWordNet graph, it is possible to use the SAMPA transcriptions to compute phonological neighbourhoods. A word A is defined to be a phonological neighbour of the word B if it is possible to obtain B by altering the transcription of A using exactly one Insertion/Deletion/Substitution operation. Computing phonological neighbourhoods provides a further aspect that can be used to query the database using various linguistic aspects at the same time. Phonological neighbourhoods are computed by comparing the SAMPA transcriptions of the words loaded into the database after the second step and by establishing a relationship of type HAS_PHONOLOGICAL_NEIGHBOUR between two words if their Minimum Edit Distance equals 1. This kind of relationship has a distance property that, in these cases, is set to 1. Relationships of type HAS_PHONOLOGICAL_NEIGHBOUR are also established between words that have the same pronunciation but have different written forms. In this case the value of the

Table 1: Summary of the fields defined for each type of label. Among verb forms, only participles have gender and number set.

Label	Name	WordNet ID	Word	Phones	Gender	Number	Degree	Conjugation
SEMANTIC_FIELD	✓							
SYNSET		✓						
NOUN			✓	✓	✓	✓		
ADJECTIVE			✓	✓	✓	✓	✓	
ADVERB			✓	✓				
VERB			✓	✓	○	○		✓

Figure 2: A subgraph resulting from the MultiWordNet expansion step. Two senses of the verb *essere* share the same conjugation. POS tags and SAMPA transcriptions are available as properties in each node.

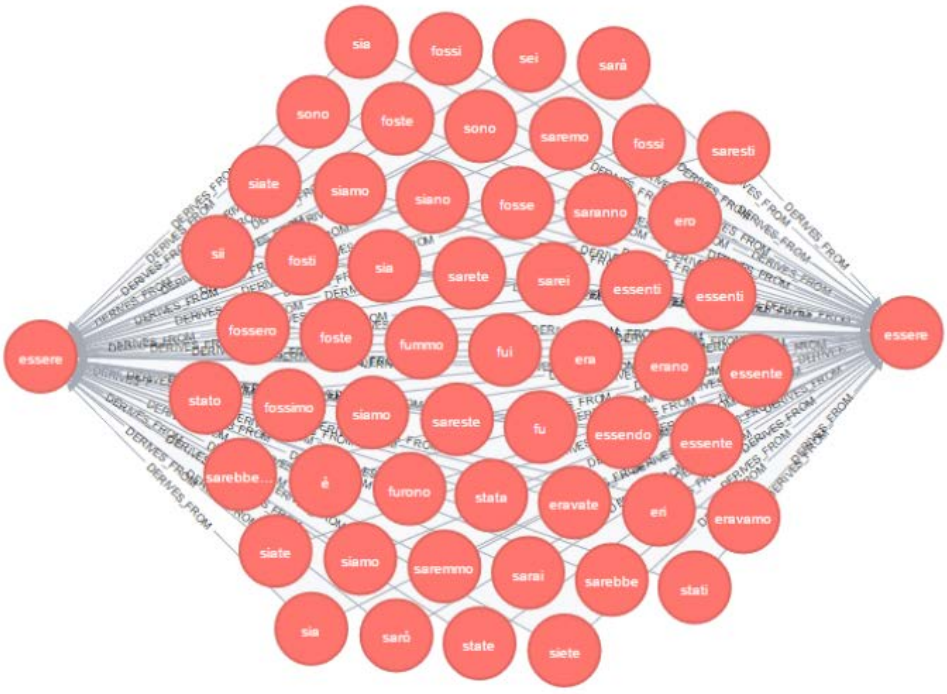


Figure 3: A subgraph resulting from the phonological neighbourhoods generation step. In this example, the phonological neighbours of the word *occhi* are shown.

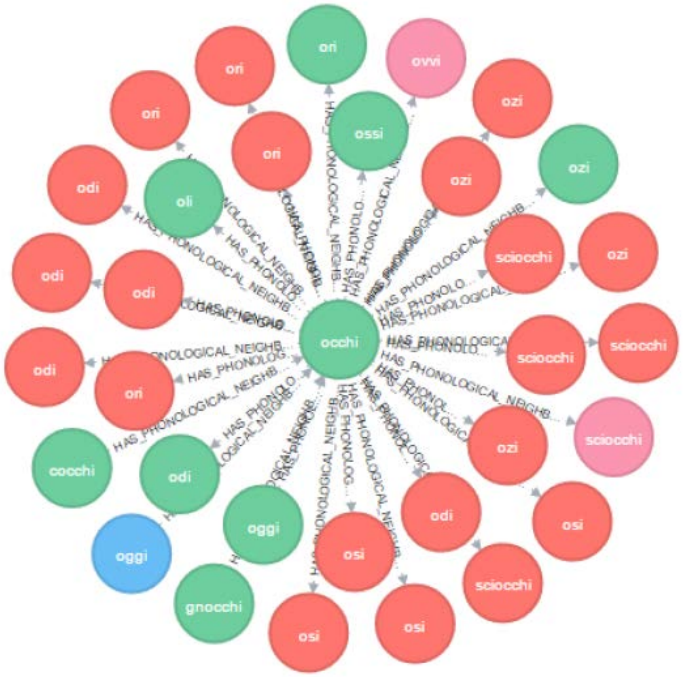
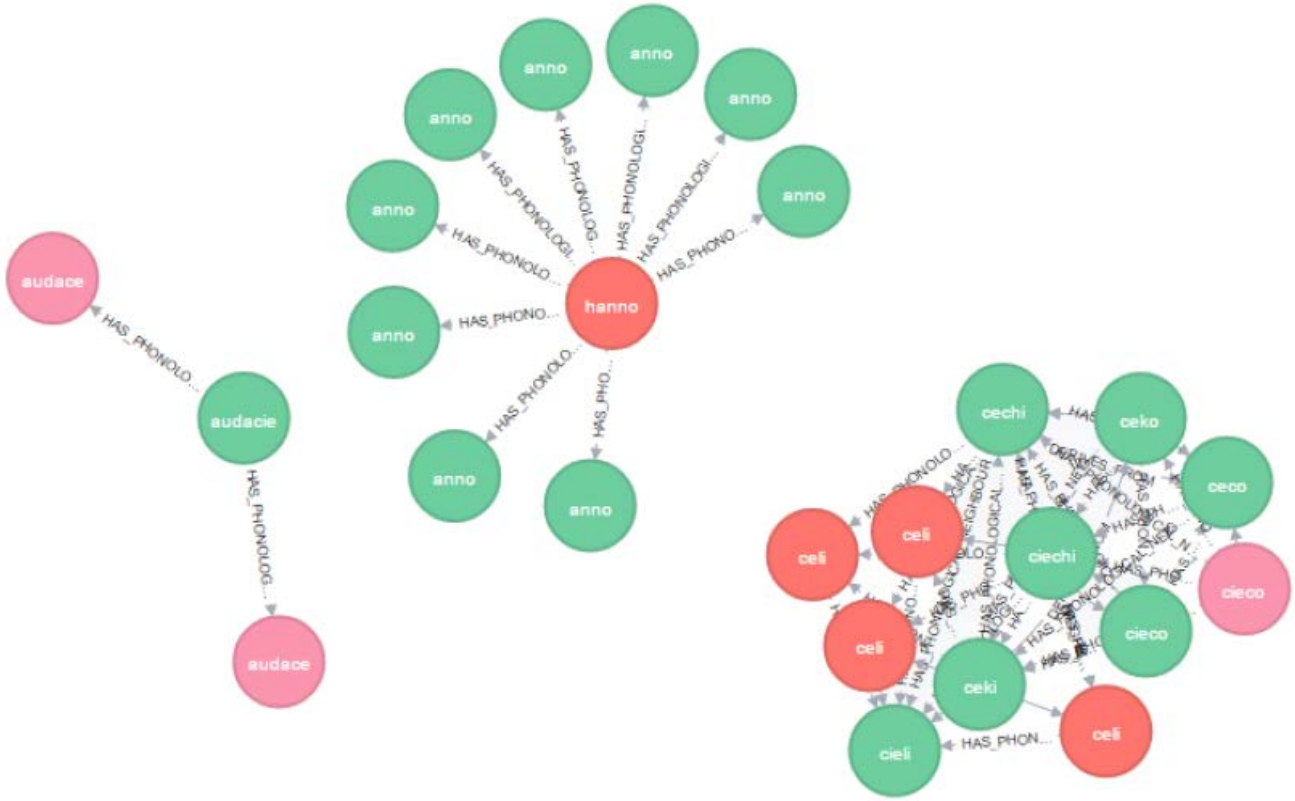


Figure 4: Excerpt of the result to the query MATCH (n)-[:HAS_PHONOLOGICAL_NEIGHBOUR distance:'0']-(m) WHERE n.word <> m.word RETURN DISTINCT n.word, m.word.



distance property is set to 0. An example of phonological neighbourhood is shown in Figure 3. It is important to note that, although the phonological neighbourhood relationship is oriented, like all the connections in Neo4J, it should be interpreted as bidirectional. The Cypher language allows to submit orientation-independent queries to the database and it is therefore recommended to use this strategy, instead of establishing two symmetric connections, to limit the size of the database.

While still in development, MWN-E contains 292,282 nodes containing 1,307,137 properties. 943,174 relationships among these nodes were imported or computed. Phonological neighbourhood relationships represent the majority of this set as 59% of these connections were found among the initial set of words. The vast majority of these reports distance 1 between the involved words.

5. EXAMPLES

At this stage of development, useful information can be extracted from MWN-E to implement, for example, corrections to the transcriptions proposed by Automatic Speech Recognition (ASR). Powerful, commercially available ASR engines are designed to be general purpose and may not be able to take advantage from contextual or domain information, thus feeding less than optimal data to a dialogue manager. This may happen, in particular, when words written in different ways share the same pronunciation. It is possible, in Cypher, to query MWN-E for the set of words that share the same pronunciation but present different written forms to handle this kind of situation, as shown by the example query in Figure 4.

Having phonological and syntactic data in the same database allows MWN-E to efficiently filter the phonological neighbours of a single word by their POS tag. This is useful in cases where it is necessary to substitute words in a given text without altering the syntactic structure of the sentence. This may be the case, for example, of intelligibility tests and of experiments on phonological discrimination capability. The example in Figure 5 shows how it is possible to obtain the phonological neighbours of the adjective *rosso* that are adjectives too and share the same gender and number of the starting word.

While, in MWN-E, we considered phonological neighbourhoods up to a maximum distance of 1 between words, it is possible to consider phonological neighbourhoods at distances higher than 1. While these may be computed and added to the database at a later time, it is currently possible to take into account a specific pattern of neighbourhood that poses an additional constraint on phonological neighbourhoods at a distance higher than 1. For example, by searching for paths composed of relationships of the type HAS_PHONOLOGICAL_NEIGHBOUR, it is possible to find couples of words at phonological distance 2 that have a common phonological neighbour. In other

Figure 5: Result to the query MATCH (n:ADJECTIVE word: 'rosso')-[HAS_PHONOLOGICAL_NEIGHBOUR distance: '1']-(m:ADJECTIVE) WHERE n.gender=m.gender AND n.number=m.number RETURN DISTINCT m.word. The adjective *rosso* appears two times because, in MultiWordNet, it has two senses.

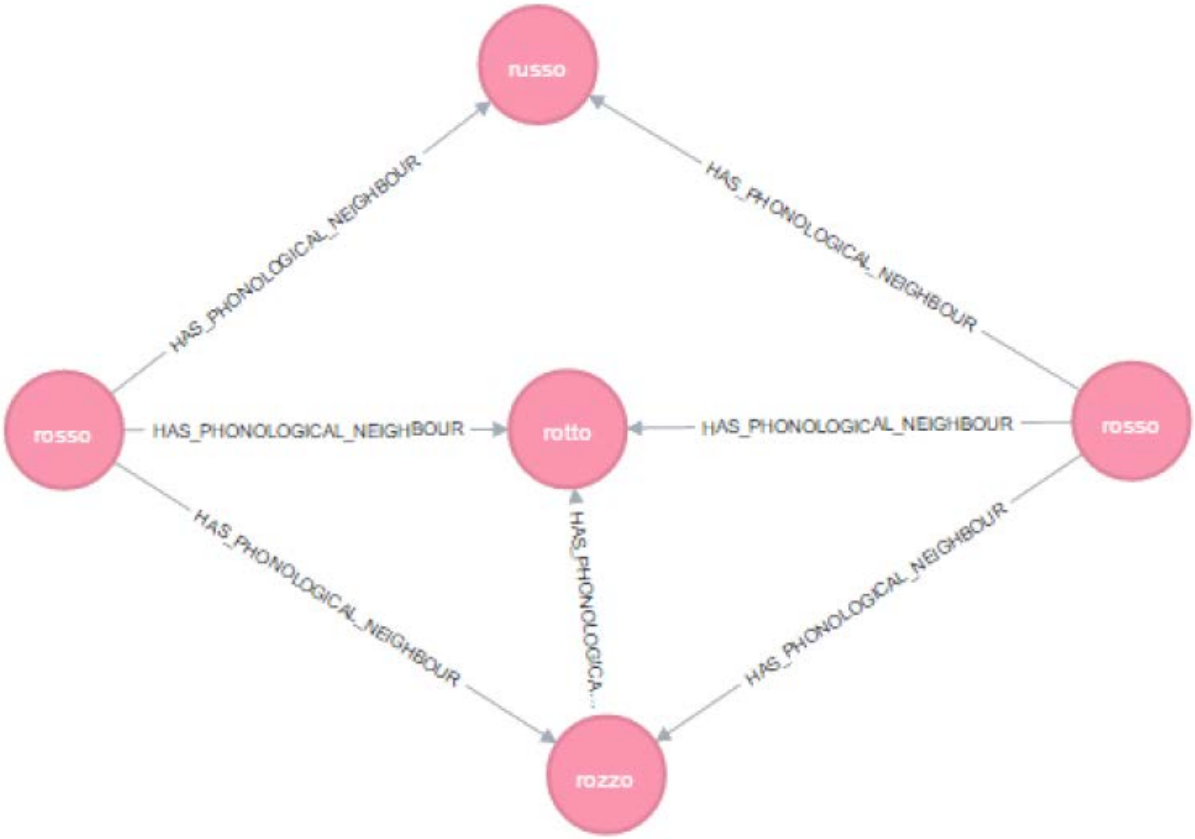


Figure 6: Excerpt of the result to the query MATCH p = (n word: 'rosso')-[:HAS_PHONOLOGICAL_NEIGHBOUR distance: '1']-(m)-[HAS_PHONOLOGICAL_NEIGHBOUR distance: '1']-(l) WHERE n.word <> l.word RETURN DISTINCT p.

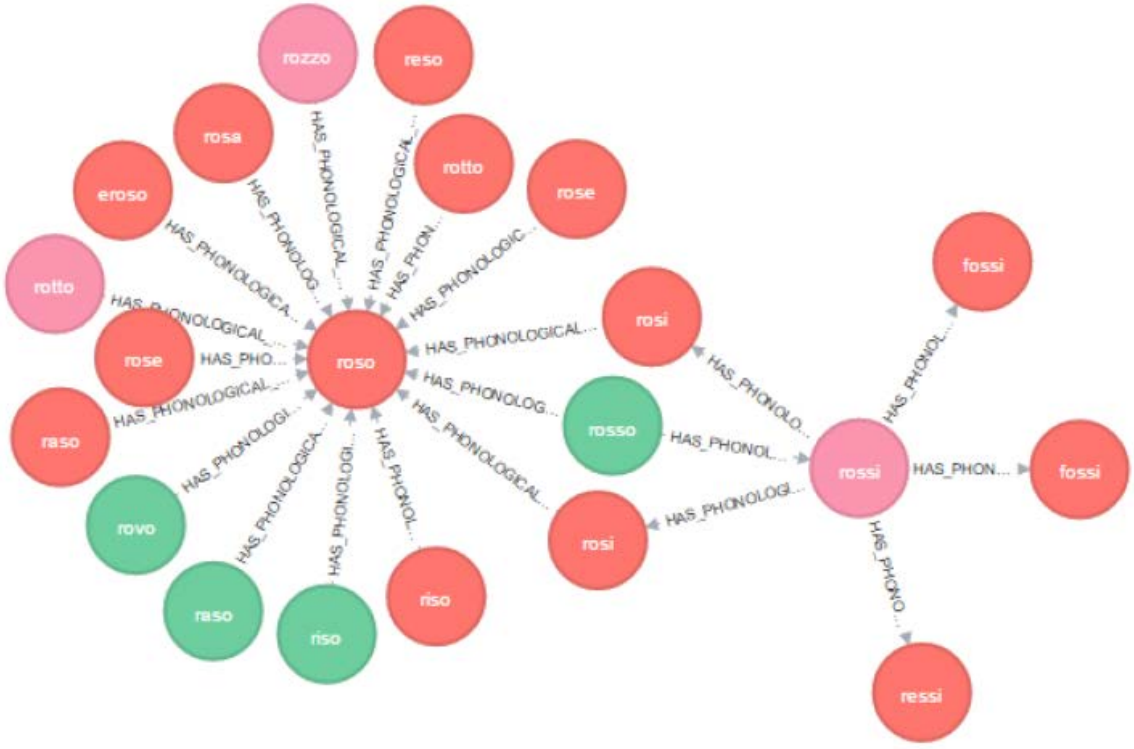
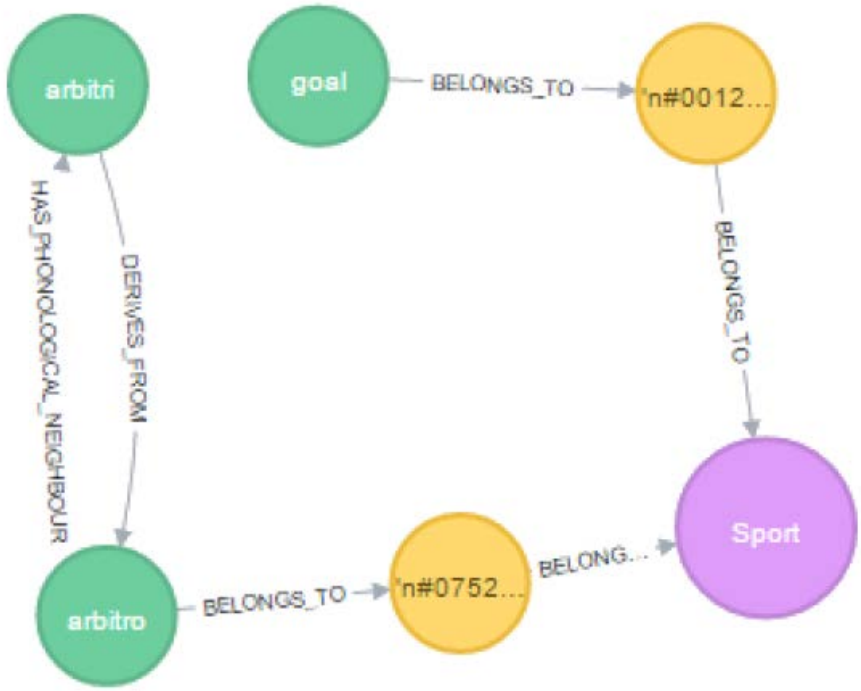


Figure 7: Result to the query MATCH p = allshortestpaths((n word: 'goal')-[*]-(m word: 'arbitri')) RETURN p.



words, it is possible to find a word A that undergoes two changes to be transformed in a word B and that, after the first operation, becomes a third, intermediate, word C. The example in Figure 6 shows the result of a query designed to find these situations.

As a last example, the possibility of looking for paths in the database while using built-in functions like searching for the shortest paths allows to investigate complex relationships between words. Since derived forms are included in MWN-E, it is possible to look for paths involving such terms in the general graph. As an example, Figure 7 shows the result of a query designed to obtain the set of shortest paths connecting the lemma *goal* to the derived term *arbitri* ('referees'). The result clearly shows the connection passing through the lemma *arbitro* and through the semantic field 'sport'. In this case, MWN-E and Cypher handle lemmatisation as part of the query itself.

6. CONCLUSIONS

We have presented the MWN-E database: a graph representation of the MultiWordNet Italian subset extended with data coming from Morph-it and the ISTC lexicon. Also, we have added automatically computed phonological neighbourhoods to represent how similar words can sound. The database merges different aspects of language into a single resource that supports cross-domain constrained queries so that, for example, homophones and acoustically similar words can be easily extracted from MWN-E and filtered using morpho-syntactic constraints. Some applications of this kind of resource are found in dialogue systems to resolve uncertainties coming from ASR engines and to design intelligibility tests for speech synthesis. As this work is still in development, we plan to add more information, like term frequency in various domains, as words properties. This approach may also be applied, in the future, to larger datasets, like the Open Multilingual Wordnet (Bond & Foster, 2013).

7. REFERENCES

Bond, F., & Foster, R. (2013). Linking and extending an open multilingual wordnet. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (pp. 1352–1362).

Cosi, P., Tesser, F., Gretter, R., Avesani, C., & Macon, M. W. (2001). Festival speaks Italian! In *Proceedings of the 7th European Conference on Speech Communication and Technology*.

Dietze, F., Karoff, J., Valdez, A. C., Ziefle, M., Greven, C., & Schroeder, U. (2016). An open-source object-graph-mapping framework for neo4j and scala: Renesca. In *Proceedings of the International Conference on Availability, Reliability, and Security* (pp. 204–218).

Drakopoulos, G., Kanavos, A., Makris, C., & Megalooikonomou, V. (2015). On converting community detection algorithms for fuzzy graphs in neo4j. In *Proceedings of the 5th International Workshop on Combinations of Intelligent Methods and Applications, CIMA*.

Jiménez, P., Diez, J. V., & Ordieres-Mere, J. (2016). Hoshin kanri visualization with neo4j. Empowering leaders to operationalize lean structural networks. *Procedia CIRP*, *55*, 284–289.

Lison, P. & Kennington, C. (2016). Opendial: A toolkit for developing spoken dialogue systems with

probabilistic rules. *Proceedings of the ACL 2016* (p. 67).

Manola, F. & Miller, E. (2004). Rdf primer. Available from https://www.w3.org/TR/2004/REC-rdf-primer-20040210/#rdfmodel.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, *38*(11), 39-41.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., & Marsi, E. (2007). Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, *13*(2), 95–135.

Pianta, E., Bentivogli, L., & Girardi, C. (2002). Developing an aligned multilingual database. In *Proceedings of the 1st International Conference on GlobalWordNet*.

Toral, A., Bracale, S., Monachini, M., & Soria, C. (2010). Rejuvenating the Italian wordnet: upgrading, standardising, extending. In *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010), Mumbai*.

Vossen, P. (Ed.) (1998). *EuroWordNet: A multilingual database with lexical semantic networks*. Norwell, MA: Kluwer Academic Publishers.

Webber, J. (2012). A programmatic introduction to neo4j. In *Proceedings of the 3rd Annual Conference On Systems, Programming, And Applications: Software For Humanity* (pp. 217–218).

Zanchetta, E. & Baroni, M. (2005). Morph-it! A free corpus-based morphological resource for the Italian language. In *Proceedings of Corpus Linguistics*.

Origlia, A., Paci, G., & Cutugno, F. (2019). MWN-E: a graph database to merge morpho-syntactic and phonological data for Italian. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 37–45). Málaga: Universidad de Málaga.

Subsidia: Tools and Resources for Speech Sciences

8

Methodological issues in the assessment of cross-language phonetic similarity

Juli Cebrian

Universitat Autònoma de Barcelona
e-mail: juli.cebrian@uab.es

Citation / Cómo citar esta publicación: Cebrian, J. (2019). Methodological issues in the assessment of cross-language phonetic similarity. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 47–53). Málaga: Universidad de Málaga.

ABSTRACT: The assessment of cross-linguistic similarity remains a crucial methodological issue in speech perception and second language acquisition research. This is so because models of second language speech base their predictions precisely on the degree of similarity between native and non-native sounds. However, it is still unclear what the best approach to cross-language comparisons is. This paper discusses a few key issues in the assessment of cross-linguistic similarity, focussing on perceptual methods, by reporting some results from a series of studies involving native and non-native speakers of English, Catalan and Spanish. The issues discussed include the type of task, the nature and amount of the stimuli used, the effect of amount of L2 experience on cross-language perception, and the use of L1 data as control data. The paper advocates for the use of multiple methods and bidirectional data, and also presents a new approach involving online processing tasks.

Keywords: second language speech; cross-linguistic similarity; speech perception.

RESUMEN: La evaluación de la similitud entre lenguas sigue planteando problemas metodológicos en la investigación sobre percepción del habla y adquisición de segundas lenguas. Esto es así porque los modelos de habla en segundas lenguas basan sus predicciones precisamente en el grado de similitud entre los sonidos nativos y no nativos. Sin embargo, todavía no está claro cuál es la mejor manera de realizar esa comparación entre lenguas. Este artículo analiza algunos problemas clave en la evaluación de la similitud entre lenguas. Para ello, se centra en los métodos perceptivos y aporta resultados de una serie de estudios con hablantes nativos y no nativos de inglés, catalán y español. Los aspectos analizados incluyen el tipo de tarea, la naturaleza y la cantidad de estímulos utilizados, el efecto de la cantidad de experiencia con la L2 sobre la percepción interlingüística, y el uso de datos de la L1 como control. El artículo defiende el uso de múltiples métodos y datos bidireccionales, y también presenta un nuevo enfoque basado en tareas de procesamiento online.

Keywords: lengua segunda; similitud entre lenguas; percepción del habla.

1. INTRODUCTION

It is well-known that second or foreign language (L2) learners tend to perceive and produce target language sounds in terms of native language categories (Best, 1995; Flege, 1995; Kuhl & Iverson, 1995). According to Trubetzkoy (1969), the L1 phonological system functions as a perceptual "sieve" filtering target language (TL) sounds that as a result are categorized in terms of the closest L1 categories, at least at initial stages in the acquisition process. Models of L2 speech try explain the relationship between the level of similarity between native and non-native sounds and success in L2 category formation. For instance, the Native Language Magnet model (Kuhl, 1991; Kuhl & Iverson, 1995) claims that in the process of acquiring the L1, a set of propotypical sound categories are developed which guide L1 perception. These prototypes also affect L2 perception, as non-native sounds are perceptually attracted to the closest L1 sound prototypes. According to the Speech Learning Model (Flege, 1995, 2003), there is a process of *equivalence classification* by which phonetically similar TL sounds are mapped on to existing L1 categories. Thus learners need to discern differences between native and target sounds in order to establish accurate categories for the L2 sounds. Best's Perceptual Assimilation Model (Best, 1995; Best & Tyler, 2007) makes a series of predictions about discriminability of TL sounds based on different patterns of perceptual assimilation of target sounds to L1 sounds.

The notion of cross-linguistic similarity, thus, is crucial in order to make predictions about the relative difficulty and learnability of target language sounds.

2. ASSESSING PHONETIC SIMILARITY

Different methods of assessing the similarity between native and non-native sounds have been suggested, including articulatory comparisons, acoustic comparisons and perceptual judgements.

Articulatory comparisons involve contrasting L1 and L2 sounds on the grounds of articulatory descriptions. While these can be informative and can provide preliminary results, they have been found to fail to reflect perceptual similarity. For example, as Strange (2007) explains, lip rounding is a redundant feature in American English for non-low vowels as non-low front vowels are unrounded and non-low back vowels are rounded. By contrast, rounding is distinctive in German and French, which distinguish between front rounded and front unrounded vowels. On the basis of this comparison, we could predict that English speakers will have trouble differentiating the French / German high front rounded and high front unrounded vowels. Results from perceptual studies, however, show that American English speakers find the contrast between French / German high back rounded and high front rounded vowels harder to discriminate than the front vowel contrast (Polka & Bohn, 1996). Strange, Levy, & Lehnholf (2004) found that American English speakers perceived front rounded vowels as English back rounded vowels rather than as front unrounded vowels, showing that perceptual judgements are better predictors of L2 discrimination ability than phonetic or articulatory descriptions.

Acoustic comparisons involve analyses of the acoustic properties of native and non-native sounds in order to determine the extent to which native and non-native categories overlap (Flege, Bohn & Jang, 1997; Flege, MacKay & Meador, 1999; Tsukada et al. 2005). For example, native and non-native vowels are typically compared in terms of their spectral properties (F1, F2, F3) and temporal properties (duration). However, discrepancies between the acoustic measurements and perceptual judgements are not infrequent (e.g., Bohn, Strange, & Trent, 1999; Stevens, Liberman, Studdert-Kennedy, & Öhman, 1996; Strange, Levy et al., 2004). For instance, Cebrian (2006) found that pairs of native and non-native vowels that had very similar degrees of acoustic similarity based on average steady state F1 and F2 measurements patterned differently when the perceptual similarity between the same pairs of vowels was examined. It is possible that additional vowel properties such as formant trajectories or f_0 need to be taken into account too in order to obtain a more complete picture of acoustic similarity. Further, acoustic characteristics vary considerably as a function of inter-speaker differences, phonetic context and prosodic context. In any case, recent proposals advocate for the use of perceptual measures, or a combination of perceptual and acoustic comparisons, as the best approach to determining cross-linguistic phonetic similarity (Bohn, 2002; Strange, 2007).

The focus of this paper is on perceptual similarity. The main issues concerning perceptual approaches to cross-language similarity are presented in the next section.

3. PERCEPTUAL MEASURES OF CROSS-LINGUISTIC SIMILARITY

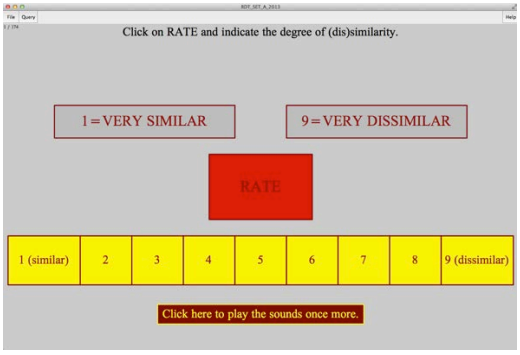
Perceptual measures of similarity involve perceptual judgements of two kinds: indirect covert comparisons and direct overt comparisons. The former are typically represented by the paired comparison technique (rated dissimilarity tasks) and the latter by interlingual identification tasks (perceptual assimilation tasks).

3.1. Rated dissimilarity task

A rated dissimilarity task (RDT) is a paired comparison task in which listeners are presented with a pair of stimuli and they have to rate the degree of (dis)similarity between the two stimuli by means of a 7-point or a 9-point Likert scale. Figure 1 presents an example of the visual display that accompanies each trial in an RDT, using Praat (Boersma & Weenink, 2016). Flege, Munro & Fox (1994) used this technique to evaluate the perceived similarity between Spanish and English vowels. Cebrian, Mora & Aliaga-García (2011) also used it to compare British English vowels and Catalan vowels.

Rated dissimilarity tasks present pairs of stimuli representing different conditions. These may include pairs of two L1 sounds, two L2 sounds or an L1 and an L2 sound. In addition, the two members of each pair may belong to the same sound category (possibly for control purposes) or to different categories, in which case they may be from adjacent categories (e.g., English /æ/ and /ɛ/) or distant categories (/æ/ and /i/). By way of illustration, Table 1 provides an example of the results obtained by Cebrian et al. (2011) indicating the mean dissimilarity rating for each type of vowel pair.

Figure 1: Example of rated dissimilation task. Display of response alternatives and rating scale.



Cebrian, J. (2019). Methodological issues in the assessment of cross-language phonetic similarity. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 47–53). Málaga: Universidad de Málaga.

Table 1: Mean dissimilarity ratings for each type of vowel pair (1 = same, 7 = different; L1 = Catalan, L2 = English) from Cebrian et al. (2011).

Type of pair	Language	Mean dissimilarity rating (SD)
same category	L1-L1	1.7 (0.5)
same category	L2-L2	2.0 (0.6)
adjacent category	L2-L2	3.4 (0.8)
adjacent category	L1-L2	3.7 (0.6)
adjacent category	L1-L1	4.2 (0.8)

Results in this case show that adjacent L2-L2 vowels are perceived as being more similar than adjacent L1-L1 vowels, possibly showing more defined categories for L1 sounds. Cebrian et al. also found that some L2-L1 pairs obtained similarity ratings that fell within the values of those obtained by same-category L1-L1 pairs, showing that some L2 vowels are perceived as near identical to L1 vowels.

Strange (2007) argues that direct overt tasks such as the paired technique comparison (Flege et al., 1994; Cebrian et al., 2011) are problematic because in these tasks listeners do not compare a given stimulus to their own mental representations of L1 phonetic categories. Instead, listeners compare two physical stimuli: an L2 sound and an L1 sound, the latter produced by a speaker who is different from the listener. Hence, the task may not involve accessing the listener's actual own internal representations. Strange advocates for tasks that present a single stimulus to be compared with the listener's own processing categories, such as an interlingual identification task.

3.2. Perceptual assimilation task

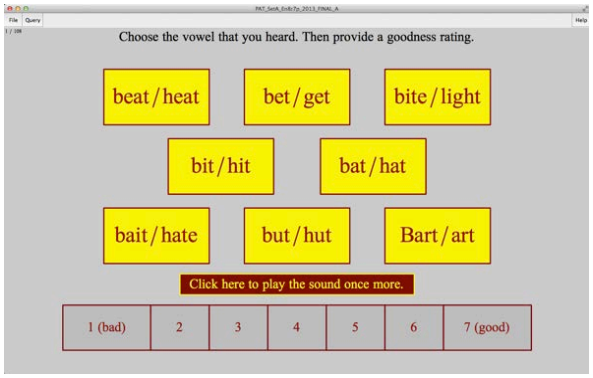
Recent research advocates for cross-language mapping tasks or interlingual identification tasks (Bohn, 2002; Strange, 2007), also known as perceptual assimilation tasks (Guion, Flege, Akahane-Yamada & Pruitt, 2000). In this task, listeners are presented with a single L2 stimulus and have to identify it in terms of L1 categories and then provide a goodness of fit rating. These tasks have been used to examine the perceptual similarity between native and non-native consonants (Guion et al., 2000; Park & de Jong, 2008; Schmidt, 1996, 2007, among others) and vowels (Lengeris, 2009; Strange, Bohn, Trent & Nishi, 2004; Strange, Levy et al, 2004, among others). Figure 2 presents an example of the visual display of a perceptual assimilation task (PAT), illustrating the L1 (English) category responses and a 7-point rating scale to indicate the goodness of fit rating, using Praat (Boersma & Weenink, 2016).

The results of a perceptual assimilation task are typically presented in a confusion matrix showing the percentage assimilation of each L2 vowel to the closest L1 vowels (i.e., identification percentage of L2 stimuli in terms of L1 categories) and the median goodness of fit rating. An example is provided in Table 2. In order to take both measures into account, Guion et al. (2000) proposed a "fit index" score, calculated by multiplying the identification percentage by the median rating value.

Table 2: Example of confusion matrix showing PAT results (adapted from Cebrian et al. 2011). Each row shows the percentage assimilation of each L2 English stimulus to a L1 Catalan vowel. Goodness ratings are given in parentheses.

	English stimuli						
	/i	ɪ	eɪ	ɛ	æ	ʌ	ɜ/
i	96 (4.6)	82 (3.2)					
e		15 (3.4)		7 (3.3)			30 (2.0)
ei			90 (3.5)				
ɛ				91 (4.7)			24 (1.7)
a					100 (4.7)	85 (3.7)	23 (1.4)
ai			8 (2.7)				
ɔ						11 (4)	9 (2.7)

Figure 2: Example of Perceptual assimilation task. Display of response alternatives and rating scale.



PATs thus offer an appropriate means of providing perceptual similarity data. However, problems remain concerning the interpretation of the results (e.g. how high or low does a "fit index" need to be to consider a given L2 vowel a good or a poor match for an L1 vowel?), as well as concerning methodological issues such as the nature of the stimuli and the inclusion or exclusion of control L1 sounds.

The next sections explore different factors that may affect the outcome of perceptual similarity measurements. These can be grouped in terms of whether they concern the task itself and the type of stimuli used or if they involve individual differences concerning the listeners.

4. TASK FACTORS

4.1. Stimuli and other task design issues

Sounds are affected by their phonetic and prosodic contexts. This fact raises the question of whether the type of context in which target stimuli are presented in tasks like RDTs and PATs will have an effect on the similarity judgements. Strange, Bohn et al. (2004), Strange, Levy et al. (2004), and Strange et al. (2005) tested the perceptual similarity of North German

vowels to American English vowels in different prosodic environments, namely words in citation form and longer words in carrier sentences. Although results were consistent in many cases, it was also found that prosodic context had an effect. For instance, when German /œ/ was presented in citation form, it was identified as an English back vowel 55% of the time and as an English front vowel 45% of the time. By contrast, when presented in a multisyllabic word embedded in a carrier sentence, the same vowel (German /œ/) was assimilated to an English back vowel 96% of the time.

The phonetic context, e.g., the nature of the segments preceding and / or following the target sound, has also been found to affect vowel assimilation patterns. Bohn & Steinlen (2003) found that Danish speakers assimilated English /i/ to Danish /e/ in glottal and alveolar contexts, but as /i/ in velar context. Further, Levy (2009) reported that assimilation patterns of French vowels to English vowels were more consistent in a bilabial context than in an alveolar context. In fact, discrepancies across studies testing the same population have been linked to the use of different phonetic contexts in different studies. Rallo Fabra (2005) and Rallo Fabra & Romero (2012) found that experienced Catalan learners of English identified English /ɪ/ mostly as Catalan /i/. By contrast, the experienced, and inexperienced, learners in Cebrian (2006) classified English /ɪ/ most frequently as Catalan /e/. Still, the studies differed in different ways. Stimuli in Rallo Fabra's studies involved sVt words, while Cebrian's stimuli consisted of vowels in isolation. Further, Rallo Fabra's stimuli were elicited from a dialectally non-homogeneous group of American English speakers, while the stimuli in Cebrian's study were produced by Canadian English speakers. Another methodological difference between the two studies was the fact that Rallo Fabra, but not Cebrian, included a "non-L1" response as a possible response alternative in the PAT, which was also chosen as a response for English /ɪ/.

Finally, studies examining perceptual similarity also differ in whether listeners are tested on non-native / L2 sounds only, or if L1 sounds are also included for comparison purposes. While the inclusion of L1 stimuli provides a useful baseline for native-like categorization, it also makes it possible for listeners to directly compare L1 and L2 sounds across trials, thus possibly interfering with the intended comparisons between the auditory stimuli and the listeners' internal mental representations.

In summary, different factors may affect the way non-native sounds are assimilated to native sounds, including the phonetic context, the prosodic context, the inclusion of control L1 stimuli in the task, the availability of a "none" response alternative or the type of native variety represented in the task.

5. LISTENER FACTORS

5.1. Amount of L2 experience

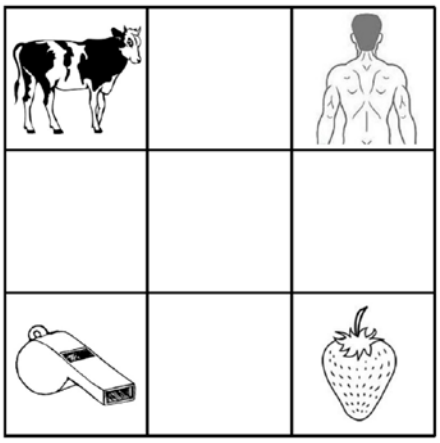
Strange (2007) argues that a complete analysis of cross-linguistic similarity should incorporate assessment of similarity by listener groups with different levels of L2 experience. Different studies have investigated if the perception of similarity between L1 and L2 sounds varies as a function of L2 experience. Recall that models like the Speech Learning Model (Flege, 1995, 2003) propose that learners need to discern differences between native and target-language sounds in order to establish authentic (target-like) categories for the L2 sounds. Flege claims that this ability is not lost as a result of maturation and that learners can eventually detect differences between native and non-native sounds given enough exposure to and experience with the target language. Few studies, however, have examined the effect of experience on cross-linguistic perception. Further, these studies have examined this issue by comparing two groups differing in experience, rather than investigating changes within the same population in a longitudinal approach.

Flege et al. (1994) tested Spanish speakers with different levels of L2 English proficiency by means of a rated dissimilarity task. They found that experience with English had little effect on the degree of perceived dissimilarity. Cebrian (2006) found that experience with the L2, understood as length of residence in the target language country, affected the identification of L1 vowels, but did not seem to affect the perception of similarity between L1 and L2 vowels. Further, Cebrian (2009) also found a fairly consistent pattern of English to Catalan vowel assimilation by two groups of Catalan speakers varying in amount of exposure to English from minimal exposure to several years of instruction.

Other studies have reported some effect of experience. Frieda & Nozawa (2007) compared the perceptual assimilation of English vowels to L1 Japanese vowels by two groups of Japanese L2 English speakers differing in level of proficiency. They found that the groups only differed in their assimilation of one of the English vowels, namely /ɪ/, showing some effect of experience, but experience did not affect the perception of the remaining vowels. Finally, Rallo Fabra (2005) and Rallo Fabra & Romero (2012) found differences in how experienced and inexperienced Catalan learners of English classified the English vowels /ɪ/ and /æ/. For instance, experienced learners tended to perceive English /ɪ/ as the L1 /i/, while the inexperienced learners classified it as the L1 /e/. Further, the experienced learners made a greater use of the "non-Catalan" response than the inexperienced group. These findings may indicate that experience may have enhanced the ability to distinguish L2 to sounds from L1 sounds, a prerequisite for more target-like L2 category formation according to most L2 speech models (e.g., Flege, 1995).

Possibly, in order to assess the effect of experience on the perceived similarity between native and

Figure 3: Visual display in an eye-tracking experiment (Cebrian & Mora, 2016).



non-native sounds, a better approach would be a longitudinal study examining potential changes or developments in cross-linguistic perception by the same group of learners as a function of increased L2 experience.

5.2. Bidirectionality

Cross-language similarity studies typically test speakers of one of the two languages involved, namely speakers of the L1 in the study. Few studies have contrasted the same data from the point of view of both speakers of the L1 and of the L2. Some exceptions are Schmidt (1996, 2007) and Cebrian (2015). Schmidt tested the perception of a series of English and Korean consonants by native speakers of Korean and native speakers of English. Cebrian (2015) contrasted the results of RDTs involving pairs of English and Catalan vowels performed by native speakers of British English and native speakers of Catalan. In both cases, the researchers advocate for the contribution of bidirectionality as a more complete approach to measuring cross-linguistic similarity.

6. NON-PERCEPTUAL TASKS

As discussed in the previous sections, while perceptual measures are currently the most frequently used methods of assessing cross-linguistic similarity, there remain a number of practical and theoretical limitations. On the one hand, the inclusion of sufficient stimuli to obtain adequate data on which to base similarity judgements already renders perceptual tasks rather long and potentially tedious for the listeners. Further, the fact that the phonetic and the prosodic contexts affect the way sounds are perceived suggests that in order to obtain a reliable measure, multiple tasks would be needed, including different types of contexts and conditions. This would of course increase the length of tasks and the probable fatigue effects. On the other hand, identification and dissimilarity ratings tasks are tasks that require the listener to reflect on the stimuli provided and pass a judgement. Such off-line tasks do not reflect the way that sounds are processed in real-life speech perception or in every-day conversations. An alternative approach would be one involving online tasks of the sort that are used in language processing research.

6.1. On-line tasks and language processing

There is evidence that L2 speakers access both L1 and L2 lexicons when processing L2 speech (Chambers & Cooke, 2009; Marian & Spivey, 2003, among others). This evidence comes from eye-tracking studies in which participants follow instructions to click on a depicted target word, presented alongside a phonological competitor and two distractors. Using this methodology, for instance, Marian & Spivey (2003) found that in the course of processing the English word *marker* (target word), Russian speakers of L2 English would look to a picture of a stamp (Russian "marku", phonological competitor from the L1) more often than to pictures of phonologically unrelated words (distractors). This indicates that when processing a given L2 word, speakers activate similar sounding L1 words, at least temporarily. Following these findings, Cebrian & Mora (2016) explored the use of such online tasks to measure phonetic or perceived similarity between L1 and L2 sounds. For instance, in order to explore the phonetic similarity between English /æ/ or /ʌ/ and Spanish or Catalan /a/, crucial trials included an English word containing each of these vowels and another word whose translation into the L1 contains vowel /a/.

One example is given in Figure 3, from Cebrian and Mora (2016). In this case, the target word is *back* (/bæk/) and the interlingual competitor is *cow* (Spanish and Catalan *vaca*, whose first three sounds are /bak-/). The other two pictures in the display show unrelated distractors. Eye gazes to the interlingual competitor (vaca) are measured as the participant hears and processes the instructions to *click on the back*. Comparing the results for this type of trial with trials involving /ʌ/ as the target vowel (e.g., target *buck* and competitor *vaca*) would provide an online measure of which of the two English vowels, /æ/ or /ʌ/, trigger more looks to the interlingual competitor (/a/).

In fact, preliminary results reported by Cebrian and Mora (2016) show a close link between PAT results and the results of an online task. The results of the PAT showed that Catalan /a/ is closer to English /æ/ than to English /ʌ/. This result went hand in hand with the finding that greater L1 competition was observed from Catalan /a/ when the target word contained English /æ/ than when it contained English /ʌ/. Online tasks thus emerge as a potentially effective method to assess crosslinguistic similarity.

7. SUMMARY AND CONCLUSIONS

This paper has reviewed the importance of assessing the phonetic similarity between native and non-native sounds in order to make appropriate predictions about the learnability of L2 sounds. The main methodological approaches have been discussed, with an emphasis on perceptual methods of assessment such as rated

Cebrian, J. (2019). Methodological issues in the assessment of cross-language phonetic similarity. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 47–53). Málaga: Universidad de Málaga.

dissimilarity tasks and perceptual assimilation tasks. Despite being the most reliable method of measuring cross-language similarity, perceptual tasks still face a number of limitations including the need to control and incorporate the effects of phonetic and prosodic contexts and their relative length and potential fatigue effects. Online tasks used in language processing offer new alternatives to the assessment of cross-linguistic similarity.

8. ACKNOWLEDGMENTS

This research was supported by a research grant from the Spanish Ministry of Economy and Competitiveness (FFI2013-46354-P) and by a grant from the Catalan Government (2014SGR61).

9. REFERENCES

Best, C. T. (1995). A direct realist view on cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). Timonium, MD: York Press.

Best, C. T., Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). Amsterdam: John Benjamins.

Boersma, P., Weenink, D. (2016). Praat: doing phonetics by computer [Computer program]. Version 6.0.22, retrieved 15 November 2016 from http://www.praat.org/

Bohn, O.-S. (2002). On phonetic similarity. In P. Burmeister, T. Piske, & A. Rohde (Eds.), *An integrated view of language development: Papers in honor of Henning Wode*. (pp. 191–216). Wissenschaftlicher Verlag Trier.

Bohn, O-S., Steinlen, A. (2003). Consonantal context affects cross-language perception of vowels. In D. Recasens, M.J. Solé & J. Romero (Eds.), *Proceedings of the 15th International Congress of the Phonetics Sciences*. Barcelona, Spain: Universitat Autònoma de Barcelona / Causal Productions.

Bohn, O.-S., Strange, W. & Trent, S.A. (1999). On what it takes to predict perceptual difficulty in cross-language vowel perception, *Journal of the Acoustical Society of America*, 105, No. 2, Pt. 2.

Cebrian, J. (2006) Experience and the use of duration in the categorization of L2 vowels. *Journal of Phonetics* 34, 372-387.

Cebrian, J. (2009). Effects of native language and amount of experience on crosslinguistic perception. Journal of the Acoustical Society of America, Vol. 125, No. 4, Pt. 2, April, p. 2775. ISSN: 0001-4966.

Cebrian, J. (2015). Reciprocal measures of perceived similarity. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: the University of Glasgow. Paper number 1041.1-9.

Cebrian, J., Mora, J.C., Aliaga-García, C. (2011). Assessing crosslinguistic similarity by means of rated discrimination and perceptual assimilation tasks. In Wrembel, M., Kul, M., Dziubalska-Kołaczyk, K. (eds.), *Achievements and Perspectives in the Acquisition of L2 Speech*. Frankfurt: Peter Lang. Vol I. 41-52.

Cebrian, J, Mora, J.C. (2016).Asymmetric lexical access and crosslinguistic perceptual similarity: An eye-tracking study. Paper presented at the 8th International Symposium on the Acquisition of Second Language Speech (New Sounds). Aarhus University, Aarhus, Denmark.

Chambers, C.G., Cooke, H. (2009). Lexical competition during second-language listening: Sentence context, but not proficiency, constrains interference from the native lexicon. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 35, 1029-1040.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language researc*h (pp. 233–277). Timonium, MD: York Press.

Flege, J. E. (2003). Assessing constraints on second language segmental production and perception. In: A. Meyer & N. Schiller (eds.), *Phonetics and Phonology in Language Comprehension and Production, Differences and Similarities*. Berlin: Mouton de Gruyter.

Flege, J. E., Munro, M. J., Fox, R. A. (1994). Auditory and categorical effects on cross-language vowel perception. *Journal of the Acoustical Society of America*, 95, 3623–3641.

Flege, J. E., Bohn, O-S. & Jang, S. (1997) Effects of experience on non-native speakers' production and perception of English vowels, *Journal of Phonetics* **25**, 437-470,.

Flege, J. E., MacKay, I. R. A. & Meador, D. (1999). Native Italian speakers' production and perception of English vowels, *Journal of the Acoustical Society of America* 106, 29763-2987.

Frieda, E. M., Nozawa, T.(2007). You are what you eat phonetically. The effect of linguistic experience on the perception of foreign vowels. Bohn, O-S. & M. J. Munro (eds.). *Language Experience in Second Language Speech Learning* (pp. 79-96). John Benjamins.

Guion, S. G., Flege, J. E., Akahane-Yamada, R., Pruitt, J.S. (2000). An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *Journal of the Acoustical Society of America* 107 (5) Pt. 1., 2711–2724.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics*, 50(2), 93–107.

Kuhl, P. K., Iverson, P. (1995). Linguistic experience and the "perceptual magnet effect. In W. Strange

(Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (p. 121–154). Timonium, MD: York Press.

Marian, V., & Spivey, M. (2003). Competing activation in bilingual language processing. *Bilingualism: Language and Cognition* 6, 97–115.

Polka, L., Bohn, O-S. (1996). A cross-language comparison of vowel perception in English-learning and German-learning infants. *Journal of the Acoustical Society of America*, 100, 577-592.

Lengeris, A. (2009). Perceptual assimilation and L2 learning: Evidence from the perception of Southern British English vowels by native speakers of Greek and Japanese. *Phonetica*, 66, 169-187.

Levy, E. S. (2009). Language experience and consonantal context effects on perceptual assimilation of French vowels by American-English learners of French. *Journal of the Acoustical Society of America*, 125, 1138–1152.

Park, H. & de Jong, K. J. (2008). Perceptual category mapping between English and Korean prevocalic obstruents: Evidence from mapping effects in second language identification skills. *Journal of Phonetics*, 36, 704-723.

Rallo Fabra, L. (2005). Predicting ease of acquisition of L2 speech sounds: A perceived dissimilarity test. *Vigo International Journal of Applied Linguistics*, 2, 75–92.

Rallo Fabra, L., Romero, J. (2012). Native Catalan learners' perception and production of English vowels. *Journal of Phonetics*, 40, 491–508.

Schmidt, A. M. (1996). Cross-language identification of consonants. Part 1. Korean perception of English. *Journal of the Acoustical Society of America*, 99, 3201–3211.

Schmidt, A. M. (2007). Cross-language consonant identification. English and Korean. Bohn, O-S. & M. J. Munro (eds.). *Language Experience in Second Language Speech Learning* (pp. 185-200). John Benjamins.

Stevens, K., Liberman, A., Studdert-Kennedy, M. & Öhman, S. (1996). Cross language study of vowel perception, *Language Speech*, 12, 1-23.

Strange, W. (2007). Cross-language similarity of vowels. Theoretical and methodological issues. Bohn, O-S. & M. J. Munro (eds.). *Language Experience in Second Language Speech Learning* (pp. 15-34). John Benjamins.

Strange, W., Bohn, O-S., Trent, S.A., Nishi, K. (2004). Acoustic and perceptual similarity of North German and American English vowels. *Journal of the Acoustical Society of America*. 115, 1791-1807.

Strange, W., Levy, E & Lehnholf, R. (2004). Perceptual assimilation of French and German vowels by American English monolinguals: Acoustic similarity does not predict perceptual similarity. *Journal of the Acoustical Society of America* 115, 2606.

Strange, W., Bohn, O-S., Nishi, K., Trent, S.A. (2005). Contextual variation in the acoustic and perceptual similarity of North German and American English vowels. *Journal of the Acoustical Society of America* 118, 1751-1762.

Trubetzkoy, N. (1969).*Principles of phonology*. Berkeley: University of California Press.

Tsukada, K., Birdsong, D., Bialystok, E., Mack, M., Sung, H., & Flege, J. (2005). A developmental study of English vowel production and perception by native English adults and children. *Journal of Phonetics*, 33, 263-290.

Cebrian, J. (2019). Methodological issues in the assessment of cross-language phonetic similarity. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 47–53). Málaga: Universidad de Málaga.

Subsidia: Tools and Resources for Speech Sciences

9

Exploiting a multimedia academic corpus for learning Spanish as a Foreign Language: *Video4ELE-UNED*

Victoria Marrero[1] and Víctor Fresno[1]

[1] Universidad Nacional de Educación a Distancia
e-mail: vmarrero@flog.uned.es

Citation / Cómo citar este artículo: Marrero, V. & Fresno, V. (2019). Exploiting a multimedia corpus for learning Spanish as a Foreign Language: *Video-4ELE-UNED*. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 55–57). Málaga: Universidad de Málaga.

ABSTRACT: Video4ELE-UNED is the first outcome of a larger project (SARILRAM-ELE) aimed to build a linguistic information retrieval system from multimedia repositories of academic nature. Our starting point is a large video database from the broadcasting of the UNED in RTVE (the public Spanish Radio and TV Corporation), which have high quality handmade subtitles. In this paper we will show the current prototype, running over a selection of 1514 of those subtitled videos. We will compare the corpus with CREA, *Corpus de Referencia del Español Actual*, from the Real Academia Española, in order to determine its usefulness in the area of Spanish as a Foreign Language (SFL). Finally, some of the didactic materials created by means of Video4ELE will be shown.

Keywords: multimedia repositories; Spanish as a Foreign Language; subtitles.

RESUMEN: Video4ELE-UNED es el primer resultado de un proyecto mayor (SARILRAM-ELE), cuyo objetivo es construir un sistema de recuperación de la información a partir de respositorios multimedia de carácter académico. Nuestro punto de partida es una gran base de datos de vídeos de las emisiones de la UNED en Radio Televisión Española (RTVE), que incluyen subtítulos de calidad procesados a mano. En este artículo mostramos el prototipo en su estado actual, que incluye una selección de 1514 de esos vídeos subtitulados. Comparamos el corpus con el *Corpus de Referencia del Español Actual* (CREA), de la Real Academia Española, para determinar su utilidad en el ámbito del Español como Lengua Extranjera (ELE). Finalmente, mostramos algunos de los materiales didácticos creados gracias a Video4ELE.

Palabras clave: repositorios multimedia; Español como Lengua Extranjera; subtítulos.

1. INTRODUCTION

The tool we introduce here is placed in the broader context of a research project named *Sistema de Acceso y Recuperación de la Información Lingüística en Repositorios Académicos Multimedia con Aplicaciones a la Enseñanza del Español como Lengua Extranjera*. SARILRAM-ELE). It aims to build a system for recovery of the linguistic information from multimedia repositories of an academic nature, generating a specific platform for research and exploitation of services, from own resources already existing.

So far, this tool has been applied to the field of SFL, but it should be also useful for many other research and applications interested in spoken standard Spanish, in an academic register.

2. THE *VIDEO4ELE* TOOL

From a whole corpus of almost 8,000 academic long videos, we have selected a sample dataset of 1,514. The subtitling associated with them (which has been developed under the supervision of experts and has a great quality) has generated 243,123 transcripts, with almost 65,000 different lexical units (*UNED-ELE* corpus). Transcripts have been morphologically labelled by means of the Freeling package (Padró & Stanislovsky, 2012), and phonetically annotated with Aucel (López Morràs, http://aucel.com/pln/). After that, each fragment of subtitle has been indexed along with its lemma, linguistic tags, video to which it belongs and its time mark (with Apache Lucene).

Finally, an information retrieval system has been created.

This way, Video4ELE allows us to extract sounds, words or phrases according to specific search pattern-based queries.

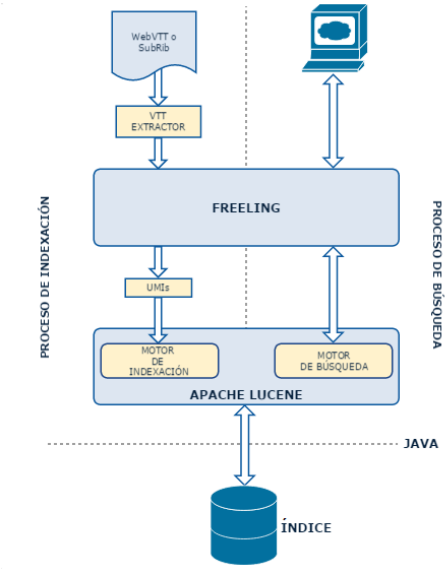
In this paper we will show the prototype with its current utilities:

- Search by free text, admitting wildcard characters or meta-characters.

Table 1. Comparison between UNED-ELE and CREA.

Corpus	Tokens	Type/token ratio
CREA	152,522,438	0.48
UNED-ELE	2,007,861	3.11

Figure 1. Processes in *Video4ELE*, from Villarejo-Aguilar (2015, p. 37).



- Search by free text, including querying by lemma and/or word stems, and allowing to establish search patterns combining the grammatical categories corresponding to the previous and later units.
- Phonetic search.

Once the chain is retrieved, in any of the three types of search, the system returns the corresponding fragments of video, audio and subtitles, perfectly synchronized, allowing the access to visual and auditory information, oral and written, phonetics, lexical and textual simultaneously.

3. CHARACTERISTICS OF THE *UNED-ELE* CORPUS. COMPARISON WITH THE *CREA* CORPUS

A comparison between our corpus (UNED-ELE) and the *Corpus de Referencia del Español Actual* (CREA), of the Real Academia Española can be seen in Table 1 (Villarejo Aguilar, 2015).

UNED-ELE is, by the moment, a corpus much smaller than *CREA*, but it shows a type/token ratio much higher. That leads us to consider that we have a higher lexical density than standard Spanish, at less as represented in the RAE corpus.

Concerning the word classes, we find almost double proportion of numbers and 50% more conjunctions, but 25% less interjections (academic language has a low emotional profile). The present is the prevailing verbal tense in *UNED-ELE* corpus much more than in *CREA*. We find also a greater use of first and second person.

Table 2: Some examples of words ordered by frequency in both corpora.

Entry	UNED-ELE position	CREA position
universidad (university)	87	508
educación (education)	121	483
investigación (research)	144	491
estudiantes (students)	187	1197
formación (training)	190	718
ciencia (science)	204	855

Table 3: Some combinations of word classes and their frequencies. *Str.* = structure; *Prep.* = preposition; *Det.* = determiner; *Adj.* = adjective; *Verb-Ind.* = verb, indicative mode; *VerbIndPr.* = verb, indicative mode, present tense.

Str.	Cat.1	Cat.2	Cat.3	Freq.
1	*Prep.*	*Det.*	*Noun*	110,124
2	*Det.*	*Noun*	*Prep.*	68,230
3	*Verb*	*Det.*	*Noun*	54,896
4	*Noun*	*Prep.*	*Det.*	49,904
5	*Det.*	*Noun*	*Adj*	45,097
6	*Noun*	*Prep.*	*Noun*	41,126
7	*Verb*	*Prep.*	*Det.*	34,293
8	*VerbInd*	*Det.*	*Noun*	34,199
9	*VerbIndPr*	*Det.*	*Noun*	26,227
10	*Det.*	*Noun*	*Verb*	22,213

The semantic fields in our more frequent words characterize this corpus as belonging to an academic and scientific domain; a specialized language use highly relevant for second language students and people interested in the Spanish-speaking universitary environment.

In Table 2 we can find different words ordered by frequency of occurrence in both corpora.

If we consider the combinations of word classes, more than 5,500 patterns have been registered, over 10,648 possible combinations.

Table 3 shows the top ten combinations of three elements, with their corresponding number of occurrences. Some examples of structures are the following:

(1) *en el ámbito* (in the field)
(2) *una universidad con* (a university with)
(3) *fue un fracaso* (it was a failure)
(4) *colaboración con la…* (collaboration with the…)
(5) *muchos estudiantes extranjeros* (many foreign students)
(6) *laboratorio desde casa* (laboratory from home)
(7) *celebrado en la…* (held in the…)
(8) *presentó un trabajo* (presented a paper)
(9) *incluye un concepto* ([it] includes a concept)
(10) *una formación adecuada* (an adequate training).

4. SOME APPLICATIONS

Finally, we will present some didactic applications for the area of the teaching of Spanish as a foreign language that our team has developed, mostly in the grammatical area (Andión-Herrero and Criado de Diego, 2017; Pau Molina, 2016), but also in the phonological, prosodic area (Martínez Acacio, 2017), see Figure 2, Figure 3 and Figure 4.

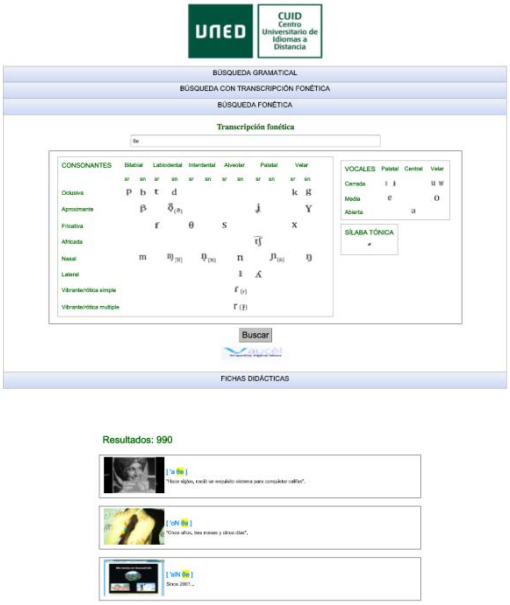
5. CONCLUSIONS

Our project is still in a very initial stage. In the next future we are planning a labelling system for higher levels of language (textual, discursive and pragmatic levels). The system will allow the participation of expert users in a collaborative workflow, with a double aim: to enrich the corpus itself and to build a teaching community in EFL.

Figure 2: Applying search patterns combining the grammatical categories.

Figure 3: Applying search patterns combining the grammatical categories.

Figure 4: Applying search patterns combining the grammatical categories.



6. REFERENCES

Andión Herrero, M.A. & Criado de Diego, C. (2017). *Claves de gramática. Cuaderno de ejercicios para el nivel B1*. Madrid: UNED.

Martínez Acacio, E. (2017). *La entonación de las interrogativas en español: Una propuesta didáctica* (Master Thesis). UNED.

Padró, L. y Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012).*

Pau Molina, S. (2016). *Nexos temporales: actividades gramaticales a partir de la aplicación Video4ELE de la UNED* (Master Thesis). UNED.

Villarejo-Aguilar, D. (2015). *Sistema de acceso y recuperación de información lingüística para su uso en la enseñanza del Español como Lengua Extranjera. Caracterización y analítica del corpus audiovisual de la UNED* (Master Thesis). UNED.

Marrero, V. & Fresno, V. (2019). Exploiting a multimedia academic corpus for learning Spanish as a Foreign Language: *Video4ELE-UNED*. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 55–57). Málaga: Universidad de Málaga.

Subsidia: Tools and Resources for Speech Sciences

Plataforma interactiva para el autoaprendizaje de la pronunciación inglesa: la enseñanza de la entonación

Eva Estebas Vilaplana

Universidad Nacional de Educación a Distancia
e-mail: eestebas@flog.uned.es

Citation / Cómo citar este artículo: Estebas Vilaplana, E. (2019). Plataforma interactiva para el autoaprendizaje de la pronunciación inglesa: la enseñanza de la entonación. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 59–64). Málaga: Universidad de Málaga.

RESUMEN: El principal objetivo de este trabajo es presentar la plataforma interactiva *Teach Yourself English Pronunciation* diseñada para el autoaprendizaje de la pronunciación inglesa en un contexto de enseñanza a distancia. La herramienta va dirigida a estudiantes hispanohablantes de inglés que aprenden fonética de manera autónoma. La plataforma contiene ocho espacios que cubren los principales temas de la fonética inglesa, desde aspectos segmentales (vocales, consonantes, combinación de sonidos y procesos de habla secuenciada) hasta rasgos prosódicos (acentos, ritmo, entonación y focalización). Los contenidos se presentan a través de microtemas que incluyen ejemplos auditivos de pares mínimos y una batería de ejercicios con *feedback* inmediato. El diseño de la plataforma supuso una revisión de los principales modelos de análisis entonativo del inglés y de su efectividad a la hora de enseñar la entonación inglesa a distancia. Se incorporó una nueva metodología, denominada *TL_ToBI* (*ToBI for Teaching and Learning*), que combina parte de las premisas del sistema de anotación prosódica *ToBI* con el análisis de las curvas melódicas del modelo británico.

Palabras clave: plataforma interactiva; pronunciación inglesa; entonación; enseñanza a distancia; *TL_ToBI*.

ABSTRACT: The main aim of this paper is to present the interactive platform *Teach Yourself English Pronunciation* specifically designed for the independent learning of English pronunciation in a distance learning context. The tool is addressed to Spanish students of English who learn phonetics autonomously. The platform is divided into eight areas which cover the main topics of English phonetics, both segmental (vowels, consonants, consonant clusters and connected speech processes) and suprasegmental (stress, rhythm, intonation and focalization). The contents are presented in tips which include audio examples of minimal pairs and a battery of exercises with immediate feedback. The design of the platform prompted a revision of the main models of intonation analysis in English and their efficacy in the teaching of English intonation in a distance learning environment. A new methodology, called *TL_ToBI* (*ToBI for Teaching and Learning*), was incorporated. The new approach combines some of the tenets of the *ToBI* system with the traditional teaching insights of the British School.

Keywords: interactive platform; English pronunciation; intonation; distance learning; *TL_ToBI*.

1. INTRODUCCIÓN

Durante los últimos años la enseñanza de la pronunciación inglesa ha ido adquiriendo cada vez más presencia en las aulas de inglés como L2 (Kissling, 2013). Dadas las notables diferencias entre los rasgos fónicos del inglés y los de otras lenguas, como es el caso del español, son ya muchos los profesores de inglés que ven conveniente despertar la conciencia de los estudiantes sobre la relevancia del aprendizaje de la fonética inglesa para conseguir una comunicación óptima en esta lengua (Jenkins, 2000). En la mayoría de los casos, los docentes se centran en la práctica de los rasgos segmentales, es decir, en la producción y en la percepción de los principales sonidos vocálicos y consonánticos del inglés. Por el contrario, la enseñanza de los rasgos suprasegmentales o prosódicos, como acento, ritmo o entonación, queda relegada a un segundo plano y son escasas las ocasiones en las que la prosodia tiene presencia en el aula. Esto se debe, por un lado, a la poca relevancia que se ha dado a la prosodia a la hora de formar a los docentes de una L2 y, por otro lado, a la falta de una metodología adecuada para la enseñanza de los rasgos suprasegmentales y, más en concreto, de la entonación. Esta situación se ve agravada en un contexto de enseñanza a distancia donde las instrucciones y los materiales de aprendizaje tienen

que ser muy claros para que el alumno asimile la materia sin la ayuda de un profesor presencial.

El principal objetivo de este trabajo es presentar la plataforma interactiva *Teach Yourself English Pronunciation (TYEP)* para el autoaprendizaje de la fonética inglesa. El mayor atributo de esta herramienta es que cubre no únicamente los rasgos segmentales del inglés sino también los suprasegmentales. En el ámbito de la prosodia, la elaboración de la plataforma puso de manifiesto ciertas carencias de los métodos tradicionales para la enseñanza de la entonación inglesa a distancia por lo que se ideó una nueva metodología que facilitara el autoaprendizaje de las características melódicas del inglés.

En este estudio se muestra, en primer lugar, la estructura general de la plataforma, sus objetivos, su diseño y su metodología, y en segundo lugar, se presenta el sistema *TL_ToBI* (*ToBI for Teaching and Learning*) creado para la enseñanza específica de la entonación inglesa a distancia.

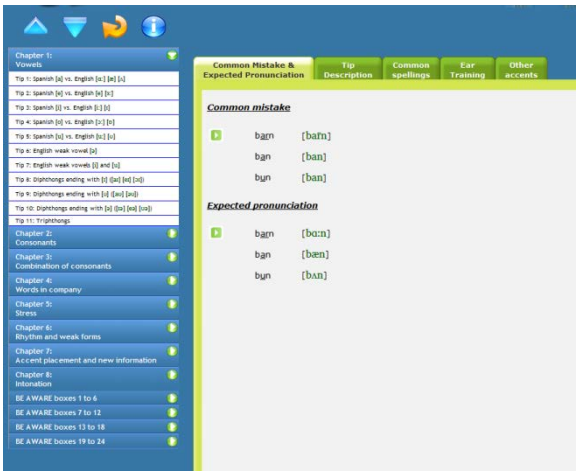
2. LA PLATAFORMA *TYEP*

La plataforma interactiva *TYEP* es una herramienta cuya finalidad es ayudar a los estudiantes de inglés con un nivel medio/alto a practicar las principales características de la pronunciación inglesa de manera autónoma. La plataforma incluye material interactivo que cubre de forma exhaustiva todo el temario de fonética inglesa en un ámbito de enseñanza superior. Se divide en ocho secciones que tratan tanto aspectos segmentales (vocales, consonantes, combinación de sonidos y procesos de habla secuenciada) como prosódicos (acento, ritmo, focalización y entonación). La plataforma presenta un diseño por microtemas. La metodología de aprendizaje se basa en la detección de errores, en la comparación entre lenguas (inglés y español) y en la adquisición de los contrastes fónicos a través de pares mínimos.

2.1. División en microtemas

Cada una de las ocho secciones en las que se divide la plataforma incluye distintos microtemas que abordan un asunto concreto de la pronunciación inglesa. Cada uno de ellos es autónomo, es decir, se puede hacer en el orden que desee el estudiante por lo que propicia la creación de distintos itinerarios de estudio según las necesidades de cada alumno. Para cada microtema, los estudiantes acceden a un área de trabajo dividida en cinco espacios que incluyen: (1) presentación de los errores comunes de los hispanohablantes sobre un determinado aspecto fónico del inglés y contraste con la producción nativa; (2) ejemplos auditivos de los fenómenos; (3) grafías (en caso de los sonidos consonánticos y vocálicos); (4) ejercicios interactivos de identificación y discriminación auditiva con *feedback* inmediato; y (5) actividades de contraste entre el acento inglés británico y el americano. La Figura 1 incluye un ejemplo del área de trabajo de la plataforma *TYEP* y de la división en microtemas para el primer capítulo, dedicado a las vocales.

Figura 1: Ejemplo del área de trabajo de la plataforma interactiva *TYEP* y su división en microtemas.



2.2. Detección de errores y comparación entre lenguas

La presentación de los distintos aspectos fonéticos del inglés se realiza a partir de la producción errónea de un determinado fenómeno por parte de un hispanohablante, seguida de la pronunciación esperada producida por un hablante inglés. A partir de ahí, se establece una comparación entre la pronunciación inglesa y la española con el fin de que los estudiantes no únicamente aprecien las discrepancias fónicas entre el inglés y el español sino que también se beneficien de sus similitudes. Una vez los estudiantes son conscientes de las diferencias entre ambas lenguas, se procede a practicar el fenómeno concreto a través de pares mínimos.

2.3. Audición de pares mínimos

Un par mínimo está constituido por dos palabras que difieren en un único sonido, como por ejemplo, *coat* (“abrigo”) y *goat* (“cabra”). Aunque inicialmente los pares mínimos se empleaban para determinar los contrastes fonológicos de una lengua, generalmente desconocida, se vio que este método era también muy efectivo para la enseñanza de la fonética de una L2. En el caso concreto del inglés, el uso de los pares mínimos para la enseñanza de la pronunciación tiene una tradición de más de medio siglo (véase O’Connor, 1967, entre otros).

En la plataforma *TYEP* el método de pares mínimos se aplica no únicamente en el ámbito de la palabra sino también en el de la frase para discernir rasgos prosódicos. Por ejemplo, *Melanie’s singing* tiene una lectura neutra con foco amplio frente a *MELANIE’S singing*, una frase marcada con foco estrecho en el sujeto.

La audición de pares mínimos en la plataforma *TYEP* tiene dos ámbitos de aplicación: (1) como ejemplos de un determinado contraste y (2) como actividades auditivas tanto de identificación como de discriminación de un determinado rasgo. En total, la plataforma *TYEP* consta de una batería de más de 2000

Estebas Vilaplana, E. (2019). Plataforma interactiva para el autoaprendizaje de la pronunciación inglesa: la enseñanza de la entonación. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 59–64). Málaga: Universidad de Málaga.

archivos de audio y más de 200 ejercicios interactivos con *feedback* inmediato para el autoaprendizaje de la pronunciación inglesa.

El sistema de pares mínimos se usa en todas las secciones temáticas de la plataforma *TYEP*, incluidas las relacionadas con la prosodia. En el caso concreto de la entonación, dada la complejidad que supone a veces captar el recorrido de la curva melódica, se estimó necesario presentar los pares mínimos de los contrastes entonativos con un sistema adicional de notación prosódica que ayudara a los estudiantes a entender los movimientos tonales.

Para ello se revisaron los modelos tradicionales de análisis entonativo y se diseñó una nueva propuesta de notación melódica más visual y accesible para la enseñanza a distancia.

3. EL SISTEMA *TL_ToBI*

El estudio y la modelización de la prosodia inglesa y, en particular, de su entonación se enmarcan dentro de dos grandes corrientes de análisis entonativo, conocidas como la Escuela Británica y la Escuela Americana. La Escuela Británica (Cruttenden, 1986; O'Connor y Arnold, 1973; Wells, 2006, entre otros) se caracteriza por un análisis configuracional del contorno entonativo, con una parte nuclear y otra prenuclear, y por un inventario de tonos definidos según sus trayectorias (ascendente, descendente, sostenido, etc.). Tanto la Escuela Americana como las propuestas más recientes derivadas de ella, concretamente el modelo Métrico-Autosegmental (Ladd, 1996; Pierrehumbert, 1980, entre otros) o el sistema *ToBI* (*Tone and Break Indices)* (Beckman y Hirshberg, 1997), analizan los contornos entonativos mediante objetivos tonales (*targets*) con valores alto (H) y bajo (L), que se asocian a las sílabas acentuadas y al final de una frase entonativa. Mientras que el sistema británico se ha usado para fines pedagógicos, los modelos americanos se han caracterizado por ser descripciones más teóricas.

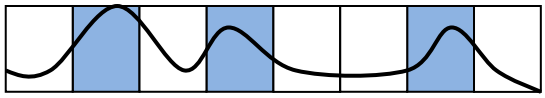
En el contexto de la enseñanza de la entonación inglesa a distancia, ambos modelos presentan ciertas carencias. En el sistema británico es difícil ver, sin un apoyo presencial, la alineación de las trayectorias de las categorías tonales con puntos concretos de la cadena segmental. El sistema *ToBI*, a su vez, presenta un inventario de tonos con diferencias de alineación tan sutiles (H*, L+H*, L*+H, H*+L, L+!H*, etc.) que pueden resultar difíciles de manejar en un entorno de autoaprendizaje.

Este trabajo presenta un nuevo modelo para la enseñanza de la entonación inglesa, *TL_ToBI* (*ToBI for Teaching and Learning*), que parte de las premisas del sistema *ToBI* inicial pero que a la vez incorpora características del modelo británico. *TL_ToBI* adopta del *ToBI* original la asociación de los tonos a las sílabas acentuadas y a los finales de la frase entonativa pero incorpora cuatro diferencias: (1) ayudas visuales (gráficas) que permiten ver la división silábica solapada con los movimientos tonales relevantes; (2) la separación entre la información métrica, señalizada mediante un asterisco en las sílabas con acento léxico, y la información tonal, descrita mediante tres tonos H, L y M (tono medio); (3) el uso de acentos tonales solo monotonales; y (4) la incorporación de tonos de frontera bitonales. Del sistema británico, hereda el concepto de un único nivel de fraseo prosódico y la interpretación del contorno según su configuración (nuclear y prenuclear). En las siguientes secciones se incluyen detalles más concretos y ejemplos del sistema *TL_ToBI*.

3.1. Información visual

En *TL_ToBI*, para cada curva melódica, el estudiante no únicamente oye el audio sino que también ve una representación gráfica del contorno que se solapa con la división silábica, representada en forma de cajas. Debajo de cada caja aparecen escritas las sílabas que forman el enunciado. Las sílabas con acento léxico están marcadas con un asterisco. Las cajas sombreadas indican que en esta sílaba hay un movimiento tonal relevante. Véase el ejemplo (1).

(1)

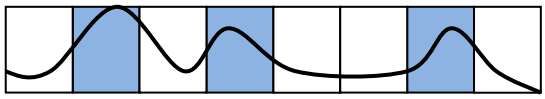


I'm *sure she'll *call you to *mor row.

3.2. Información métrica e información tonal

Mientras que en el sistema *ToBI* cada etiqueta contiene información tonal (H, L) y métrica (* indica un tono asociado a una sílaba con acento léxico y % se usa para un tono de frontera), en la propuesta *TL_ToBI*, la información métrica y la tonal están separadas. El asterisco, en lugar de aparecer junto al tono, se ubica delante de las sílabas con acento léxico en la cadena segmental. De esta forma, los estudiantes ven qué sílabas pueden potencialmente estar vinculadas a un movimiento tonal relevante. Para la descripción del contorno melódico, los tonos se sitúan debajo de las sílabas tónicas pertinentes. La sílaba con el último acento tonal se subraya para indicar que los tonos posteriores corresponden a tonos de frontera. El ejemplo (2) contiene tres acentos tonales altos (H) y un tono de frontera bajo (L).

(2)



I'm *sure she'll *call you to *mor row.
H H H L

La separación entre la información métrica y la tonal sirve de ayuda para entender que no todas las sílabas con acento léxico están asociadas a un acento tonal, como se ejemplifica en (3).

(3)

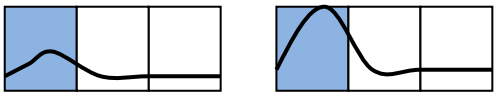
I'm *sure she'll *call you to *mor row.
H L

Estebas Vilaplana, E. (2019). Plataforma interactiva para el autoaprendizaje de la pronunciación inglesa: la enseñanza de la entonación. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 59–64). Málaga: Universidad de Málaga.

3.3. Niveles tonales

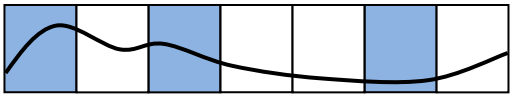
El sistema *ToBI* describe los movimientos melódicos mediante dos tonos, H (alto) y L (bajo). Mientras que los tonos bajos suelen tener unos valores de f_0 bastante constantes, los tonos H muestran más variabilidad y pueden aparecer con escalonamiento descendente (*downstepped*) o ascendente (*upstepped*), indicados con los signos exclamativos ! (!H) y ¡ (¡H) respectivamente. El sistema *TL_ToBI* ve necesario anotar la presencia de tonos altos con distinto escalonamiento pero, en lugar de usar los diacríticos de exclamación, incorpora un tono M (medio) por ser más transparente a la hora de entender su funcionalidad. Este tono ya se usó para el inglés en el modelo de Liberman (1975), y se ha incorporado en otros sistemas *ToBI* como el español (Beckman, Díaz Campos, McGory y Morgan, 2002). Además, el modelo británico de análisis entonativo incluye en su inventario de tonos categorías contrastivas como *high-fall* (descenso alto) o *low-fall* (descenso bajo) que, en cierta manera, vienen a reflejar la presencia de tres niveles tonales, ya que el *high-fall* equivaldría a HL y el *low-fall* a ML. En el sistema *TL_ToBI*, los tonos H, M y L indican tanto un acento tonal como un tono de frontera, como se aprecia en los ejemplos (4) y (5).

(4)



*Mel an ie. *Mel an ie.
M L H L

(5)

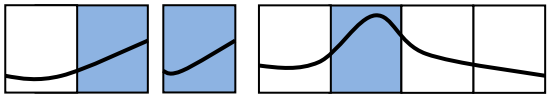


*Would you *like a ba *na na?
H H L M

3.4. Dominios prosódicos

El sistema *ToBI* distingue dos niveles de dominio prosódico, uno mayor (la *frase entonativa* a la que se asocian los tonos de frontera) y otro menor (la *frase intermedia* delimitada por los acentos de frases). A pesar del valor fonológico de esta distinción, *TL_ToBI* solo reconoce un único nivel prosódico, ya que para la enseñanza de la entonación lo más relevante es saber identificar el final de un grupo entonativo independientemente de su jerarquización. En este sentido, *TL_ToBI* se asemeja a la propuesta de la Escuela Británica que solo reconoce un único nivel de frase entonativa. Los estudiantes, por tanto, solo tienen que identificar el final de una frase prosódica delimitada por un tono de frontera. Véase el enunciado enumerativo de (6). Como se aprecia en este ejemplo, cuando una frase entonativa acaba con una palabra aguda, como en los dos primeros grupos entonativos del enunciado, el tono de frontera (en este caso, M) se sitúa en la misma sílaba que el acento tonal (en este caso, L).

(6)



She's *nice, *tall and *el e gant.
LM LM H L

3.5. Acentos nucleares y prenucleares

Una de las principales diferencias entre la Escuela Británica y los modelos americanos de análisis entonativo es la descripción del último acento tonal de un enunciado. En la tradición británica, el último acento tonal, también denominado *nuclear*, no únicamente indica la melodía de la sílaba acentuada sino que también especifica la trayectoria final del enunciado. En este modelo, por tanto, los tonos que aparecen en posición nuclear son distintos de las categorías tonales prenucleares ya que deben modelar el final de la frase entonativa. En el sistema *ToBI*, por el contrario, las categorías tonales asociadas a los acentos prenucleares son las mismas que las del último acento, ya que la melodía final de un enunciado se describe mediante los tonos de frontera.

Al igual que en el sistema *ToBI*, en la propuesta *TL_ToBI* se utilizan los mismos tonos tanto en la posición nuclear como en la prenuclear y se mantiene la presencia de tonos de frontera para describir el contorno final. Sin embargo, *TL_ToBI* incorpora dos diferencias relevantes en relación al sistema *ToBI* inicial:

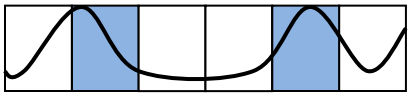
- los acentos tonales solo son monotonales (L, M, H), al contrario del sistema *ToBI* que incluye tanto acentos monotonales (L*, H*) como bitonales (L*+H, H*+L, etc.).
- los tonos de frontera pueden ser monotonales (L, M, H) y bitonales (HL y LH), a diferencia de la notación de *ToBI* que solo considera tonos de frontera monotonales (L%, H%).

La propuesta de usar solo acentos monotonales en las sílabas tónicas relevantes tiene relación con las diferencias en la carga semántica de los acentos prenucleares y los nucleares. A pesar de que los acentos prenucleares son importantes para determinar el tipo de contorno melódico, el peso semántico de la configuración nuclear es mucho mayor ya que es la responsable del significado final de un enunciado (véase Face, 2007). Por esta razón, se ha reducido el número de acentos tonales a tres (H, M, L) con el fin de que el alumno se centre en la configuración final del contorno al ser la más relevante en términos semánticos.

En el sistema *ToBI*, los tonos ubicados al final de un enunciado solo son monotonales ya que, al mantener dos niveles de fraseo prosódico, se distingue entre tonos de frontera (situados al final de una frase entonativa) y acentos de frase (asociados al final de una frase intermedia). Dado que en *TL_ToBI* solo se propone un único dominio fraseológico, se estima necesario incorporar tonos de frontera bitonales para describir los movimientos complejos del final de un enunciado, es decir, trayectorias de f_0 con más de un objetivo tonal,

como el *fall-rise* o el *rise-fall*, según la nomenclatura de la Escuela Británica. Los tonos de frontera bitonales se han usado para lo modelización de la entonación de lenguas como el español (Estebas-Vilaplana y Prieto, 2008). El enunciado presentado en el ejemplo (7) ilustra un contorno con un tono de frontera bitonal.

(7)



It's *ra ther ex *pen sive
H H LH

El método *TL_ToBI*, por tanto, simplifica y adapta la propuesta de notación prosódica del sistema *ToBI* con la finalidad de ayudar a los estudiantes de fonética inglesa a entender e interiorizar las curvas melódicas de los principales enunciados del inglés. A pesar de que *TL_ToBI* reduce muchos de los detalles fonéticos presentes en las categorías entonativas del sistema *ToBI* inicial, los tonos propuestos son capaces de describir los principales contrastes entonativos del inglés y encuentran una equivalencia con el inventario de tonos sugerido por la Escuela Británica, tal y como se expone en la Tabla 1.

En la plataforma *TYEP*, el método de anotación prosódica recogido en *TL_ToBI* se ha usado para describir la entonación neutra y marcada de cuatro tipos de frases: declarativas, interrogativas (absolutas, parciales y *tag-questions*), exclamativas e imperativas. La Figura 2 muestra un ejemplo del método *TL_ToBI* para presentar la entonación inglesa de frases declarativas neutras y de enunciados declarativos no neutros (con matices de enfado y de reservas).

Aunque la descripción de la entonación del inglés o de cualquier otra lengua es mucho más compleja que lo que puede llegar a representar cualquier modelo entonativo, la propuesta de *TL_ToBI* constituye un primer acercamiento a la entonación inglesa, destinado a cubrir las necesidades de un alumnado a distancia no especializado en temas fonéticos.

Tabla 1: Equivalencias entre el inventario de tonos de la Escuela Británica y el propuesto por *TL_ToBI*.

Escuela Británica	***TL_ ToBI***		**Escuela Británica**	***TL_ ToBI***
Tonos nucleares	Acentos tonales	Tonos de frontera	Tonos pre-nucleares	Acentos tonales
High-fall	H	L	*High*	H
Low-fall	M	L	*Low*	M o L
High-rise	M	H		
Low-rise	L	M		
Mid-level	M	M		
Fall-rise	H	L H		
Rise-fall	L	H L		

Figura 2: Ejemplo de anotación prosódica de frases declarativas neutras y marcadas según el método *TL-ToBI*.

4. EL SISTEMA *TL_ToBI* EN LA ENSEÑANZA DE LA ENTONACIÓN INGLESA A DISTANCIA

Para valorar la validez y el impacto del sistema *TL_ToBI* en la enseñanza de la entonación inglesa a distancia frente a la efectividad de los sistemas tradicionales, se llevó a cabo un estudio comparativo de las curvas entonativas del inglés producidas por dos grupos de estudiantes hispanohablantes de la asignatura "Pronunciación de la Lengua Inglesa" del *Grado en Estudios Ingleses: Lengua, Literatura y Cultura* de la Universidad Nacional de Educación a Distancia (UNED), correspondientes a los cursos 2012-2013 y 2013-2014, en los que se enseñó entonación mediante el sistema británico y el sistema *TL_ToBI* respectivamente (véase Estebas-Vilaplana, 2015, para más detalles). En ambos casos, los alumnos tuvieron que enfrentarse al aprendizaje de la entonación de manera autónoma pero con una metodología docente distinta (modelo británico *vs. TL_ToBI*). Para cada grupo de alumnos se examinaron tres parámetros: (1) la producción de los acentos prenucleares; (2) la producción del último acento tonal y del tono de frontera (configuración nuclear); y (3) la producción de todo el contorno entonativo (configuración nuclear y configuración prenuclear). Se analizó la entonación tanto neutra como marcada de los siguientes tipos de frases: declarativas, interrogativas absolutas, interrogativas parciales e imperativas.

Los resultados de este estudio demostraron una mejoría significativa en la producción de los contornos entonativos ingleses por parte del grupo de estudiantes hispanohablantes que usó *TL_ToBI* frente al grupo que trabajó con el modelo británico. Esta mejoría se observó tanto en el contorno completo como en la configuración prenuclear y en la configuración nuclear por separado. Los porcentajes de producción de las curvas melódicas con la entonación esperada se reproducen en la Tabla 2, que muestra, para cada uno de los parámetros, un incremento de alrededor de un 15 % de mejoría en la producción de los contornos entonativos del inglés por parte de los estudiantes que recibieron instrucción con *TL_ToBI*.

Los resultados según el tipo de frase y según el tipo de entonación (neutra o marcada) se reproducen en la

Tabla 2: Porcentaje de producción de las curvas melódicas inglesas por parte de los estudiantes hispanohablantes instruidos según las premisas de la Escuela Británica (EB) y las de *TL_ToBI*.

	Contorno completo		**Configuración Nuclear**		**Configuración Prenuclear**	
	EB	*TL_ ToBI*	EB	*TL_ ToBI*	EB	*TL_ ToBI*
%	52.2	65.1	57.8	73.1	53.3	67.1

Figura 3: Porcentaje de curvas melódicas con la entonación inglesa esperada producidas por los estudiantes hispanohablantes instruidos según la Escuela Británica (EB) y según *TL_ToBI* para los distintos tipos de frases (D: declarativa; IA- interrogativa absoluta; IP: interrogativa parcial; IM: imperativa) con lectura neutra (N) y marcada (M).

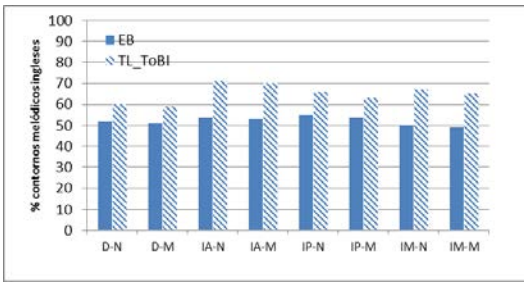


Figura 3, que muestra un incremento significativo en la producción adecuada de las curvas melódicas del inglés por parte de los estudiantes que han utilizado el sistema *TL_ToBI* frente a los alumnos que usaron el modelo británico. Sendos *t-tests* indicaron que las diferencias son significativas en todos los casos ($p < 0.01$).

Estos resultados corroboran la efectividad del método *TL_ToBI* para enseñar la entonación del inglés e indican que, en un contexto de enseñanza a distancia, un sistema más visual y más preciso en lo que respecta la alineación de las entidades tonales con la cadena segmental, como el que plantea *TL_ToBI*, favorece el aprendizaje de la entonación inglesa.

5. CONCLUSIONES

En este trabajo se ha presentado la plataforma interactiva *TYEP* diseñada para el autoaprendizaje de la pronunciación inglesa en un contexto de enseñanza a distancia y se ha hecho especial hincapié en la necesidad de usar un sistema de enseñanza de la entonación inglesa más visual e intuitivo que el que han estado utilizando los modelos tradicionales. Para facilitar la enseñanza de la entonación inglesa a distancia, este trabajo ha presentado un sistema de transcripción prosódica, denominado *TL_ToBI*, que supone una adaptación del sistema *ToBI* con fines docentes. Los resultados de un estudio donde se comparó la producción de las curvas entonativas inglesas por parte de dos grupos de estudiantes hispanohablantes instruidos según la Escuela Británica de entonación y el sistema *TL_ToBI* respectivamente, demostraron que los alumnos que siguieron el nuevo método consiguieron mejores resultados que los que trabajaron con el modelo tradicional. En un futuro es preciso validar la eficiencia del sistema *TL_ToBI* aplicándolo a otros de tipos frases, estilos y discursos más complejos. Asimismo, este método se pretende usar para la enseñanza de la entonación de otras lenguas, concretamente, para el autoaprendizaje de las curvas melódicas del español como lengua extranjera.

6. AGRADECIMIENTOS

Este trabajo ha sido subvencionado por el proyecto *Evaluación automática de la pronunciación del español como lengua extranjera para hablantes japoneses* (Junta de Castilla y León. Clave del proyecto: VA145U14).

7. REFERENCIAS

Beckman, M. y Ayers-Elam, G. (1997). Guidelines for ToBI labelling. Consultado en **¡Error! Referencia de hipervínculo no válida.**_ToBI/.

Beckman, M., Díaz Campos, M., McGory, J. T. y Morgan A. T. (2002). Intonation across Spanish in the Tones and Break Indices framework. *Probus, 14*, 9–36.

Cruttenden, A. (1986). *Intonation*. Cambridge: Cambridge University Press.

Estebas-Vilaplana, E. (2015). The learning of English intonation by Spanish speakers in a distance education environment. *Actas de Phonetics Teaching and Learning Conference 2015* (pp. 40–43). Consultado en https://www.ucl.ac.uk/pals/study/cpd/cpd-courses/ptlc/proceedings_2015/PTLC_2015_Estebas.pdf.

Estebas-Vilaplana, E. y Prieto, P. (2008). La notación prosódica del español: una revisión del Sp_ToBI. *Estudios de Fonética Experimental, 17*, 265–283.

Face, T. (2007). The role of intonational cues in the perception of declaratives and absolute interrogatives in Castilian Spanish. *Estudios de Fonética Experimental, 16*, 185–226.

Kissling, E. (2013). Teaching pronunciation: is explicit phonetics instruction beneficial for FL learners? *The Modern Language Journal, 97*, 720–744.

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford: Oxford University Press.

Ladd, R. (1996). *Intonational phonology.* Cambridge: Cambridge University Press.

Liberman, M. (1975). *The intonational system of English* (Tesis doctoral). Cambridge, MA: MIT.

O'Connor, J. D. (1967). *Better English pronunciation*. Cambridge: Cambridge University Press.

O'Connor, J. D. y Arnold, G. F. (1973). *Intonation of colloquial English*. London: Longman.

Pierrehumbert, J. (1980). *The Phonology and phonetics of English intonation*. Bloomington: Indiana University Press.

Wells, J. (2006). *English intonation: An introduction*. Cambridge: Cambridge University Press.

Subsidia: Tools and Resources for Speech Sciences

11

Dumloquor hora fugit: aprendizaje autónomo y autorregulado de la pronunciación del catalán a través de las *Guies de pronunciació del català*

Josefina Carrera-Sabaté[1], Jesús Bach Marqués[1] y Mar Mir Campillo[1]

[1] Universitat de Barcelona
e-mail: jcarrera@ub.edu

Citation / Cómo citar esta publicación: Carrera-Sabaté, J., Bach Marqués, J. y Mir Campillo, M. (2019). *Dumloquor hora fugit:* aprendizaje autónomo y autorregulado de la pronunciación del catalán a través de las *Guies de pronunciació del català*. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 65–73). Málaga: Universidad de Málaga.

RESUMEN: El aprendizaje y enseñanza de lenguas ha sufrido un cambio acorde con la sociedad, y actualmente el uso de las TIC está en boga ya que permiten llevar a cabo un aprendizaje autónomo y autorregulado. Dada la escasez de recursos TIC existentes para aprender la pronunciación del catalán de manera autónoma, se crearon las *Guies de pronunciació del català* (http://www.guiesdepronunciacio.cat) con el objetivo de ofrecer a los estudiantes cuya L1 no es el catalán la posibilidad de mejorar su pronunciación a partir de ejercicios basados en los elementos segmentales y suprasegmentales del catalán. Esta propuesta, que se fundamenta tanto en aspectos lingüísticos como en aspectos comunicativos y hace uso de materiales procedentes de la cultura catalana, está pensada para que cada usuario escoja un itinerario concreto en función de su L1. Las acciones de innovación docente que se han apoyado en los materiales de la web permiten confirmar que estos resultan muy útiles para los usuarios interesados en la mejora de la pronunciación del catalán.

Palabras clave: pronunciación; prosodia; aprendizaje de segundas lenguas; aprendizaje autónomo; e-learning; catalán.

ABSTRACT: Learning and teaching of languages has suffered a change according to society. Nowadays, the use of ICT is trendy because it allows learners to take an autonomous and self-regulated learning. Due to the lack of existent ICT resources for learning Catalan pronunciation autonomously, the site *Guies de pronunciació del català* (http://www.guiesdepronunciacio.cat) was created. Its main aim is to offer students whose L1 is not Catalan the possibility of improving their pronunciation by doing exercises based on the segmental and suprasegmental elements of Catalan. This proposal, which is based both on linguistic and communicative aspects and makes use of materials from the Catalan culture, has been designed so that each learner can choose a specific itinerary according to his L1. The innovative teaching actions which have sought for support in the web materials permit to confirm that they are so useful for those users interested in improving the pronunciation of Catalan.

Keywords: pronunciation; prosody; second language learning; autonomous learning; e-learning; Catalan.

1. INTRODUCCIÓN

1.1. Nuevos paradigmas en aprendizajes de lenguas

El aprendizaje de lenguas ha conllevado una transición importante, acorde con la evolución de la sociedad: se ha pasado de los tradicionales intercambios cara a cara en clase a los actuales intercambios online con el uso de las TIC (Murphy, 2015). Las TIC, pues, han permitido la aparición de una gran cantidad de recursos para aprender lenguas de una manera autónoma y autorregulada. Actualmente, existen dos tipos de propuestas e-learning, según Clark y Mayer (2016, p. 8): (1) las que están diseñadas para un aprendizaje bajo demanda (e-learning asincrónico) y (2) las que están diseñadas para un público abierto que puede seguir clases virtuales desde cualquier parte del mundo (e-learning sincrónico). Además, las dos modalidades pueden contener propuestas informativas (inform courses) y propuestas prácticas (perform courses).

1.2. Recursos tecnológicos para el aprendizaje de la pronunciación de segundas lenguas[1]

Los recursos relacionados con la enseñanza de idiomas son especialmente atractivos entre los usuarios de segundas lenguas. Algunos especialistas incluso destacan su carácter claramente estimulante, así como su utilidad en ausencia de inmersión lingüística o en casos en que los usuarios presentan problemas individuales (Campillos, 2010). Sin embargo, estos recursos se entienden como una ayuda para el aprendizaje, pues se estima necesaria la supervisión de un tutor durante el aprendizaje, ya que el conocimiento y la visión que aporta no pueden ser suplidos por las tecnologías del habla (Gil, 2007).

Los recursos o tecnologías del habla destinadas a la enseñanza de lenguas se pueden dividir en dos grupos: En primer lugar, aquellos que se basan en contenido visual, es decir, que utilizan oscilogramas, espectrogramas, curvas melódicas —datos sobre la onda de sonido— o material audiovisual relacionado con los órganos fonadores y sus movimientos y posiciones en el proceso de articulación. Estos últimos son relevantes para el aprendizaje de la lengua, especialmente por lo que respecta al movimiento de los labios (resulta ser crucial para cualquier aprendizaje). Hay que tener en cuenta, además, que la naturalidad es especialmente importante en el habla de las personas visualizadas en cualquier interfaz.

En segundo lugar, existen recursos que se utilizan para la grabación, reconocimiento, síntesis y reproducción de la voz. Las tecnologías de grabación y reproducción de la voz, que aparecieron a mediados del siglo XX, permitían al usuario grabarse hablando, reproducir lo que había dicho y compararlo con los modelos que ofrecían estos mismos recursos para poder mejorar su pronunciación y entonación (Llisterri, 2006). Actualmente, existen algunas propuestas de reconocimiento de voz que permiten un feedback automatizado a través de móviles para mejorar la pronunciación (Cavus, 2016), aunque el reconocimiento automático del habla todavía le falta camino para ser utilizado como procedimiento de corrección automática de errores de pronunciación (Witt, 2012).

1.3. Folklore musical y literario para pronunciar segundas lenguas

En el aprendizaje de cualquier lengua es fundamental el conocimiento de su folklore musical y literario, porque en él se refleja la esencia de la comunidad de dicha lengua, caracterizada por una cultura, costumbres, valores, creencias y forma de pensamiento propias. Ser conocedor e incluso partícipe de todo este bagaje sociocultural puede facilitar, en mayor o menor grado, la comprensión, aprendizaje y comunicación en la segunda lengua.

Utilizar el folklore musical y literario para el aprendizaje de segundas lenguas puede ser de gran utilidad para aproximar el ámbito sociocultural de la lengua al hablante, para motivarlo y, a la vez, trabajar la fonética de la lengua con materiales significativos. Por otro lado, como observa Álvarez (2010), hay una serie de recursos memorísticos que son intrínsecos a la literatura popular y folklórica, como repeticiones o acumulaciones de encabalgamientos, los cuales favorecen el aprendizaje tanto de la lengua en sí como de su pronunciación, ritmo y entonación. De igual manera actúa la canción: la música, gracias al ritmo, musicalidad o rima de las cadencias, la melodía y otros recursos que se le asocian, facilitan la adquisición de la fonética de una segunda lengua, como también del resto de propiedades y características de dicha lengua (ritmo, léxico, sintaxis, etc.). En este sentido, son numerosas las propuestas que hacen uso de poemas, canciones o historias para poder ayudar a los debutantes de una lengua. Un ejemplo se encuentra en las propuestas para aprender inglés del British Council: http://learnenglish.britishcouncil.org/en/stories-poems?utm_source=facebook&utm_medium=social&utm_campaign=bc-learnenglish).

1.4. Aportaciones de la inteligencia emocional y la gamificación a los aprendizajes de segundas lenguas

Es sabido que en todos los individuos coexisten dos tipos de inteligencia: la cognitiva y la emocional. En todos los aprendizajes, las emociones juegan un papel muy relevante: hay estados anímicos que favorecen los aprendizajes (alegría, entusiasmo) y otros que los obstaculizan (tristeza, miedo, rabia). La emoción no se puede separar de los procesos cognitivos, puesto que la información percibida a través de los sentidos pasa por el cerebro emocional (sistema límbico) antes de ser procesada por la corteza cerebral. Según Bueno (2015), cuando la parte emocional del cerebro está activada, el aprendizaje es mucho más completo y cuando los aprendizajes presentan una carga emocional, quedan mejor almacenados en la memoria y se pueden recuperar y utilizar más eficientemente.

Además, según Damasio (2005), la emoción, el pensamiento y el cuerpo están interrelacionados a través de circuitos bioquímicos y neurales, de manera que todo lo que ocurre en una dimensión afecta a las otras. Teniendo esto en cuenta, se puede incidir sobre las emociones con acciones concretas sobre el propio cuerpo; así pues, cambiando la actitud corporal a través de la respiración consciente, la relajación o la visualización, se puede influir directamente sobre el estado de ánimo (Darder, 2013).

Por otro lado, y dado que el juego es una característica inherente a la especie humana, algunos estudios recientes han demostrado que los aprendizajes a través de juegos son muy efectivos para captar la atención del alumno y despertar su motivación. Lee (2011) destaca los beneficios cognitivos, emocionales y sociales que implican las diferentes técnicas de gamificación existentes. Ya Piaget (1986), desde la

[1]En esta propuesta utilizamos indistintamente los términos segunda lengua (L2) y lengua extranjera (LE) para referirnos a la lengua no nativa, sin entrar en los contextos de aprendizaje de una lengua no nativa.

psicología cognitiva, concede al juego un lugar predominante en los procesos de desarrollo y relaciona el desarrollo de los estadios cognitivos con el desarrollo de la actividad lúdica. Huizinga (1987), desde el punto de vista antropológico, concibe el juego como una función humana tan esencial como la reflexión o el trabajo y lo entiende como una cualidad intrínsecamente motivadora.

Relacionado con lo dicho, la gamificación es conocida como el uso de dinámicas mecánicas y estéticas propias del juego en entornos no lúdicos para adquirir, desarrollar o mejorar un comportamiento determinado (Kapp, 2012). El uso del juego en el ámbito de la docencia no es nuevo. Hay una gran diversidad de experiencias que confirman la potencialidad del juego como una estrategia bajo la cual profesorado y formadores diseñen y usen un gran número de técnicas, recursos y dinámicas vinculados al juego (Hays, 2005; Kapp, 2012; Randel, Morris, Wetzel y Whitehill, 1992; Sitzmann, 2011). Así pues, es beneficioso despertar las emociones al inicio, durante y al final de los aprendizajes, ya que son los motores del juego y crean un ambiente propicio para que se produzca el aprendizaje. Cada juego puede potenciar un tipo determinado de aprendizaje —de conceptos, habilidades, actitudes, o todo a la vez— con el fin de ayudar a resolver problemas y poder alcanzar un aprendizaje significativo.

1.5. La prosodia en la enseñanza de segundas lenguas

Los estudios sobre la adquisición de segundas lenguas corroboran la importancia de un abordaje centrado en la prosodia y en los elementos que están relacionados con esta, como la pragmática, de acuerdo con la dificultad que parece conllevar el aprendizaje de la prosodia en segundas lenguas (Delmonte, 1988). Esta dificultad está directamente conectada a la diversidad de parámetros que influyen en la prosodia —las pausas, el volumen, la cualidad de la voz, etc.—, a diferencia de lo que pasa con la pronunciación de segmentos, que resulta mucho más sencilla.

Por otro lado, los resultados de investigaciones recientes relacionadas con el aprendizaje de la prosodia subrayan que las tareas de percepción auditiva y, sobretodo, visual son eficaces para aprenderla, especialmente si cuentan con el refuerzo de la gestualidad; además, la contextualización pragmática es otro elemento fundamental para abordar la prosodia (Hurley, 1992).

1.6. Recursos para el aprendizaje del componente fónico de la lengua catalana

Recientemente, el interés por el aprendizaje de la lengua catalana ha crecido y se ha globalizado, en cierto modo. A raíz de este hecho, han aparecido y se han creado nuevos materiales para aprender catalán desde otras lenguas. Para empezar, el Departament de Benestar i Família de la Generalitat de Catalunya ofrece unas publicaciones comparativas de la gramática del catalán con las de una considerable diversidad de lenguas, entre ellas el chino, el panyabí o el wólof. En relación a la fonética, se han realizado trabajos sobre producción y percepción en comunidades plurilingües, y, además, estudios sobre la producción de sonidos en estudiantes universitarios extranjeros que son usuarios de catalán (Creus y Julià-Muné, 2010).

Por otro lado, los materiales y recursos e-learning existentes en el mercado para aprender la pronunciación del catalán son escasos. Algunos ejemplos se encuentran en los libros de Dalmau, Miró y Molina (1985) y de Bau, Pujol y Rius (2007), la guía de corrección fonética de Biblioni (biblioni.cat/correcciofonetica) o los proyectos *Pronunciem* de Pi (2008) y GALÍ (http://clic.xtec.cat/gali).

A raíz de este estado de la cuestión se han creado dos interfaces con el propósito de ofrecer materiales para facilitar el aprendizaje del componente fónico de la lengua catalana:

(1) El año 2006, *Els sons del català*, (http://www.ub.edu/sonscatala/ca/), de la mano de Solà, Carrera-Sabaté y Pons (Carrera-Sabaté, 2012). Esta herramienta, imprescindible para aquellos que quieren adentrarse en el conocimiento de la fonética catalana, presenta los sonidos del catalán de manera descriptiva y desde diferentes ámbitos de estudio (articulación, acústica, transcripción fonética, comparación, etc.). Este sitio web no pretende proporcionar pautas para mejorar la pronunciación del catalán.

(2) El año 2010, atendiendo a las peticiones de diversos usuarios de la página *Els sons del català*, y de acuerdo con las necesidades de los alumnos interesados en aprender la pronunciación del catalán, el grupo de innovación docente FONCAT de la Universidad de Barcelona empezó la creación de una nueva web a fin de ofrecer una serie de recursos que pudieran satisfacer las necesidades mencionadas. En sus inicios, este portal web, *Guies de pronunciació del català* (GPC) (http://www.ub.edu/guiesdepronunciacio/) [2], que partió de una propuesta de Carrera-Sabaté (en prensa), pretendía ser una ayuda para castellanohablantes peninsulares que aprendieran catalán.

2. *GUIES DE PRONUNCIACIÓ DEL CATALÀ*: OBJETIVOS, ESTRUCTURA Y FUNCIONAMIENTO

El portal GPC (*Guies de pronunciació del català*; http://www.ub.edu/guiesdepronunciacio), como ya

[2] El grupo lo formaban entonces Maria Cabrera, Imma Creus, Ana M. Fernández, Roser Güell, Joan Julià-Muné, M. Rosa Lloret, Clàudia Pons, Gemma Reguant, Paolo Roseano y Josefina Carrera-Sabaté (IP). También se obtuvo la colaboración externa de Joan Borràs, Dolors Font, Pilar Prieto, Agnès Rius o Francina Torres. El proyecto también ha contado con varios becarios durante estos años, como Jesús Bach, Enric Blanco, Eva Bosch, Rosanna Costantini, Lea Feliu, Mar Mir y Josep Pons, y además con varios estudiantes voluntarios de los últimos cursos del grado de Filología Catalana de la Universidad de Barcelona: Laia Benavent, Marta Busquets, Anna Costa, Leia Jiménez, Sara Mortreux, Meritxell Sabaté y Montserrat Sala. En el proyecto colaboran profesores de tres universidades (UB, UdL y UPF) y del Institut del Teatre de Barcelona, y ha recibido financiación de la UB (PMID y Redice), de la UPF y de La Caixa.

Carrera-Sabaté, J., Bach Marqués, J. y Mir Campillo, M. (2019). *Dumloquor hora fugit:* aprendizaje autónomo y autorregulado de la pronunciación del catalán a través de las *Guies de pronunciació del català*. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 65–73). Málaga: Universidad de Málaga.

hemos visto, tiene como objetivo presentar recursos para mejorar la pronunciación del catalán a partir de un conjunto de materiales organizados por itinerarios lingüísticos de origen, según la L1 de los usuarios y teniendo en cuenta el componente prosódico de la lengua y la cultura catalana popular. En este momento, solo está disponible el itinerario para hablantes de español peninsular, pero ya se están terminando otros, como los que irán dirigidos a hablantes de español de América y de inglés (británico y americano). En cuanto a la forma, se presentan tanto aspectos segmentales como suprasegmentales, y se ofrecen propuestas con distintos niveles de dificultad para dar respuesta a necesidades derivadas de los distintos niveles de conocimiento del catalán de los usuarios.

La página web contiene un menú lateral con los distintos apartados de la misma, que se detallarán a continuación. En la página de acceso se ofrece una breve descripción de los contenidos de la interfaz y se presenta la estructura del portal, así como aspectos de gestión y financiación (Figura 1). El segundo apartado es "*equip / equipo*", donde se detallan los profesores, profesionales de la lengua y asesores lingüísticos, músicos, actores y becarios que han colaborado o colaboran con el proyecto.

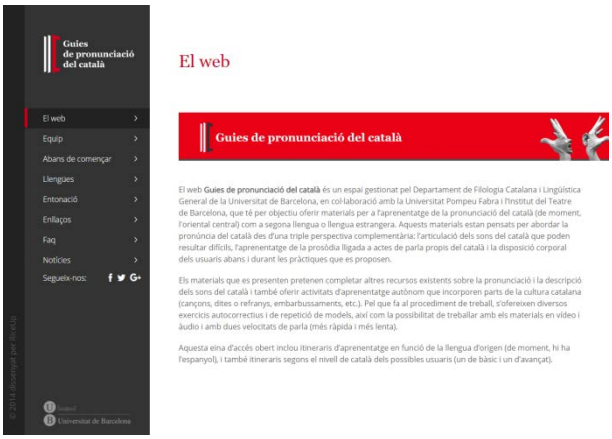
Las secciones que siguen son "abans de començar / antes de empezar", "llengües / lenguas" y "entonació / entonación", pero se tratarán con más precisión y profundidad más adelante. En el apartado "enllaços / enlaces" los usuarios pueden encontrar gran parte de los recursos web que hay actualmente en el mercado y que están relacionados con la pronunciación y la prosodia del catalán o de alguna otra lengua. La siguiente sección está en construcción, y albergará un conjunto de explicaciones teóricas sobre la selección de los contenidos disponibles en el portal. Finalmente, en el apartado "notícies / noticias" el usuario puede encontrar información sobre eventos, congresos, jornadas o talleres que tienen que ver con las GPC. Esta interfaz tiene presencia y hace divulgación de sus actividades en varias redes sociales, a las que se puede acceder desde los respectivos iconos dispuestos al pie del menú lateral.

Los apartados que se comentarán a continuación, abans de començar / antes de empezar", "llengües / lenguas" y "entonació / entonación", son los tres que recogen los materiales destinados a la práctica de la pronunciación y la prosodia del catalán.

2.1. "*Abans de començar*": ejercicios de sensibilización auditiva

El apartado "*abans de començar*" (http://www.guiesdepronunciacio.cat/abans-de-començar) ofrece al usuario unos ejercicios previos a la práctica de la pronunciación. Se trata de una propuesta de ejercicios de sensibilización auditiva y relajación corporal diseñada por Gemma Reguant, profesora del Institut del Teatre de Barcelona, con la finalidad de preparar el cuerpo para el aprendizaje posterior (apartado 1.4).

Figura 1: Página principal.



El oído a menudo recurre a mecanismos de desensibilización para protegerse del ruido o de la contaminación acústica exterior. Además, el estrés o la ansiedad perjudican la atención y, a su vez, la sensibilidad auditiva. Es por eso que no solo se trabaja con el oído, sino también con los otros sentidos, a fin de relajar todo el cuerpo y, especialmente, hacer que el oído sea más sensible a la percepción y producción de los ejercicios de práctica que se proponen para mejorar la pronunciación del catalán. En este apartado el usuario puede encontrar una explicación teórica del porqué del ejercicio, las instrucciones y los audios necesarios para llevarlo a cabo y una relación de los posibles efectos que se puede experimentar después de terminar la relajación (Figura 2).

2.2. "*Llengües*": itinerarios según la lengua de origen

En esta sección el usuario puede encontrar distintos itinerarios de aprendizaje de la pronunciación del catalán según su lengua de origen (http://www.guiesdepronunciacio.cat/catala-espanyol). Hoy por hoy solo está disponible el itinerario español peninsular-catalán (Figura 3), pero los que parten del español de América o del inglés (habrá un itinerario para el inglés americano y otro para el británico) ya están en proceso de implementación.

Cuando el usuario entra en un itinerario concreto, lo primero que encuentra es una tabla fonética para los sonidos consonánticos y otra para los sonidos vocálicos. Cada símbolo fonético tiene unas correspondencias ortográficas con palabras para poder asociar cada sonido con su símbolo fonético. Además de estar representados todos los sonidos de ambas lenguas (lengua meta y lengua origen), los sonidos aparecen coloreados siguiendo un criterio inspirado en los colores de un semáforo. Así pues, el verde indica que el sonido se encuentra en las dos lenguas y presenta la misma distribución o aparece en contextos similares; el naranja indica que el sonido aparece en las dos lenguas, pero la distribución o contextos donde aparece son distintos; finalmente, el rojo indica que el sonido es propio del catalán pero no de la L1 del usuario. Además, hay algunos sonidos que están de color azul: son aquellos

Figura 2: Apartado "*abans de començar*".

que solo se encuentran en la L1 del usuario y no forman parte del inventario de fonemas del catalán.

Los recuadros en verde, es decir, los de los sonidos propios de ambas lenguas y con igual distribución, están enlazados a las páginas correspondientes del portal *Els sons del català*, donde se pueden consultar sus características articulatorias y acústicas. En último lugar, los recuadros en naranja o rojo, que son aquellos con los sonidos que pueden presentar más dificultades a los usuarios, contienen un enlace que los dirige a una nueva página, donde hay una gran variedad de propuestas para pronunciar el sonido seleccionado. En estas propuestas podemos encontrar dos tipos de material multimedia: vídeos y audios de las palabras o construcciones que se van a trabajar. La finalidad de este doble recurso es que los usuarios puedan escoger entre solo escuchar o escuchar y visualizar los materiales que se proponen, atendiendo a la manera como aprenden mejor (auditiva / visual).

Según la dificultad o las características del fragmento pronunciado, a veces podemos encontrar, además del archivo de audio, hasta dos tipos distintos de grabación audiovisual: vídeos con un plano detalle que enfocan solo la boca y los labios del hablante y vídeos con un plano medio, que enfocan todo el torso del hablante. Estos segundos suelen aparecer en actividades con un carácter más comunicativo, como las relativas a frases, trabalenguas, refranes de la lengua o canciones, con la voluntad de que el usuario pueda observar también los movimientos corporales y la comunicación no-verbal que acompañan a las propuestas. Por otro lado, el asistente de reproducción de los vídeos permite al usuario modificar el volumen de reproducción, ampliar la imagen a pantalla completa y ralentizar los vídeos con la opción "*ralentitza / ralentiza*" (Figura 4). Esta opción posibilita que el usuario pueda ralentizar los vídeos y observar con más claridad la posición y el movimiento de la boca y de los labios, cosa que es de utilidad cuando se aprende una lengua extranjera. Este recurso que ofrece el asistente de reproducción permite que los usuarios puedan imitar y llegar más fácilmente a una pronunciación similar a la proporcionada en el modelo. Por otro lado, todos los ejercicios relacionados con los segmentos tienen dos iconos al lado de los vídeos: un micrófono y un reproductor. Estos símbolos pueden utilizarse en tiempo real para que el usuario pueda grabarse y reproducir sus producciones lingüísticas a fin de compararlas con las del modelo (Figura 5).

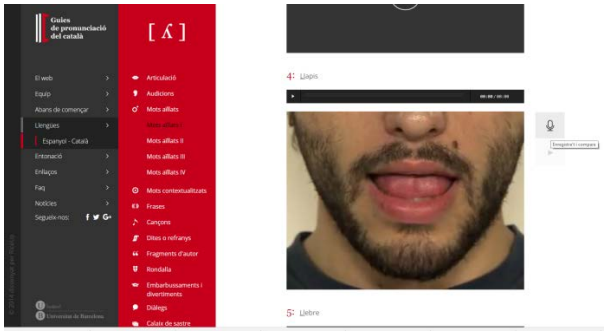
Todos los vídeos han sido grabados por actores y exalumnos del Institut del Teatre y por becarios de la UB, siempre bajo la supervisión de la profesora del Institut del Teatre que colabora con el proyecto, Roser Güell.

Los ejercicios que se proponen para cada sonido (marcado previamente en rojo y naranja) en las GPC se clasifican en varios subapartados según sus características formales o su grado de dificultad. Las propuestas son totalmente independientes entre ellas pero contienen una gradación de "abajo a arriba" para poder ofrecer una visión global de cada sonido supuestamente difícil: desde la articulación aislada a los ajustes articulatorios pasando por aspectos rítmicos y prosódicos, aspectos que se trabajan de forma integrada a través del folklore catalán.

Figura 3: Apartado "*llengües*". Itinerario español peninsular-catalán.

Figura 4: Detalle del botón "*ralentitza*" del reproductor.

Figura 5: Detalle del botón de grabación del reproductor.



Carrera-Sabaté, J., Bach Marqués, J. y Mir Campillo, M. (2019). *Dumloquor hora fugit:* aprendizaje autónomo y autorregulado de la pronunciación del catalán a través de las *Guies de pronunciació del català*. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 65–73). Málaga: Universidad de Málaga.

La primera entrada es "*articulació / articulación*", que contiene hasta tres tipologías distintas de ejercicios. El primer ejercicio es de articulación; a partir de explicaciones escritas y de material audiovisual se muestra a los usuarios cómo pronunciar el sonido que se está practicando. A continuación aparecen ejercicios de contraste fonético, en los que el alumno puede escuchar varias palabras con el sonido a trabajar y compararlas con otras similares o con las que constituyen pares mínimos. Se pretende así que los usuarios se familiaricen con la distinción y pronunciación de sonidos semejantes. Finalmente, hay un sencillo ejercicio de audición con palabras de gran simplicidad que contienen el sonido a trabajar.

La segunda sección es "*audicions / audiciones*". Presenta unos ejercicios autocorrectivos de audiciones que están preparadas para una tarea de discriminación del tipo ABX, con el sonido percibido en diferentes posiciones silábicas y acentuales, tanto en palabras aisladas como contextualizadas (Figura 6).

La sección contiene cuatro niveles distintos según el grado de dificultad de la tarea de discriminación y de las palabras y también según la frecuencia de uso que tienen en catalán. Los ejercicios propuestos en esta sección (Figura 5) están pensados para que los sonidos sean discriminados auditivamente (sin embargo, los usuarios también tienen acceso a los vídeos, que pueden reproducir si lo precisan). A la derecha de cada ejercicio aparecen varias opciones de sonidos, y los usuarios tienen que marcar cuál creen que es el que contiene la palabra que acaban de oír o ver. Como hemos indicado, estos son ejercicios autocorrectivos, por lo que, si la respuesta es correcta, el sonido queda marcado de color verde; si es incorrecta, de color rojo.

Una característica destacable de esta actividad de audición es que los usuarios, de entrada, no pueden visualizar la palabra escrita que se pronuncia en el material audiovisual. Así se evita que la grafía facilite o interfiera en la discriminación de los sonidos. Sin embargo, los usuarios tienen disponible la opción "*mostra la paraula / muestra la palabra*" para visualizar la grafía de la palabra que se ha escuchado.

Las tres secciones que siguen son "*mots aïllats / palabras aisladas*", "*mots contextualitzats / palabras contextualizadas*" y "*frases*". En cada sección se va introduciendo el sonido de interés en contextos cada vez más largos y complejos. En "*mots aïllats / palabras aisladas*" el usuario encuentra palabras aisladas con el sonido a pronunciar en diferentes posiciones contextuales, silábicas y acentuales y siguiendo cuatro divisiones internas. En estas divisiones las palabras están ordenadas de menos a más longitud silábica, complejidad articulatoria y frecuencia de uso, para que se puedan escuchar y repetir aisladamente según las necesidades de los usuarios. En segundo lugar, en "*mots contextualitzats*" hay dos divisiones internas que siguen un criterio similar al de la sección anterior. Aquí hay ejemplos en que se pueden observar fenómenos de asimilación de rasgos fonéticos entre sonidos de distintas palabras, que a veces pueden afectar al sonido que se practica. Finalmente, "*frases*" es una sección única con oraciones completas que contienen uno o varios vocablos con el fonema a trabajar (Figura 7). Para todas estas secciones, el portal ofrece la posibilidad de que el usuario se grabe a sí mismo y después reproduzca su articulación.

A continuación hay una serie de secciones que están desarrolladas atendiendo a las ventajas que ofrece el folklore, la tradición y la cultura popular en el aprendizaje de lenguas (véase el apartado 1.3.). La sección "*cançons / canciones*" contiene canciones tradicionales catalanas con una amplia presencia de los sonidos que presentan problemática. Las canciones tradicionales tienen acompañamiento de piano con armonizaciones poco convencionales que suelen atraer al usuario (Figura 8).

Los otros apartados contienen citas célebres de autores conocidos, dichos y refranes, rondallas, trabalenguas y otros juegos orales que nos vienen de la tradición (Figura 9). La musicalidad asociada a la mayoría de estos géneros así como otras de sus características intrínsecas permiten que el usuario pueda aprender los sonidos que presentan dificultades con más facilidad, en contextos más amplios y con más naturalidad, al mismo tiempo que se dan a conocer determinados aspectos de la cultura y la literatura catalanas. La mayoría de estas secciones incluyen audio y dos vídeos (plano detalle de los labios y plano medio con la posibilidad de ralentizar la velocidad) por cada ejercicio, y todas ellas permiten la grabación y posterior reproducción de la voz del usuario, que puede repetir las construcciones propuestas tantas veces como quiera.

Figura 6: Ejemplo de ejercicio de audición.

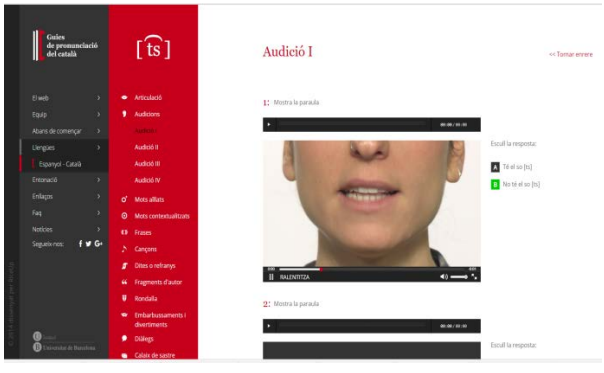


Figura 7: Ejemplo de práctica con frases.

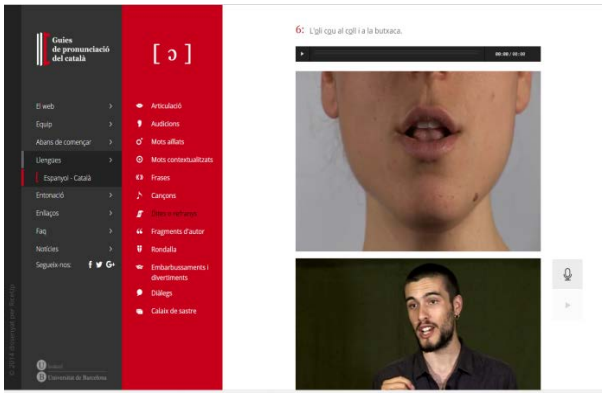
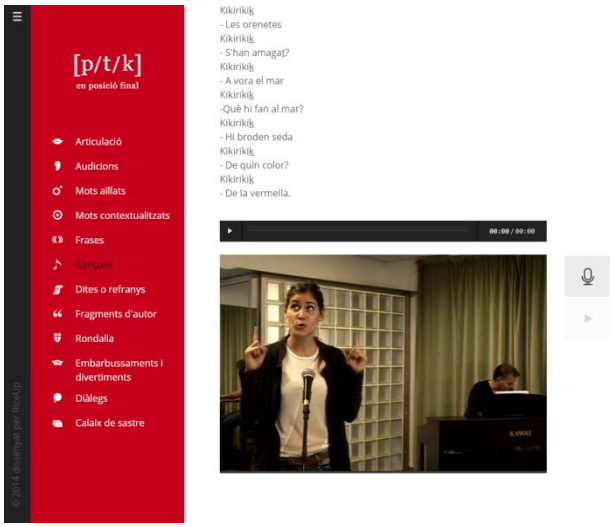


Figura 8: Ejemplo de canción.



Después, en el apartado "*diàlegs / diálogos*", el usuario puede encontrar diálogos de fragmentos de la serie *La Riera* de TV3. Del mismo modo que en los ejercicios anteriores, el usuario puede practicar los sonidos problemáticos en un contexto de más naturalidad comunicativa.

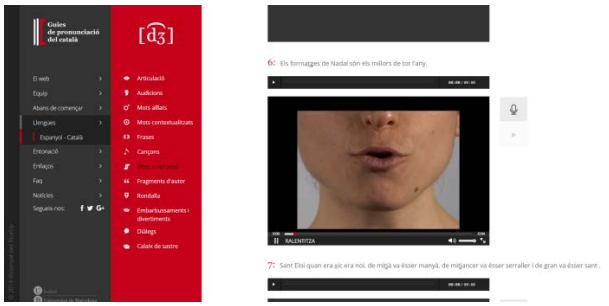
Finalmente, el apartado "*vols jugar / quieres jugar*" (a punto de ser implementado) está dedicado al aspecto del juego y la gamificación. Por ello, se han creado unos juegos de discriminación auditiva con diferentes niveles de dificultad que simulan los viajes que los usuarios hacen por el mundo con el equipaje que necesitan para poder realizarlos.

2.3. "*Entonació / entonación*": ejercicios de prosodia del catalán

Esta sección ha sido desarrollada por investigadores del *Grupd'Estudis de Prosòdia* (GrEP) de la Universitat Pompeu Fabra[3]. Se ha incluido en el portal de acuerdo con los estudios que demuestran la relevancia de la entonación en el aprendizaje de segundas lenguas (véase el apartado 1.5). El punto de partida del abordaje de la prosodia son tres divisiones de los hechos de habla: aserciones, peticiones y preguntas (Figura 10). A su vez, cada uno de ellos contiene varias secciones. Así pues, las aserciones pueden ser neutras, exclamaciones, obviedades, dudas o correcciones; las peticiones, órdenes o ruegos, y las preguntas, totales —neutra, de sorpresa, de confirmación o de repetición— o parciales —neutra, de sorpresa, de ruego, de orden o de repetición—(http://www.guiesdepronunciacio.cat/entonacio).

Para trabajar la entonación, el usuario puede encontrar vídeos de actores pronunciando frases. Los vídeos van acompañados de la curva de entonación, que va cambiando de color a medida que se reproduce el vídeo, en tiempo real (Figura 11).

Figura 9: Ejemplo de práctica con refranes.

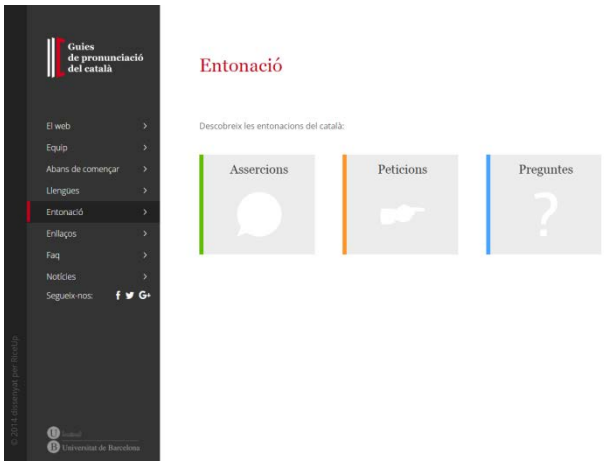


Los hechos de habla se presentan en cuatro propuestas distintas: "*1. Com sona? Mira i escolta / ¿Cómo suena? Mira y escucha*", "*2. Practica*", "*3. Igual o diferent? / ¿Igual o diferente?*" y "*4. Adequat o no? / ¿Adecuado o no?*". El primer apartado es un ejercicio de descripción: el usuario escucha y pone atención a la entonación de las frases en un contexto determinado. El segundo es de repetición: se trata de que el usuario repita la misma entonación que se reproduce en el vídeo. El tercero se basa en la comparación y discriminación de pronuncias: el usuario debe marcar en la plataforma si las dos entonaciones propuestas son iguales o diferentes. Finalmente, el último ejercicio consiste en escuchar ciertas entonaciones y decidir si son adecuadas a los contextos que se proponen o no. Los dos últimos son ejercicios que se pueden llevar a cabo en el mismo sitio web y son autocorrectivos (Figura 12).

3. APLICACIÓN PRÁCTICA DE LAS GUIES DE PRONUNCIACIÓ DEL CATALÀ

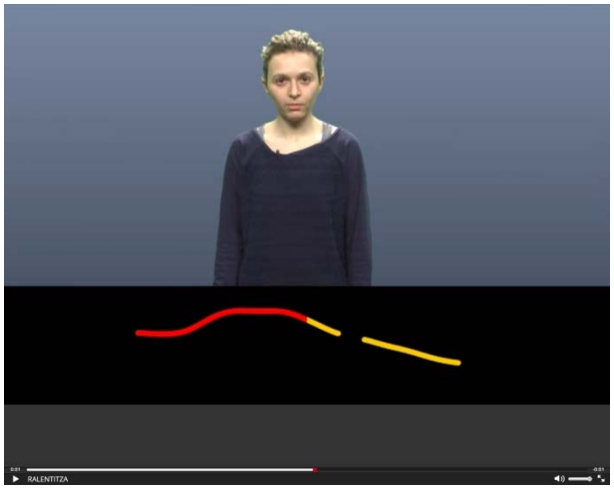
Como acabamos de explicar, las GPC ofrecen una amplia variedad de recursos para poder mejorar la pronunciación de los sonidos más problemáticos para aquellos hablantes cuya L1 no es el catalán, aunque también pueden ser de gran utilidad para catalanohablantes que tengan dificultades con la articulación de algunos sonidos concretos.

Figura 10: Página sobre entonación.



[3]Joan Borràs-Comes, Pilar Prieto y Paolo Roseano

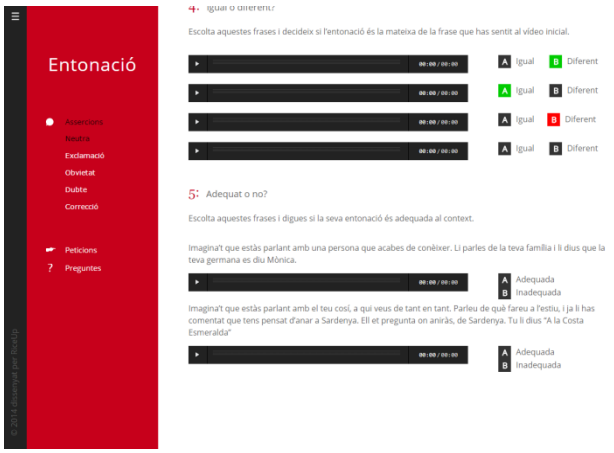
Figura 11: Curva de entonación cambiando de color a tiempo real.



Teniendo en cuenta la relevancia de la aplicabilidad del trabajo en el mundo educativo, los materiales han sido utilizados en diferentes acciones de innovación docente para ayudar a alumnos de diferentes grados a mejorar su pronunciación (Carrera-Sabaté, Bach, y Mir, 2016; Carrera-Sabaté, Creus, Fernández, Blanco, y Pons, 2014; Carrera-Sabaté, Valls, Royo, y Lloret, 2016) en asignaturas donde se trabaja la lengua oral.

Después de observar alumnos de procedencias lingüísticas diversas que cursan asignaturas de lengua oral, se detectan interferencias de otras lenguas en la pronunciación del catalán que resultan muy problemáticas, hasta el punto de comprometer la fluidez dela expresión oral. Normalmente, las carencias detectadas en los alumnos que hablan catalán son aspectos de interferencia de la fonética de la L1 en la del catalán y también aspectos relacionados con problemas de desarrollo en la adquisición de sonidos de la L1 (este es el caso, por ejemplo, de la lateral palatal en catalanohablantes que no presentan ningún otro problema de pronunciación). Las dificultades en la pronunciación de los alumnos pueden presentarse en distintos grados: pocos sonidos con dificultad para ser pronunciados —2 o 3— o dificultades en bastantes sonidos —7 u 8—.

Figura 12: Ejemplo de ejercicios de entonación.



En todas las experiencias se ha puesto de manifiesto que el uso de las GPC facilita el aprendizaje de la pronunciación. Seguramente esto se ve facilitado por contener materiales que abordan la comunicación desde "arriba a abajo", como propone Gil (2007, p. 13), que integran la prosodia en la pronunciación dentro de un marco comunicativo global.

Al margen de estos aspectos, merece la pena considerar otros factores que están conectados con la mejora de la pronunciación delos alumnos observados: efectivamente, la cantidad de pronunciaciones no propias del catalán por alumno y las características fónicas de estas, junto con la cantidad de horas invertidas en el trabajo autónomo de las GPC y también los aspectos relacionados con la propia subjetividad de los alumnos, condicionan su rendimiento. De esta manera, el progreso final entre ellos, como ya se había detectado en otros estudios (Carrera-Sabaté *et al.*, 2014), es necesariamente diferente.

4. OBSERVACIONES FINALES

Las GPC son la primera propuesta de autoaprendizaje e-learning de la pronunciación del catalán con una integración de la prosodia. Pretende ser, además, una propuesta útil para el aprendizaje de la pronunciación del catalán que incluye la cultura catalana como entramado que sostiene la lengua.

Uno de los grandes retos las GPC es transmitir la complejidad que subyace en las propuestas de mejora de la pronunciación del catalán como segunda lengua sin una excesiva simplificación de los contenidos, cosa que distanciaría a los usuarios de la realidad.

Finalmente, el hecho de que el portal tenga acceso abierto y pueda consultarse des de diferentes interfaces le otorga una versatilidad que lo acerca a la sociedad. Además, la organización de los materiales por niveles de dificultad con la integración en la cultura catalana permite que las GPC puedan utilizarse tanto por alumnos como por profesores de catalán, por profesionales de la logopedia, la corrección o los *mass media*.

5. REFERENCIAS

Álvarez, V. (2010). El valor didáctico de la literatura folclórica para enseñar inglés. *Pedagogía Magna, 5*, 243-250.

Bau, M.; Pujol, M.; Rius, A. (2007). *Curs de pronunciació*. Barcelona: Publicacions de l'Abadia de Montserrat.

Bueno, D. (Ed.) (2015) Llengua, societat i comunicació. *Cervell i llenguatge, 13*.

Campillos, L. (2010). Tecnologías del habla y análisis de la voz. Aplicaciones en la enseñanza de la lengua. *Diálogo de la Lengua, 2*, 1–41.

Carrera-Sabaté, J. (2012). Els sons del català. Una propuesta para la docencia de la fonética catalana. M. Cruz y M. Trenchs (Eds.), *Experiencias de innovación docente en la enseñanza universitaria de las Humanidades* (pp. 73–89). Barcelona: Octaedro.

Carrera-Sabaté, J. (en prensa). *Eines de fonètica catalana*.

Carrera-Sabaté, J., Bach, J., y Mir, M. (2016). La pronúncia del català en una experiència d'aprenentatge i servei i de tutories entre iguals. *Congrés FIET*, Girona.

Carrera-Sabaté, J., Creus, I., Fernández, A. M., Blanco, E., y Pons, C. (2014). Guies de pronunciació del català: tools and reflections to modify articulatory habits at University. *6th International Conference on Education and New Learning Technologies (EDULEARN 14)*.

Carrera-Sabaté, J., Valls, E., Royo, L., y Lloret, M. R. (2016). Millorar la pronúncia del català en graus de Mestre i de Comunicació: una acció d'innovació docent mitjançant el web Guies de pronunciació del català. *IX Congrés Internacional de Docència Universitària i Innovació*, Bellaterra, Universitat Autònoma de Barcelona.

Cavus, N. (2016). Development of an intelligent mobile application for teaching English pronunciation. *Procedia Computer Science*, *102*, 365–369.

Clark, R. C. y Mayer, R. E. (2016). *E-learning and the science of instruction* (4.ª ed.). New Jersey: Wiley.

Creus, I. y Julià-Muné, J. (2010). Multilingüisme i fonètica contrastiva: els «eurosons» i la pronúncia del català. *Actes del Quinzè Col·loqui Internacional de Llengua i Literatura Catalanes* (Universitat de Lleida, 7–11 de setembre de 2009) (pp. 447–460). Barcelona: PAM.

Dalmau, M., Miró, M., y Molina, D.-À. (1985). *Correcció fonètica. Mètode verbo-tonal.* Vic: Eumo.

Damasio, A. (2005). *En busca de Spinoza. Neurobiología de la emoción y los sentimientos*. Barcelona: Crítica.

Darder, P. (2013) Emocions i educació, una integració necessària. En P. Darder *et al.* (Eds.) *Aprendre i ensenyar amb benestar i empatia* (pp. 11–22). Barcelona: Octaedro.

Delmonte, R. (1988). Analisi Automatica delle Strutture Prosodiche. R. Delmonte, G. Ferrari y I. Prodanoff (Eds.), *Studi di lingüística computazionale* (pp. 109–162). Padova: Unipress.

Gil, J. (2007). *Fonética para profesores de español: de la teoría a la práctica*. Madrid: Arco/Libros.

Hays, R. T. (2005). *The effectiveness of instructional games: A literature review and discussion*. Patuxent River, MD: Naval Air Warfare Center Training Systems Division.

Huizinga, J. (1987). *Homo Ludens.* Madrid: Alianza.

Hurley, D. S. (1992). Issues in Teaching Pragmatics, Prosody, and Non-Verbal Communication. *Applied Linguistics*, *13*, 259–280.

Kapp, K. (2012). *The gamification of learning and instruction. Game-based methods and strategies for training and education.* San Francisco: Pfeiffer-Wiley.

Lee, J. J. y Hammer, J. (2011). Gamification in Education: What, How, Why Bother? *Academic Exchange Quarterly*, *15*(2).

Llisterri, J. (2006). La enseñanza de la pronunciación asistida por ordenador. *Actas del XXIV Congreso Internacional de AESLA. Aprendizaje de lenguas, uso del lenguaje y modelación cognitiva: perspectivas aplicadas entre disciplinas* (pp. 91–120).

Murphy, L. (2015). Online language teaching: the learner's perspective. En R. Hampel y U. Stickler (Eds.), *Developing online language teaching* (pp. 45–62). New York: Palgrave Macmillan.

Piaget, J. (1986). *La formación del símbolo en el niño*. México: Fondo de Cultura Económica.

Randel, J. M., Morris, B. A., Wetzel, C. D. y Whitehill, B. V. (1992). The effectiveness of games for educational purposes: A review of recent research. *Simulation and Gaming*, *23*(3), 261–276.

Sitzmann, T. (2011). A meta-analytic examination of the instructional effectiveness of computer based simulation games. *Personnel Psychology*, *64*(2), 489–528.

Witt, S. M. (2012). Automatic Error Detection in Pronunciation Training: Where we are and where we need to go. En O. Engwall (Ed.), *IS ADEPT: International Symposium on Automatic Detection of Errors in Pronunciation Training* (pp. 1–8). Stockholm: KTH.

Webgrafía

British Council, Stories & Poems. http://learnenglish.britishcouncil.org/en/stories-poems?utm_source=facebook&utm_medium=social&utm_campaign=bc-learnenglish

Gabriel Biblioni. biblioni.cat/correcciofonetica

Proyecto GALÍ. http://clic.xtec.cat/gali

Els sons del català. http://www.ub.edu/sonscatala/ca/

Guies de pronunciació del català. http://www.ub.edu/guiesdepronunciacio/

Guies de pronunciació del català. Itinerario español peninsular-catalán. http://www.guiesdepronunciacio.cat/catala-espanyol

Guies de pronunciació del català. Abans de començar. http://www.guiesdepronunciacio.cat/abans-de-començar

Guies de pronunciació del català. Entonació. http://www.guiesdepronunciacio.cat/entonacio

Carrera-Sabaté, J., Bach Marqués, J. y Mir Campillo, M. (2019). *Dumloquor hora fugit:* aprendizaje autónomo y autorregulado de la pronunciación del catalán a través de las *Guies de pronunciació del català*. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 65–73). Málaga: Universidad de Málaga.

Subsidia: Tools and Resources for Speech Sciences

12

Explicit and implicit training methods for the learning of stress contrasts in Spanish

Sandra Schwab[1,2] and Volker Dellwo[1]

[1] Universität Zürich
[2] Université de Genève
e-mail: Sandra.Schwab@unige.ch, Volker.Dellwo@uzh.ch

Citation / Cómo citar esta publicación: Schwab, S. & Dellwo, V. (2019). Explicit and implicit training methods for the learning of stress contrasts in Spanish. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 75–80). Málaga: Universidad de Málaga.

ABSTRACT: In the present paper, we propose two computer-assisted training methods—one explicit and another implicit—for the learning of stress contrasts in Spanish L2. The explicit training approximates the learning environment of a L2 phonetics classroom, in which the learners receive metalinguistic explanations about the Spanish accentual system and perform various exercises commonly used in corrective phonetics. In the implicit training, the learners do not receive any metalinguistic explanations and always perform the same task, namely a shape / word matching task. We evaluated the efficiency of both training methods with French and German learners. The results showed that both training methods allowed French and German learners to significantly improve their perception of Spanish lexical stress. Moreover, the differences we observed between the progressions with explicit and implicit training methods were subtle, which suggests that both immersion settings and metalinguistic explanations can be beneficial in a L2 phonetics classroom.

Keywords: lexical stress perception; training; explicitness of the instructions.

RESUMEN: En este artículo, proponemos dos métodos de entrenamiento asistido por ordenador —uno explícito y otro implícito— para el aprendizaje de los contrastes de acento léxico en español como L2. El entrenamiento explícito se aproxima a las circunstancias de aprendizaje de un aula de fonética de L2, en la que los aprendices reciben explicaciones metalingüísticas sobre el sistema acentual español y realizan varios ejercicios habitualmente utilizados en la fonética correctiva. En el entrenamiento implícito, los estudiantes no reciben ningún tipo de explicación metalingüística y siempre realizan la misma tarea, que consiste en unir palabras con figuras. Hemos evaluado la eficiencia de ambos métodos de entrenamiento con estudiantes hablantes de francés y de alemán. Los resultados muestran que ambos métodos permitieron a todos los aprendices mejorar significativamente su percepción del acento léxico en español. Las diferencias observadas entre la progresión con el método explícito y con el implícito son sutiles, lo que sugiere que tanto los contextos de inmersión como las explicaciones metalingüísticas pueden ser beneficiosos en una clase de fonética de L2.

Palabras clave: percepción del acento léxico; entrenamiento; explicitud de las instrucciones.

1. INTRODUCTION

In this paper, we propose two computer-assisted training methods—one explicit and another implicit—for the learning of accentual contrasts in L2-Spanish. According to Norris & Ortega (2000), explicit instructions show a certain advantage over implicit instructions in the learning of morphological, syntactic, and pragmatic aspects. However, such an advantage is not clearly found in pronunciation aspects (e.g., Kissling, 2013; Saito, 2013). Moreover, researchers have rarely focused their attention on the phonetic training of suprasegmental features (Thomson & Derwing, 2015), giving most of the time the priority to segmental features. Furthermore, the few studies that have indeed dealt with suprasegmental elements have mainly examined tone acquisition (e.g., Wang, Spence, Jongman & Sereno, 1999). To the best of our knowledge, only a couple of papers have investigated the efficiency of a training in the acquisition of lexical stress. On the one hand, they have shown that the perception of lexical stress improved after training (Carpenter, 2015; Schwab & Llisterri, 2011). On the other hand, they have indicated that an explicit training method improved comprehensibility in L2. In other words, explicit explanations about lexical stress had an

impact on the production skills in L2 (Gordon, Darcy & Ewert, 2013). Similar studies need to be conducted to address the perception of lexical stress. In the present paper, we fill this gap by proposing two training methods (explicit and implicit) and by testing their efficiency with French and German learners.

2. SETTINGS

Two computer-assisted trainings methods that focus on the perception of lexical stress in Spanish are presented in the next sections. The first one, explicit, approximates the situation of an L2 phonetics classroom, while the second one, implicit, resembles a situation of immersion. Each training consists of eight 30-minute sessions, for a total of approximately 4 hours.

The two training methods are dedicated (for the moment) to French- or German-speaking learners of Spanish (i.e., the instructions and/or explanations are given in French or German). They are however easily portable to other languages.

Both training methods are developed as Praat plugins (Boersma & Weenink, 2016). All the responses given by the learners are saved in a txt file, to allow the teacher/experimenter to examine, for example, the learners' progression along the training.

3. EXPLICIT TRAINING

As mentioned earlier, the explicit training, although computer-assisted, approximates the learning environment of a L2 phonetics classroom. The learners receive explicit metalinguistic explanations about the Spanish accentual system and perform various exercises commonly used in corrective phonetics.

3.1. Part I

In the first part of the explicit training, learners watch a video (approximately 9 minutes) in Spanish, with either French or German subtitles. In the video, the Spanish accentual system is described with examples of the different stress patterns (i.e., proparoxytone, paroxytone and oxytone). The French and the German versions of the explanations can be found on youtube under the following links:
https://www.youtube.com/watch?v=4rCND8IOaWo (French subtitles)
https://www.youtube.com/watch?v=V5t2ItZEr64 (German subtitles).

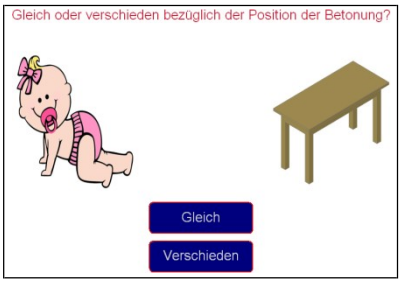
3.2. Part II

The second part of the explicit training is composed of eight sessions. Each session is divided into five exercises (described in the next sections) and lasts approximately 30 minutes.

The entire training includes 96 Spanish words, each associated with a drawing to make the training more natural. Among these words, 24 are proparoxytone (12 trisyllabic and 12 quadrisyllabic), 36 are paroxytone (12 bisyllabic, 12 trisyllabic and 12 quadrisyllabic) and 36 are oxytone (12 bisyllabic, 12 trisyllabic and 12 quadrisyllabic).

Figure 1: Explicit training. Localization of a given stress pattern.

Figure 2: Explicit training. Discrimination of stress patterns.



The difficulty of the tasks increases towards the end of the training as follows: the words are produced by two different speakers, the number of words presented in the trials or/and the number of syllables in the words increases, and the words are presented in a declarative or interrogative carrier sentence.

3.2.1. Exercise 1: localization of a given stress pattern

In the first exercise (composed of 30 questions), the learners have to localize a given accentual pattern. They hear words (2, 3 or 4 words depending on the session), each one associated with a drawing, and have to answer to the following question: "Which word has stress on the X syllable". For example, as shown in Figure 1, the learners hear the words "zapatillas"[1] (eng. sneakers) and "melocotón" (eng. peach) and have to indicate which word has stress on the penultimate syllable (i.e., "zweitletzen silbe" in German). The learners indicate their response by clicking on the drawing corresponding to the word with the given stress pattern. They receive a feedback about their response (here: correct response = zapatillas) and have the possibility to hear the words again.

3.2.2. Exercise 2: discrimination of stress patterns

In the second exercise (composed of 30 questions), the learners hear words (2, 3 or 4 words depending on the session), associated with drawings, and have to indicate whether the stress position of the words is the same or different. For example (see Figure 2), the learners hear the words "bebé" (eng. baby) and "mesa" (eng. table) and indicate whether the two words present the same or a different stress pattern. After giving their response, the

[1] The stressed syllable is underlined.

Schwab, S. & Dellwo, V. (2019). Explicit and implicit training methods for the learning of stress contrasts in Spanish. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 75–80). Málaga: Universidad de Málaga.

Figure 3: Explicit training. Repetition.

Figure 4: Explicit training. Identification of the correct stress pattern.

learners receive feedback (here: correct response = different) and have the possibility to listen to the words again.

3.2.3. Exercise 3: repetition

In the third exercise (composed of 15 questions), the learners are asked to repeat words (in isolation or in a carrier sentence, depending on the session). Despite the fact that the learners' productions are not recorded, it is important to include in the training a production task, since this task is frequently used in a phonetics classroom. As shown in Figure 3, the learners click on the drawing to hear the word they have to repeat (for example, "café"). The stress pattern is provided (the red dot represents the stressed syllable and the blue one(s) the unstressed syllable(s)) in case the learners do not perceive the stressed syllable correctly.

3.2.4. Exercise 4: identification of the correct stress pattern

In the fourth exercise (composed of 15 questions), the learners have to identify the correct stress pattern of a given word. They see the drawing of a word that they learned in the preceding exercises or sessions, and different pronunciations of the word are presented. The pronunciations differ in the stress pattern and the learners are asked to indicate the correct pronunciation. Figure 4 gives the example of the word "viento" (eng. wind), where, for example, the first pronunciation has stress on the penultimate syllable and the second on the final syllable. The learners indicate their response by clicking on the pronunciation with the correct stress pattern. They receive a feedback about their response (here: correct response = first pronunciation) and have the possibility to hear the pronunciations again.

Figure 5: Explicit training. Test (discrimination).

3.2.5. Exercise 5: test (discrimination)

The fifth and last exercise (composed of 10 questions) consists in a short test in which the learners do not receive any feedback. This exercise is identical in the eight sessions and serves to evaluate the learners' progression along the training. The learners perform a discrimination task (as in exercise 2) in which they hear 4 words in the interrogative carrier sentence "Has dicho X?" (eng. Did you say X?), the word X being associated with a drawing. They are asked to indicate whether the words present the same or a different stress pattern. For example, (see Figure 5), the learners hear "Has dicho corazón?" (eng. Did you say heart?), "Has dicho reloj?" (eng. Did you say alarm clock?), "Has dicho televisor?" (eng. Did you say tv?) "Has dicho champú?" (eng. Did you say shampoo?) and have to indicate whether the stress pattern of the last word is the same or different in the four sentences (here: correct response = same).

3.3. Summary

At the end of the explicit training, the learners have learned 96 Spanish words that were produced by two speakers, in isolation and in declarative and interrogative carrier sentences.

4. IMPLICIT TRAINING

In the computer-assisted implicit training, the learners do not receive metalinguistic explanations about the Spanish accentual system and no mention to stress patterns is made in the instructions. Like the explicit training, the implicit training is divided into eight 30-minute sessions (total = 4 hours). However, contrary to the explicit training, the learners always perform the same task, namely a shape / word matching task. Each session is composed of 6 identical parts with feedback, except the last one that serves to evaluate the progression along the training.

4.1. Shape / word matching task

The learners hear a word and 4 shapes appear on the screen (see Figure 6). They have to click on the shape they think corresponds to the word they hear. After giving their response, they receive feedback: they hear the word again and only the correct shape stays on the screen. The feedback enables the learners to learn the correspondence between the words and the shapes.

Figure 6: Display of the implicit training.

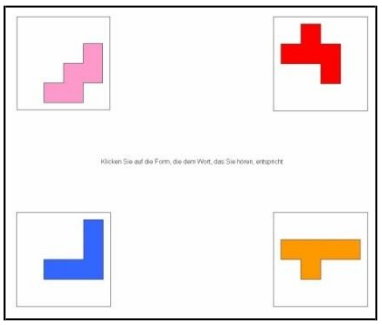


Figure 7: Shapes and words used in the implicit training.

Shape	Word
	cáscara
	cascara
	cascará
	género
	genero
	generó

4.2. Words and shapes

Two triplets of trisyllabic Spanish words are used in the implicit training: cáscara (eng. shell), cascara (that he cracked), cascará (eng. he will crack) and género (eng. gender), genero (eng. I generate), generó (eng. he generated) associated to 6 the shapes as shown in Figure 7.

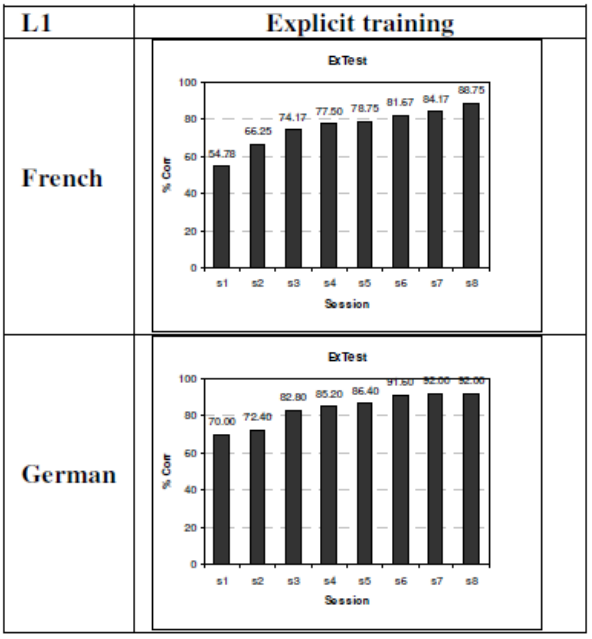
As can be observed, each triplet is composed of a proparoxytone (i.e., cáscara and género), a paroxytone (i.e., cascara and genero) and an oxytone word (i.e., cascará and generó). The six words are produced by two speakers with a falling and a rising intonation. In total, 24 auditory stimuli (6 words x 2 speakers x 2 intonations) are used in the implicit training.

4.3. Design

As mentioned earlier, each of the eight session is composed of 6 identical parts. Each word (e.g., cáscara) is presented six times per part (6 words x 6 times = 36 times per part). Depending on the session, each word is spoken by one of the two speakers, with a falling or a rising intonation.

Among the four shapes that appear on the screen, only one corresponds to the word and the three others are distractors. Among the three distractors, one corresponds to a word with a different stress pattern (e.g., cascara) and the two other shapes correspond to two words from the other triplet (e.g., género and genero). Each shape appears the same number of times in each part, as well as in the four positions on the screen.

Figure 8: Percent correct as a function of the training sessions of the explicit training for the French (top) and German learners (bottom).



4.4. Summary

At the end of the implicit training, the learners have learned six Spanish words that were produced by two speakers with a falling and a rising intonation.

5. EVALUATION OF THE TRAINING METHODS

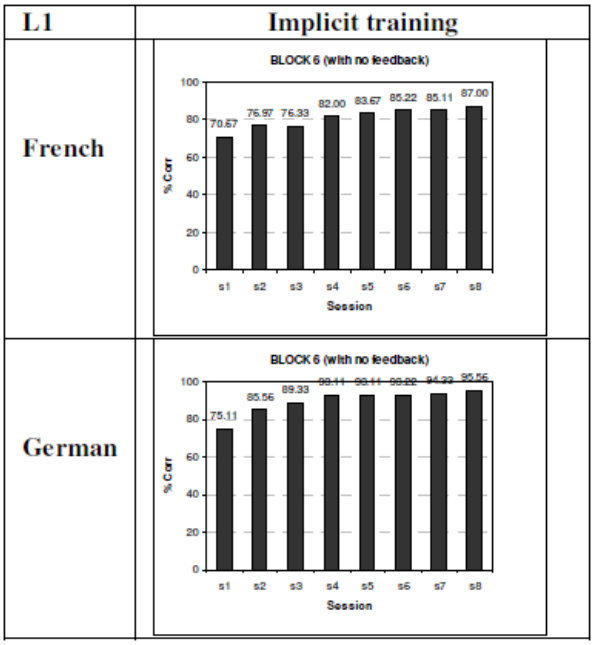
We report in this section the results of an evaluation we carried out to test the effect of the training methods. For this, we ran an experiment composed of 3 phases: (1) Pre-test; (2) Training; (3) Post-test (identical to the pre-test). In the pre- and post-tests, the participants performed a stress identification task. They heard words produced by two speakers with a falling or a rising intonation and had to indicate the stressed syllable. Two groups of learners took part in the experiment. The first group was composed of 49 Swiss French listeners with no knowledge of Spanish or other Romance language (from the University of Neuchâtel). The second group was composed of 50 Swiss German listeners with no knowledge of Spanish or other Romance language, except French (from the University of Zurich). Half of the learners received the explicit training and the other half the implicit one.

In the next sections, we report the results about the progression along both trainings (section 3.1) as well as the results related to the efficiency of both trainings (section 3.2).

5.1. Progression along the training

Figure 8 presents the progression along the explicit training for the French (top) and the German learners (bottom). The data corresponds to the learners' performance in the fifth exercise of each of the eight sessions. As mentioned in section 3.2.5, the learners performed in this exercise a discrimination task with no

Figure 9: Percent correct as a function of the training sessions of the implicit training for the French (top) and German learners (bottom).



feedback. We observe that, although the French learners present a lower performance than the German learners (at least at the beginning of the training), they show a strong progression along the training (from 55% to 89%). The German learners also present a progression along the training (from 70% to 92%).

Figure 9 presents the progression along the implicit training for the French (top) and the German learners (bottom). The data corresponds to the learners' performance in the last part (i.e., Block 6 with no feedback) of each of the eight sessions. Both groups of learners present an improvement along the training. The French performance increases from 71% to 87%, whereas the German one goes from 75% to 96%.

Taken together these findings indicate that both training methods lead to an improvement in the tasks relative to the perception of lexical stress (i.e., discrimination and shape / word matching task).

5.2. Efficiency of the training

We present here the results related to the learners' ability to generalize the knowledge they have learned during the training to another task and to other words. We examined, by means of mixed-effects logistic regression models (separate models for the French and the German learners), the effect of test (i.e., difference between pre- and post-tests) and the effect of the training method (i.e., explicit vs implicit) on the stress identification performance. We also investigated whether the progression from pre- to post-test was different with both training methods (i.e., interaction between test and training method).

As far as the French learners are concerned, results (see Table 1) show a main effect of test ($\chi^2(1) = 237.51$, $p = .000$), but no main effect of training method ($\chi^2(1) = 0.1$, $p = .75$). Interestingly, we note a marginal interaction between test and training method: the progression is marginally better for the explicit training than for the implicit one (14% and 11%, respectively; $\chi^2(1) = 3.6$, $p = .06$).

Table 1: Percent correct for the identification task in the pre- and post-test for the explicit and implicit training and for French and German learners, as well as for Spanish listeners.

L1	Training	Pre-test	Post-test	an
French	Explicit	56	70	63
	Implicit	56	67	62
	Mean	*56*	*68*	*62*
German	Explicit	72	81	77
	Implicit	74	86	80
	Mean	*73*	*84*	*78*

As for the German listeners, we also observe a main effect of test ($\chi^2(1) = 258.03$, $p = .000$) and no effect of training method ($\chi^2(1) = 1.1$, $p = .29$). Moreover, we observe an interaction between test and training method. However, contrary to French listeners, the progression is better for the implicit training than for the explicit one (12% and 9%, respectively; $\chi^2(1) = 9.29$, $p = .002$).

5.3. Discussion

Results revealed that a 4-hour training improved the perception of Spanish lexical stress contrasts by French and German listeners with no knowledge of Spanish. Although both types of training (explicit versus implicit) led to a significant improvement, their effect was not similar in both groups of listeners.

French listeners showed a marginally larger progression with the explicit than with the implicit training, which suggests that receiving metalinguistic explanations constitutes a (small) advantage for the perception of lexical stress in a second language, especially for learners who lack the concept of lexical stress in their native language (i.e., French). On the contrary, German listeners, who are used to process stress contrasts in their native language, did not show any preference for the explicit training. Indeed, they showed a larger progression with implicit than with explicit training method, which is difficult to explain.

To ensure that the progression we observed for both training methods was not due to the fact that the listeners performed the same task twice (i.e., before and after training), 23 French-speaking listeners, who did not receive any training, performed twice the identification task at a 2-week interval. Results showed no progression (52.43% vs 52.46%), which confirms that the improvement we observed with explicit and implicit training methods was caused by the training phase and not by the repetition of the task.

6. CONCLUSION

This research shows that a 4-hour training on Spanish accentual contrasts allows French- and German-speaking learners with no knowledge of Spanish to significantly improve their perception of Spanish lexical stress contrasts. Moreover, the differences we observe between the progressions with explicit and implicit

Schwab, S. & Dellwo, V. (2019). Explicit and implicit training methods for the learning of stress contrasts in Spanish. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 75–80). Málaga: Universidad de Málaga.

training methods are subtle, which suggests that both immersion settings and metalinguistic explanations can be beneficial in a L2 phonetics classroom, especially if the learners' native language lack the L2 features under study (i.e., lexical stress contrasts).

7. ACKNOWLEDGMENTS

We would like to thank L. Baqué, M.ª A. Barquero, M. Carranza, J. M. Lahoz, J. Llisterri, M. Machuca, and A. Ríos for their helpful advice during the elaboration of this experiment. S. Schwab's work was supported by the Swiss National Science Foundation (grant Ambizione PZ00P1_148036/1).

8. REFERENCES

Boersma, P., & Weenink, D. (2016). Praat: doing phonetics by computer [computer program]. Retrieved from http://www.praat.org/

Carpenter, A. (2015). Phonetic training significantly mitigates the stress 'deafness' of French speakers. *International Journal of Linguistics, 7*.

Gordon, J., Darcy, I., & Ewert, D. (2013). Pronunciation teaching and learning: Effects of explicit phoneticin struction in the L2 classroom. In J. Levis & K. LeVelle (Eds.), *Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference* (pp. 194–206).

Kissling, E. M. (2013). Teaching pronunciation: Is explicit phonetics instruction beneficial for FL learners? *Latin American, Latino and Iberian Studies Faculty Publications. Paper 8*. Available at http://scholarship.richmond.edu/lalis-faculty-publications/8.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning, 50*, 417–528.

Saito, K. (2013). Reexamining effects of form-focused instruction on L2 pronunciation development: The role of explicit phonetic information. *Studies in Second Language Acquisition, 35*, 1–29.

Schwab, S., & Llisterri, J. (2011). Are French speakers able to learn to perceive lexical stress contrasts? In W.-S. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 1774–1777).

Thomson, R. I., & Derwing, T. M. (2015). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics, 36*, 326–344.

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *Journal of the Acoustical Society of America, 106*, 3649– 3658.

Subsidia: Tools and Resources for Speech Sciences

Els sons del català, una herramienta digital para aprender fonética y fonología catalanas en la red

Clàudia Pons-Moll[1], Josefina Carrera-Sabaté[1]

[1] Universitat de Barcelona
e-mail: jcarrera@ub.edu

Citation / Cómo citar esta publicación: Pons-Moll, C., Carrera-Sabaté, J. (2019). Els sons del català, una herramienta digital para aprender fonética y fonología catalanas en la red. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 81–87). Málaga: Universidad de Málaga.

RESUMEN: En este artículo presentamos y describimos el sitio web Els sons del català, una interfaz digital para aprender fonética y fonología catalanas, de acceso libre en la red (http://www.ub.edu/sonscatala/), que se publicó por primera vez en 2009 y que se ha ido renovando hasta la fecha. Los objetivos generales de la interfaz, que nació como iniciativa de innovación docente, son a) facilitar la enseñanza de la fonética y de la fonología catalanas en todas aquellas asignaturas de la enseñanza superior, de la enseñanza secundaria y de otras enseñanzas que tienen esta disciplina por objeto de estudio o medio de trabajo; b) favorecer que el alumnado pueda aprender esta disciplina de manera autónoma, de acuerdo con las directrices del Espacio Europeo de Educación Superior; c) mejorar la enseñanza del catalán como lengua extranjera; d) proporcionar recursos y materiales de apoyo a los profesionales del asesoramiento lingüístico, del periodismo y de la logopedia, entre otros.

Palabras clave: fonética; fonología; catalán; aprendizaje y enseñanza; Internet.

ABSTRACT: In this article we present and describe the website Els sons del català, a digital interface for learning Catalan phonetics and phonology, of free access in the net (<http://www.ub.edu/sonscatala/>), which was first published in 2009 and has been updated to date. The general goals of the interface, which was originally conceived as a teaching innovation initiative, are a) to make the teaching of Catalan phonetics easier within the subjects of Higher Education, Secondary Education and other educational programs that include the study of Phonetics; b) to encourage students to learn this discipline in an autonomous way, in accordance with the European Higher Education Area guidelines; c) to improve the teaching of Catalan as a foreign language (by furnishing the learner of Catalan as L2 with the basic devices to learn the main features concerning the production and perception of Catalan sounds); d) to provide Internet support materials and resources to those professionals working in fields such as language advising, journalism and speech therapy, among others.

Keywords: phonetics; phonology; Catalan; learning and teaching; Internet.

1. INTRODUCCIÓN

En este artículo presentamos y caracterizamos el sitio web Els sons del català, una interfaz digital para aprender fonética y fonología catalanas, de acceso libre en la red (http://www.ub.edu/sonscatala/), que se publicó por primera vez en 2009 y que se ha ido renovando hasta la fecha. Los objetivos generales de la interfaz, que nació como iniciativa de innovación docente, son (1) facilitar la enseñanza de la fonética y de la fonología catalanas en todas aquellas asignaturas de la enseñanza superior, de la enseñanza secundaria y de otras enseñanzas que tienen esta disciplina por objeto de estudio o medio de trabajo; (2) favorecer que el alumnado pueda aprender esta disciplina de manera autónoma, de acuerdo con las directrices del Espacio Europeo de Educación Superior; (3) mejorar la enseñanza del catalán como lengua extranjera; (4) proporcionar recursos y materiales de apoyo a los profesionales del asesoramiento lingüístico, del periodismo y de la logopedia, entre otros. Actualmente, el espacio incluye los sonidos del catalán central, noroccidental, del valenciano (central) y del balear (mallorquín), los cuales constituyen una fase de un proyecto más amplio que recogerá información sobre otras variedades dialectales del catalán, como el alguerés o el septentrional.

El artículo se organiza de la forma siguiente. En el § 2 se introduce y contextualiza el sitio web. En el § 3 se describen los aspectos más generales de la web y se

expone el sistema de navegación que requiere para el usuario. En el § 4 se detallan las características más específicas del sitio, así como las convenciones gráficas utilizadas. En el § 5 se destacan las propiedades de la página desde el punto de vista educativo. En el § 6 se presenta de forma breve la difusión y el impacto que ha tenido la herramienta desde que se publicó en Internet por primera vez, y en el § 7 se hace referencia a la tipología potencial y real de usuarios de la página. El § 8, finalmente, se dedica a las conclusiones.

2. ELS SONS DEL CATALÀ

El espacio web Els sons del català se concibió originalmente con dos finalidades. Por un lado, para que los estudiantes de fonética y fonología catalanas pudieran encontrar recursos básicos y suficientes para comprender tanto el proceso de producción de los sonidos del catalán como el contraste articulatorio entre sonidos. Para alcanzar este primer objetivo, se buscó realismo, claridad y transparencia en las designaciones y las descripciones de los sonidos y de los símbolos fonéticos que se utilizan para representarlos (véase la Figura 1) y también en las imágenes y representaciones del tracto vocálico (véase la Figura 2). Por otro lado, también se concibió con la finalidad de que los usuarios que quisieran profundizar en la materia, pudieran encontrar información más detallada sobre la producción y transmisión de los sonidos del catalán. Por ello, la interfaz contiene, también, espectrogramas y oscilogramas (véase la Figura 3), palatogramas (véase la Figura 4), e imágenes y vídeos de los articuladores y de los movimientos articulatorios, capturados a partir de resonancia magnética dinámica (véase la Figura 5).

3. DESCRIPCIÓN GENERAL DE LA WEB

En este apartado se describen los aspectos más generales y relevantes de la web y se expone el sistema de navegación que requiere para el usuario. La web, que dispone de dos niveles de navegación básicos, se organiza de la siguiente manera.

3.1. Primer nivel de navegación

Por medio de la pestaña “Tablas de sonidos” situada en el menú principal y desplegable en función de las diferentes variedades dialectales (balear mallorquín, central, noroccidental y valenciano central), se accede a las tablas de sonidos correspondientes. En esta sección, se pueden encontrar cuatro tablas animadas que contienen los símbolos fonéticos correspondientes a los sonidos consonánticos y vocálicos de cada uno de los dialectos, y a los diacríticos y los suprasegmentales (véase la Figura 6). La tabla de consonantes está dividida, como es usual, en función del modo de articulación, del punto de articulación y de la sonoridad del sonido consonántico. Los sonidos sordos aparecen a la izquierda y los sonoros, a la derecha. Cuando se sitúa el ratón encima de cada modo y punto de articulación, se despliega una ventanilla en la que se indica qué órganos intervienen en la producción del sonido, cómo lo hacen (modo de articulación; véase la Figura 7) y en

Figura 1: Tabla de símbolos consonánticos con el ejemplo ilustrativo de la descripción de un sonido, y explicación del origen del símbolo que lo representa.

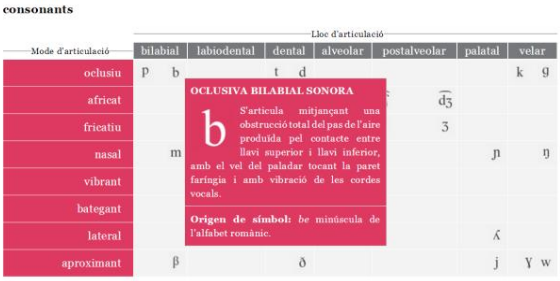


Figura 2: Diagrama articulatorio animado de la consonante fricativa alveolar sonora del catalán.

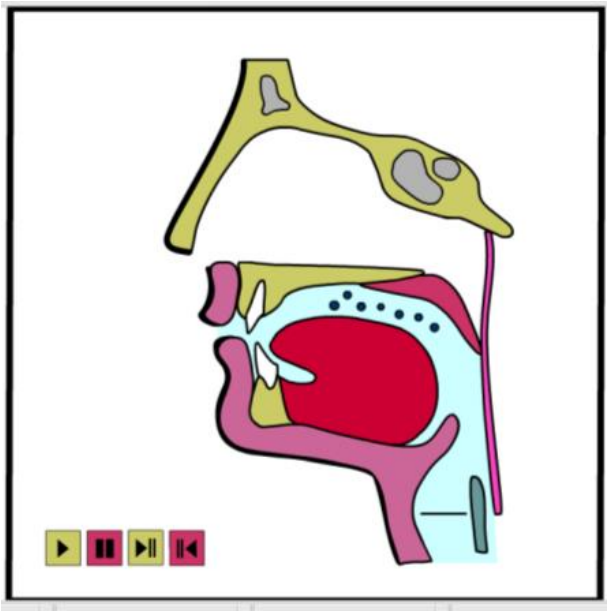


Figura 3: Representación oscilográfica y espectrográfica de la palabra pujol, con la fricativa prepalatal sonora delimitada por dos líneas rojas.

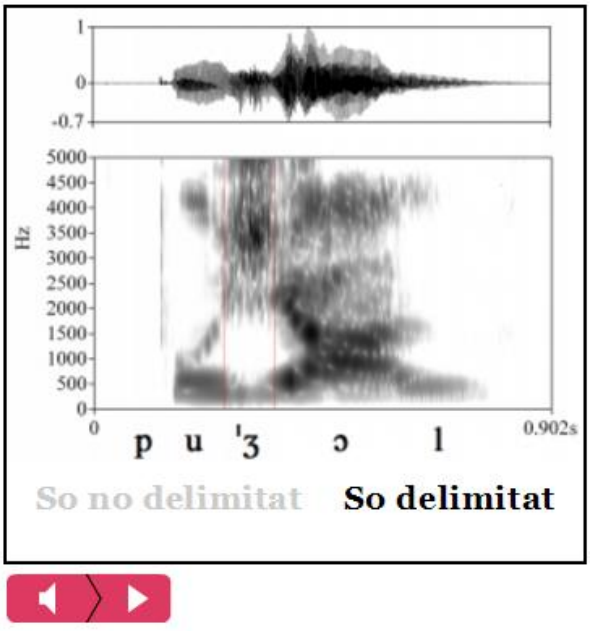


Figura 4: Palatograma correspondiente a una secuencia de la articulación de la consonante fricativa prepalatal.

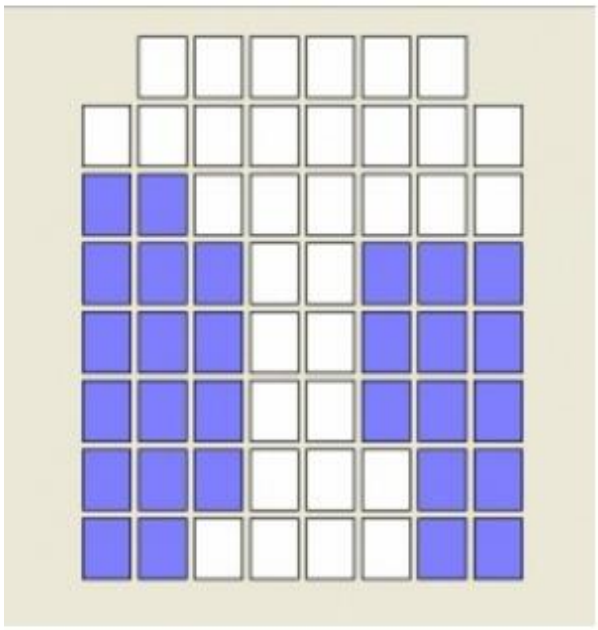


Figura 5: Vídeo obtenido con resonancia magnética dinámica de la articulación de la consonante fricativa prepalatal sonora.

qué punto del tracto vocal lo hacen (punto de articulación; véase la Figura 8). La tabla de vocales está dividida, como es tradicional, en función del grado de elevación lingual y del grado de avance lingual. Las ventanillas desplegables, en este caso, aportan información sobre los grados de elevación y de avance posibles.

Los símbolos contenidos en las tablas de diacríticos y suprasegmentales también van acompañados de ventanillas que contienen información sobre el uso y la posición de estos elementos con respecto al sonido al que afectan.

En todas las tablas, cuando se desplaza el ratón encima de un símbolo, se despliega una ventanilla con una descripción de las características articulatorias del sonido correspondiente, así como una breve referencia al origen y las características del símbolo fonético que lo representa.

3.2. Segundo nivel de navegación

Cuando se clica en cada uno de los símbolos fonéticos

Figura 6: Tabla de consonantes, vocales, diacríticos y suprasegmentales.

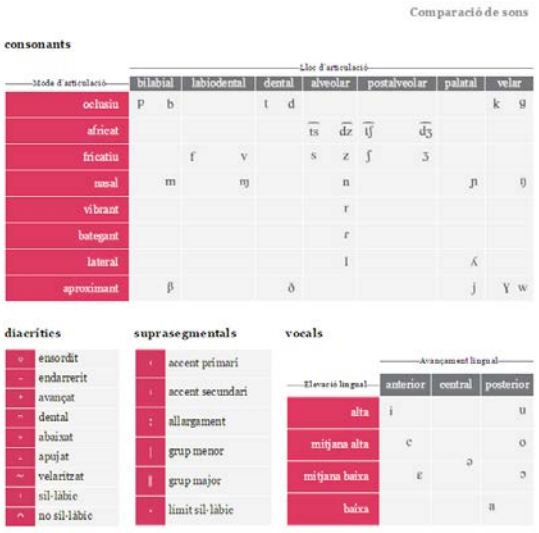


Figura 7: Ejemplo de descripción del modo de articulación.

Figura 8: Ejemplo de descripción del punto de articulación

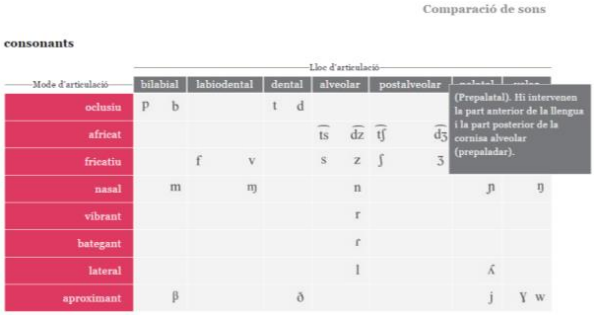
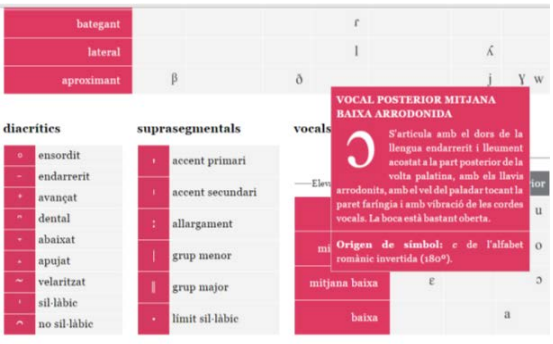
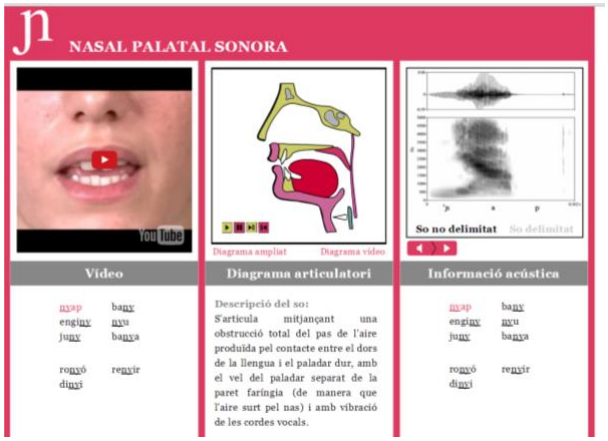


Figura 9: Ejemplo de descripción de la vocal posterior mediana baja y del símbolo que lo representa.



de las tablas de consonantes y vocales, se accede a una ficha con información sobre el sonido en cuestión. Esta ficha contiene la siguiente información.

Figura 10: Información que aparece al clicar en cada sonido.



En la franja de la izquierda, aparece un vídeo en el que dos hablantes nativos de la variedad dialectal articulan el sonido seleccionado en diferentes posiciones dentro de la palabra (inicial, medial y final), siempre que ello es posible (véase la Figura 10).

En la franja central, aparece un diagrama articulatorio animado que representa la trayectoria del aire y de los movimientos de los órganos articulatorios que intervienen en la producción del sonido en cuestión. Este diagrama articulatorio tiene dos formatos: vídeo y Flash (véase la Figura 10). Cuanto más oscuras son las bolas que representan el flujo del aire, más obstruyente es el sonido. Inversamente, cuanto más claras son, más sonorante es. Para más detalle sobre las convenciones gráficas usadas en los diagramas articulatorios, véase el § 4.1.

En la franja de la derecha, aparece un espectrograma y un oscilograma acompañados de un archivo de sonido que se corresponden con cada una de las palabras pronunciadas por los hablantes nativos. Existe la posibilidad de visualizar el sonido seleccionado delimitado por líneas rojas para distinguirlo del resto de sonidos de la palabra (véase la Figura 10).

En la zona central, debajo el diagrama articulatorio, aparece el enlace "Palatogramas e IRMD", que contiene, para la mayoría de sonidos, información sobre el grado de contacto entre la lengua y el paladar, e imágenes y vídeos (correspondientes a la articulación de cada uno de los sonidos) obtenidos por medio de la resonancia magnética dinámica (véanse la Figura 4 y la Figura 5, ya presentadas).

En la parte superior derecha, se encuentra el enlace "Partes del tracto vocal", que contiene un diagrama articulatorio interactivo con las designaciones que reciben las diferentes partes del tracto vocal, incluidas las cavidades y los articuladores (para más detalle sobre esta aplicación en concreto, véase el § 4.2).

4. DESCRIPCIÓN ESPECÍFICA DE LA WEB

4.1. Los diagramas articulatorios

Para poder plasmar gráficamente las características de los diferentes sonidos, se establecieron un conjunto de convenciones gráficas para los diagramas articulatorios, las cuales detallamos a continuación:

- Representación del flujo de aire. El flujo de aire se indica con color azul celeste. En los diagramas articulatorios animados de la Figura 11 y la Figura 12, se puede observar como la zona que recorre el flujo de aire una vez sale de los pulmones queda marcada en color azul. En la Figura 11, por ejemplo, se puede observar que la zona de color azul abarca solamente la cavidad bucal, y que por tanto se trata de un sonido no nasal. En la Figura 12, en cambio, la zona de color azul abarca las cavidades bucal y nasal, lo que indica que el flujo de aire también es expulsado al exterior a través de las fosas nasales y que, por tanto, se trata de un sonido nasal.
- Representación de la ausencia o la presencia de sonoridad. Es sabido que en la producción de los sonidos sordos, las cuerdas vocales están ligeramente abiertas y que el flujo de aire pasa a través de las cuerdas vocales sin que vibren. En los diagramas articulatorios, la ausencia de sonoridad se ha representado mediante un triángulo, cuyo interior representa la glotis (véase la Figura 11). En los sonidos sonoros de los diagramas articulatorios la vibración de las cuerdas vocales se representa con una alternancia entre el triángulo descrito anteriormente y una línea horizontal, que simboliza el momento de cierre de las cuerdas vocales (véase la Figura 12).
- Representación del modo de articulación. La salida del flujo del aire al exterior a través de las cavidades supraglóticas es bien variada. En algunos casos, los articuladores obstruyen la salida del aire; es el caso de las consonantes oclusivas, en que las partículas de aire, representadas por bolitas, aparecen de color negro (véase la Figura 11). En otros casos, el acercamiento de los articuladores produce una constricción débil y el aire es expulsado al exterior con mayor facilidad; este es el caso de las aproximantes, en que las bolitas aparecen de color azul celeste (véase la Figura 13). En las vocales, el flujo de aire es expulsado sin casi ningún impedimento, y ello se indica con la ausencia de relleno en las bolitas (véase la Figura 14).

4.2. Otros recursos

Al margen de estos recursos, para que los neófitos puedan comprender las nomenclaturas utilizadas e identificar los sonidos representados en los diagramas articulatorios, es importante que conozcan tanto las partes vinculadas a la articulación de los sonidos como el vocabulario usado. Con este objetivo, se creó un esquema interactivo que permite localizar gráficamente las zonas y los órganos implicados en el proceso de producción de los sonidos (véase la Figura 15) y también un glosario de términos fonéticos y fonológicos, que fue adaptado a partir de Juliá-Muné (2003) por Esteve Valls; este glosario, que incluye alrededor de 450 conceptos —seguramente es de los más exhaustivos que hay en la red—, se ha

Figura 11: Representación de una oclusiva bilabial sorda.

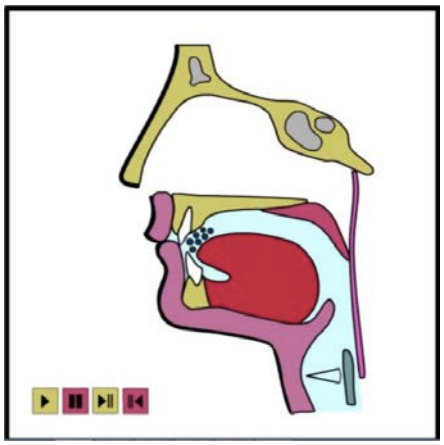


Figura 12: Representación de una nasal bilabial (sonora).

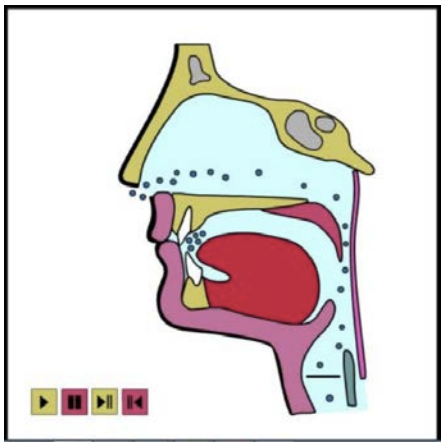
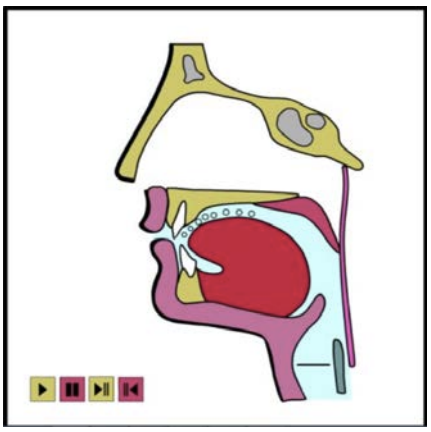


Figura 13: Representación de un modo de articulación aproximante.

Figura 14: Representación de un modo de articulación vocálico.



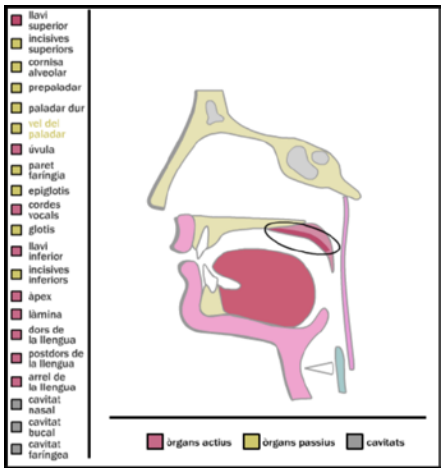
implementado de tal manera que se puede consultar alfabéticamente, lo que permite una consulta eficaz.

5. PROPIEDADES DE LA WEB

Desde un punto de vista del aprendizaje, se han buscado estrategias para que este sea interactivo, dinámico y para que acoja diferentes niveles de aproximación a la fonética. Por ello, en la web encontramos:

- Dinamismo. Con el objetivo de reforzar el dinamismo de la página, cada figura articulatoria animada va acompañada de unos comandos que permiten iniciar y suspender los gestos articulatorios de cada sonido (véanse las Figuras 11, 12, 13 y 14); además, también existe la posibilidad de ver el movimiento de los articuladores lentamente. La interfaz permite también comparar tres sonidos al mismo tiempo, una vez seleccionados a través de las tablas de consonantes y vocales.
- Interactividad y practicabilidad. Con la finalidad de que el usuario pueda afianzar parte de los contenidos, la web contiene también un conjunto de ejercicios de transcripción fonética y de reconocimiento de diagramas articulatorios. Los ejercicios de transcripción fonética incluyen diferentes niveles de dificultad, y permiten escuchar y transcribir sonidos en palabras aisladas, tanto reales como no reales (logótomos), y transcribir sonidos en secuencias de palabras. El usuario tiene la posibilidad de elegir el tipo de sonido que quiere practicar (de acuerdo con el modo de articulación, a saber, oclusivo, fricativo, africado, rótico, etc.), puede hacer la transcripción de las secuencias vacías a través de la misma tabla de sonidos de la web, y obtiene información inmediata sobre si la respuesta que ha dado es acertada o no. Los ejercicios de reconocimiento de diagramas articulatorios, que también tienen diferentes niveles de dificultad, permiten practicar el lugar de articulación y el modo de articulación de forma aislada o de forma conjunta. La autocorrección, obviamente, resulta efectiva para el trabajo autónomo.
- Niveles educativos. Hay que tener en cuenta, además, que el sitio web pretende ser útil para usuarios que buscan información más especializada. Por ello, las figuras articulatorias pueden visualizarse a través de imágenes y vídeos captados con resonancia magnética (véase la Figura 5) y también a través de palatogramas (véase la Figura 4). Por otra parte, la información acústica de los sonidos se presenta tanto con oscilogramas como espectrogramas y, en estos últimos, existe la posibilidad, como se indicaba, de visualizar, delimitado por líneas rojas, el sonido seleccionado (véase la Figura 3).

Figura 15: Diagrama articulatorio interactivo.



- Dialectología. Al inicio, la web solo contenía información relativa al catalán central, y poco después se incorporó información sobre catalán noroccidental. En los últimos años, se han incorporado dos variedades dialectales más: el valenciano central y el balear mallorquín. Para incorporar estas variedades, el proyecto ha contado con la colaboración de expertos de la Universidad de Valencia y de la Universidad de las Islas Baleares. Cabe decir que esta información es muy valiosa para las clases de fonética y de fonología. Por un lado, porque permite que el alumno conozca variedades dialectales diferentes a la suya y adquirir sensibilidad al respecto. Por otro lado, porque favorece que la web se emplee en las diferentes universidades de habla catalana, más allá de las establecidas en Barcelona. Y, finalmente, porque ayuda a proyectar el catalán en su totalidad entre la comunidad de usuarios extranjeros. A corto plazo, el objetivo es ampliar los contenidos, con información del catalán septentrional, el alguerés y otras variedades dialectales. Por el momento, se está trabajando en el alguerés.
- Traducción al inglés. Los contenidos de tipo metalingüístico de la web se han traducido de manera íntegra al inglés. La traducción de la web al inglés ha sido una acción esencial, en la medida que permite una fácil explotación del recurso desde los centros extranjeros donde se imparte enseñanza del catalán. De hecho, esta acción ha tenido una repercusión muy positiva en la accesibilidad y en la usabilidad de la web, como revelan los siguientes datos: de enero de 2014, cuando se publicó la versión en inglés, a la actualidad (diciembre de 2016), la web ha sido visitada por unos 58.118 usuarios (principalmente de España [50 325 usuarios], pero también de los Estados Unidos [2 962 usuarios], del Reino Unido [724 usuarios], de Alemania [536 usuarios], Francia [492 usuarios], Rusia [73 usuarios], Italia [245 usuarios], China [228 usuarios], Méjico [144 usuarios], Canadá [174 usuarios], etc.) y se han consultado cerca de 284 948 páginas (véase el § 6).
- Visita guiada. La web también contiene una visita guiada, que ayuda al nuevo usuario a conocer los espacios que contiene, comentados en los apartados anteriores, y que lo orienta sobre cómo usar los materiales. La visita guiada es un recurso opcional, del cual el usuario puede prescindir en caso de que no lo necesite.
- Accesibilidad. La web también se ha remodelado para que sea accesible desde nuevas plataformas digitales, como las tabletas y los teléfonos móviles, para que sea visible desde cualquier tipo de pantalla, y para que sea consultable desde los sistemas operativos más comunes, esto es, tanto desde Mac como desde PC. Esta acción ha implicado, por ejemplo, convertir todos los diagramas articulatorios que estaban en formato Flash a formato de vídeo (YouTube). Esta tarea, que ya está terminada, ha repercutido y repercutirá muy positivamente en la accesibilidad y la usabilidad de la web.

6. IMPACTO Y DIFUSIÓN

El impacto del espacio Els sons del català desde su presentación oficial en el año 2009 ha sido significativo. Apareció en los principales medios de comunicación catalanes (TV3, 324 [Espacio Internet], Cataluña Radio, RAC1, Vilaweb, etc.), y desde entonces está referenciado en un número considerable de portales, como el ésAdir, Softcatalà, PuntTIC, Edu3.cat, entre otros. Para obtener información detallada sobre la presencia de la web en los medios de comunicación en general y en Internet en particular, se puede visitar http://www.ub.edu/sonscatala/ca/noticies. El impacto de la web ha sido también notable en el ámbito internacional; es por ello que las responsables del proyecto han sido invitadas en varias ocasiones para presentarlo en el contexto de los programas internacionales de enseñanza de catalán, como los lectorados ofrecidos por el Instituto Ramon Llull. Las visitas a la web son también bastante elocuentes. Durante el último año, la web ha registrado un total de 26 132 usuarios, 36 976 sesiones y 116 121 páginas visitadas (fuente: Google Analytics). En cuanto a la procedencia geográfica de las visitas, resulta muy variada. La mayoría de visitas provienen de España, pero, como se ha indicado en el apartado anterior, también hay muchas de otras partes del mundo.

7. USUARIOS Y APLICACIONES

Desde los inicios, la herramienta Els sons del català ha sido concebida pensando en una tipología muy variada de usuarios. Desde profesores y alumnos de la enseñanza superior, pasando por profesores y alumnos de secundaria, maestros y alumnos de primaria, profesores de catalán para adultos, aprendices de catalán, comunidad inmigrante de Cataluña, profesores y alumnos de fuera de Cataluña, hasta profesionales de la corrección, de los medios de

comunicación o de la logopedia. Con todo, actualmente la web se utiliza, de manera prioritaria, en la enseñanza de la fonética y de la fonología catalanas de los diferentes grados de las universidades catalanas que imparten estas materias, esto es, en la enseñanza superior, por lo que no solo repercute positivamente en el seno de la comunidad científica, sino también en la transmisión del conocimiento de esta comunidad a los alumnos. Sabemos, también, que Els sons del català se utiliza en universidades de fuera de Cataluña como herramienta para la enseñanza de catalán como lengua extranjera. Y eso contribuye, obviamente, a la proyección del catalán en todo el mundo.

Hay que tener en cuenta, de hecho, que la iniciativa surgió para dar respuesta a las nuevas necesidades de los docentes, a raíz de la intensificación de las relaciones internacionales en el ámbito universitario y post-universitarios (mediante programas de movilidad y convenios) y los cambios impulsados en el marco del Espacio Europeo de Educación Superior, que ha fomentado, para mejorar el rendimiento académico del alumnado, la explotación de las nuevas tecnologías, la elaboración de materiales didácticos en lengua inglesa y el estudio y aprendizaje autónomos. También surgió para mejorar la enseñanza del catalán y adaptarla a las necesidades vigentes del alumnado en sus diferentes etapas, y para incrementar la proyección de la lengua catalana en todo el mundo. Finalmente, para facilitar el uso y conocimiento del catalán entre las comunidades de no catalanohablantes en Cataluña, a raíz de la amplia transformación demográfica, étnica, cultural y lingüística que ha experimentado la sociedad catalana en los últimos años. Hay que tener en cuenta, en este sentido, que según datos de 2014 del IDESCAT, el número población extranjera en Cataluña supera el millón de habitantes y la cifra de alumnos de origen extranjero en Cataluña se acerca a los 165 000.

8. OBSERVACIONES FINALES

En resumen, el proyecto en que se inscribe Els sons del català ha sabido hacer confluir el conocimiento que tenemos de la fonética y fonología catalanas y una metodología de enseñanza de estas materias adaptada a las necesidades del alumnado actual en un espacio web de acceso libre. Diversos estudios realizados sobre la efectividad del portal (Anónimo / Varios autores 2013; Anónimo / Varios autores 2014) demuestran que la herramienta resulta eficaz como material de refuerzo y de trabajo tanto en asignaturas de descripción general de la lengua catalana como en asignaturas de fonética y fonología catalanas.

La herramienta digital Els sons del català puede entenderse como un avance significativo en el ámbito de los recursos en abierto en el área de la enseñanza y aprendizaje de la fonética y la fonología del catalán, puesto que hasta la fecha el material para aprender estas materias se circunscribía sobre todo a recursos en soporte papel y el material accesible en Internet estaba esencialmente focalizado a la enseñanza del catalán escrito.

Con estos materiales y los que van a completar el espacio web, se espera seguir avanzando en el ámbito de la transferencia de conocimientos de la fonética y de la fonología a la sociedad, puesto que las disciplinas, poliédricas por naturaleza, están presentes, sea de manera directa o indirecta, en el día a día de muchas personas.

9. REFERENCIAS

Julià-Muné, Joan (2003). *Diccionari de fonètica*. Barcelona: Edicions 62.

Subsidia: Tools and Resources for Speech Sciences

14

Bayesian strategies for likelihood ratio computation in forensic voice comparison with automatic systems

Daniel Ramos[1], Juan Maroñas-Molano[1] and Alicia Lozano-Diez[1]

[1] Universidad Autónoma de Madrid
e-mail: daniel.ramos@uam.es, jmaronasm@gmail.com, alicia.lozano@uam.es

Citation / Cómo citar esta publicación: Ramos, D., Maroñas-Molano, J., & Lozano-Diez, A. (2019). Bayesian strategies for likelihood ratio computation in forensic voice comparison with automatic systems. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 89–95). Málaga: Universidad de Málaga.

ABSTRACT: This paper explores several strategies for Forensic Voice Comparison (FVC), aimed at improving the performance of the LRs when using generative Gaussian score-to-LR models. First, different anchoring strategies are proposed, with the objective of adapting the LR computation process to the case at hand, always respecting the propositions defined for the particular case. Second, a fully-Bayesian Gaussian model is used to tackle the sparsity in the training scores that is often present when the proposed anchoring strategies are used. Experiments are performed using the 2014 i-Vector challenge set-up, which presents high variability in a telephone speech context. The results show that the proposed fully-Bayesian model clearly outperforms a more common Maximum-Likelihood approach, leading to high robustness when the scores to train the model become sparse.

Keywords: likelihood ratio; forensic voice comparison; automatic speaker recognition; anchoring; Gaussian; fully-Bayesian.

RESUMEN: Este artículo explora varias estrategias de comparación forense de voces, con el objetivo de mejorar la adecuación de las razones de verosimilitud cuando se usan modelos generativos gaussianos. En primer lugar, se proponen varias estrategias de anclaje para adaptar el proceso de computación de las razones de verosimilitud a cada caso, y siempre respetando las proposiciones definidas para el caso particular. En segundo lugar, se usa un modelo gausiano completamente bayesiano para abordar la escasez de datos que suele caracterizar los entrenamientos cuando se usan las estrategias de anclaje propuestas. Presentamos una serie de experimentos que parten de las condiciones propuestas para el reto i-Vector de 2014, que presenta una gran variabilidad en un contexto de habla telefónica. Los resultados muestran que el modelo completamente bayesiano resulta mucho mejor que un enfoque más habitual basado en la Máxima Verosimilitud, lo que significa un modelo más robusto cuando los datos que se usan para entrenar el modelo son escasos.

Palabras clave: razón de verosimilitud; comparación forense de voces; reconocimiento automático de locutor; anclaje; gaussiano; completamente bayesiano.

1. INTRODUCTION

In forensic voice comparison (FVC) using automatic systems, a score is typically transformed into a likelihood ratio (LR) by using some probabilistic model (Drygajlo & Haraksim, 2016; Gonzalez-Rodriguez et al., 2007). Recently, this methodology has been proposed as the recommended way of reporting in court using automatic speaker recognition systems, in the context of the speech and audio laboratories belonging to the European Network of Forensic Science Institutes (ENFSI) (Drygajlo et al., 2016). Also, this is the typical strategy for LR computation in other biometric systems, where a score is the usual output when comparing two biometric specimens (Ramos et al., 2016). This approach can be also used in general when a model or method previously computes a score among two evidential materials, such as two handwritten documents or two chemical profiles (Bolck, Ni, & Lopatka, 2015; Hepler et al., 2011).

Many techniques have been proposed in the past to do this score-to-LR transformation. Perhaps the most straightforward way is to assign a probability distribution[1] to the score from the automatic system, given each one of the propositions in the forensic case. As score normalization techniques with Gaussianization properties are frequently an intrinsic stage of automatic speaker recognition systems (Navratil & Ramaswamy,

[1] In a domain of continuous scores, it will be a probability density function; and a probability mass function if the scores are discrete.

2003), the use of Gaussian distributions appears as a sensible choice (Villalba & Brümmer, 2014). Anyway, probabilistic distributions are typically assigned to scores using training strategies like Maximum Likelihood (ML) or Maximum a Posteriori (MAP) (Alberink, de Jongh, & Rodriguez, 2014; Ramos et al., 2006). Another popular approach is logistic regression, also trained with ML or MAP (Brümmer et al., 2006), because of its good robustness properties over the whole score range, and the possibility of regularization, the latter affecting the calibration of the final LRs negatively. Recently, fully-Bayesian strategies have been proposed, in order to cope with the sparsity in the amount of scores available for training, while showing good calibration performance (Villalba & Brümmer, 2014).

Another main problem in FVC is the use of the specimens and databases to compute a better LR in a given case. It is well known that speech is extremely variable according to many conditions (such as phonetic content, acoustic environment, emotional state, transmission channel and so on), and it has been also observed that this variability affects the score range quite seriously (Mandasari et al., 2013; Perez-Gomez et al., 2010). Therefore, in order to compute a LR that presents good performance for a given case, two actions may be considered. First, the automatic speaker recognition system must implement powerful session variability compensation techniques in order to reduce the variability of the scores to be transformed into LRs. Second, the LR model itself must take into account the conditions of the speech in each case, while respecting the propositions and conditioning information of the case itself, because the distribution of the training scores has to fit the distribution of the score in the case. The LR model can address the latter by the selection of the data to compute the scores used to train the LR model. This data selection process and the subsequent strategy to compute the scores has been dubbed *anchoring* (Alberink, de Jongh, & Rodriguez, 2014; Hepler et al., 2011). In this paper, we will propose some anchoring strategies, showing experiments that support their use in challenging FVC scenarios.

An additional problem with some *anchoring* strategies is that they typically result in a very small amount of scores to train the LR model. Unfortunately, these schemes are the most typically proposed ones in FVC (Drygajlo & Haraksim, 2016). The sparsity in the amount of scores typically affects the likelihood of the prosecution proposition, in the numerator of the LR, because the available data from a given suspect is often very limited in a case. In order to overcome this problem, we will test the use of a Gaussian fully-Bayesian model to cope with the uncertainty due to data sparsity.

Thus, the contribution of this work is two-fold: first, different anchoring schemes are tested, to show the adequacy of taking into account the conditions of the case in the anchoring process. Second, fully-Bayesian Gaussian models are proposed as an efficient way of reducing the problem of data sparsity in the training scores. Both experimental contributions are set-up in a highly challenging scenario of speech variability under telephone conditions: the NIST2014 i-Vector Challenge.

The paper is organized as follows: Section 2 describes the anchoring schemes used in this paper, and their motivation. Section 3 will introduce the models to be compared in this article, namely Gaussian ML and Gaussian Fully-Bayesian. Finally, Section 4 describes the database and experimental protocol in the highly variable context of the 2014 NIST i-Vector Challenge competition, and shows the results supporting the research hypothesis. Conclusions are finally drawn in Section 5.

2. HANDLING SPEECH DATA IN FVC USING AUTOMATIC SYSTEMS

2.1. Propositions in FVC

A FVC case has the following typical elements. On the one hand, there are some questioned speech materials (namely, one or more questioned recordings) of disputed origin, also known as *trace*, and in general of an incriminatory nature. On the other hand, a suspect identified on the basis of other information, and from whom some control materials are recorded, namely one or more *reference* or *control* recordings. The problem consists in expressing the value of such evidence with respect to two propositions defined in the case. In this context, the identity of the suspect is typically known, as well as some other features that can be extracted from contextual information in the case, or from the speech materials themselves. This motivates the following typical definition of the propositions in FVC:

- H_p: the questioned materials come from the suspect.
- H_d: the questioned materials do not come from the suspect, but another individual from a given population of potential sources.

Note that H_p and H_d differ in the assumption of the origin of the trace q, which is unknown in the case. Therefore, conditioning in the propositions will imply a change of the source of the trace in the generation of training scores. Also, it is assumed that the suspect, and no other possible individual, has originated the reference materials in both cases, because the suspect is referred to in both propositions themselves. Such propositions have been dubbed *source-specific* or, if the potential source is a person, respectively *person-specific* (Ramos et al., 2016). They are the most common in FVC, where a suspect of known identity has been usually accused by a court of law.

We define H as the random variable representing the proposition, with alphabet $\{H_p, H_d\}$. Since the score generated by the comparison between the questioned and control materials must be evaluated in the context of those competing propositions, the likelihood ratio (LR) formula naturally arises (Drygajlo & Haraksim, 2016; Drygajlo et al., 2016; Gonzalez-Rodriguez et al., 2007):

Ramos, D., Maroñas-Molano, J. and Lozano-Diez, A. (2019). Bayesian strategies for likelihood ratio computation in forensic voice comparison with automatic systems. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 89–95). Málaga: Universidad de Málaga.

(1) $$LR = \frac{p(s_c|H_p)}{p(s_c|H_d)}$$

where s_c is the score, *observed* from the comparison of the questioned and control materials, as given by the automatic system; and $p(.|.)$ denotes a conditional probability density. In this way, the value of the evidence is given by the LR, and can be reported according to common procedures (Drygajlo et al., 2016).

2.2. Generating scores: *anchoring*

In order to assign the densities in (1), two sets of training scores must be generated. Each training score will be accompanied by a *class label*, where classes are the H_p and H_d values from variable H in this forensic scenario. A training score s_i and its corresponding class label H_i are represented as (s_i,H_i). The total set of training scores will be $S = \{S_p,S_d\}$, where $S_p = \{(s_p^{(1)},H_p),\ldots,(s_p^{(Np)},H_p)\}$ for the density in the numerator (conditioned to H_p) and $S_d = \{(s_d^{(1)},H_d),\ldots,(s_d^{(Nd)},H_d)\}$ for the density in the denominator (conditioned to H_d). The process for speech data selection and score generation of training scores is known as anchoring (Alberink, de Jongh, & Rodriguez, 2014; Hepler et al., 2011), and must consider the following facts:

(1) The definition of the propositions will define how the scores S_p and S_d must be generated.
(2) The high variability of the speech signal previously mentioned seriously compromise the distribution of the scores used to train LR models (Mandasari et al., 2013; Perez-Gomez et al., 2010). Therefore, the generation of the scores must be done in the most similar speech conditions as those in the case, otherwise the densities will not represent the observed score s_c in the case at hand (Drygajlo & Haraksim, 2016).

Let s_c be generated from the comparison of the questioned speech q and the reference speech r. Thus, $s_c = \Delta(q,r)$, where $\Delta(\ldots)$ represents the score computation algorithm of the automatic speaker recognition system[2]. Also, it is assumed that $\Delta(q,r) = \Delta(r,q)$, as it happens with many speaker recognition systems (Matejka et al., 2011) based on i-Vectors. Then, from both facts described above, some conclusions can be extracted for a FVC case using automatic systems.

(1) In order to satisfy H_p, for the density in the numerator of the LR, the training scores must be generated using traces from the suspect, known as *speech controls* of *pseudo-traces*, to be compared with reference speech also from the suspect. Moreover, in FVC a *person-specific* proposition implies that using same-person scores coming from other individuals different from the suspect will not be adequate, since the distribution of scores of different individuals is known to be highly variable (Doddington et al., 1998). Therefore, comparing speech controls $\{q_p^{(i)}\}$ to other reference speech materials from the suspect, namely $\{r_p^{(j)}\}$, $S_p = \{\Delta(q_p^{(i)},r_p^{(j)})\}$ is obtained, with $|S_p| = N_p$.

(2) In order to fit the conditions of the case, both the trace and the reference speech used to generate training scores should be selected according to the following criteria.
- Reference speech: As the identity of the suspect is known to be the one in r, $\{r_p^{(j)}\}$ should consist of speech data from the suspect in conditions as close as possible to the conditions of r. In the limit, the best possible fitting is $r_p^{(j)} = r$, and to use a number N_p of speech controls $q_p^{(i)}$.
- Trace: The speech controls $q_p^{(i)}$ must have conditions as close as possible to q. Otherwise, the model obtained using S_p will not represent the model that could generate $s_c = \Delta(q,r)$ if r and q come from the same source, because a change in the conditions of the trace q will most probably affect the score distribution.

(3) In order to satisfy H_d, for the density in the denominator of the LR, the training scores must be generated using traces coming from individuals being potential origins of the trace. In forensic practice, these potential origins are assumed to be drawn from a so-called *population* of individuals. Also, since the definition of the proposition is *person-specific*, it is assumed that the suspect, and no other individual, has generated the reference speech recordings in the case. Therefore, the scores S_d must be generated by traces $q_d^{(i)}$ generated by individuals from the population of potential origins, compared to reference speech segments $r_d^{(j)}$ from the suspect. Thus, $S_d = \{\Delta(q_d^{(i)},r_d^{(j)})\}$, with $|S_d| = N_d$.

(4) In order to fit the conditions of the case, the same rationale as for S_p applies. Thus:
- Reference speech: As the identity of the suspect is known to be the one in r, the best fitting in the conditions of $\{r_d^{(j)}\}$ is to use speech data from the suspect in conditions as close as possible to the conditions of r. In the limit, the best possible fitting is to always use $r_d^{(j)} = r$, and a number N_d of traces $q_d^{(i)}$.
- Trace: The $q_d^{(i)}$ traces must have conditions as close as possible as in q. Otherwise, the model obtained using S_d will not represent the model that could generate $s_c = \Delta(q,r)$ if q comes from other individual than the suspect.

According to the aforementioned conclusions, in this work we propose two anchoring schemes.

(1) *Suspect-anchored* (SA):

[2] Although the notation might suggest similarity, scores in speaker recognition systems most often also include population models and typicality. In fact, i-Vector PLDA systems output log-likelihood-ratios, but presenting poor calibration (Matejka et al., 2011).

- Scores in $S_p = \{\Delta(q_p^{(i)},r_p^{(j)})\}$ are anchored to the speaker, and therefore $\{q_p^{(i)}\}$ are speech controls from the suspect in conditions as close as possible to the ones in q, and $\{r_p^{(j)}\}$ are speech segments from the suspect in conditions as close as possible to the ones in r.
- Scores in $S_d = \{\Delta(q_d^{(i)},r_d^{(j)})\}$ are also anchored to the speaker, and therefore $\{q_d^{(j)}\}$ will be speech segments from other speakers from the population of potential origins, in conditions as close as possible to the ones in q; and $\{r_d^{(j)}\}$ will be speech segments from the suspect in conditions as close as possible to the ones in r. In fact, $\{r_p^{(j)}\} = \{r_d^{(j)}\}$.

(2) *Reference-anchored* (RA):

- Scores in $S_p = \{\Delta(q_p^{(i)},r)\}$ are anchored to the reference speech r, and therefore $\{q_p^{(i)}\}$ are speech controls from the suspect in conditions as close as possible to the ones in q, and $r_p^{(j)} = r$.
- Scores in $S_d = \{\Delta(q_d^{(i)},r)\}$ are also anchored to the reference speech r, and therefore $\{q_d^{(j)}\}$ will be speech segments from other speakers from the population of potential origins, in conditions as close as possible to the ones in q; and $r_d^{(j)} = r$ [3].

One of our research hypotheses is that RA will offer better performance than SA, since the fitting to the conditions in r is the most perfect one for RA. Also, as it can be seen, the use of these anchoring schemes may lead to a substantially high number of scores N_d for assigning the denominator of the LR, because $\{q_d^{(i)}\}$ can be found to be large. However, N_p can be very low, because in FVC the amount of data from the suspect is usually very limited. Therefore, data sparsity must be addressed mainly for S_p.

3. GENERATIVE MODELS FOR LR COMPUTATION FROM SCORES

Computing the LR can be done following discriminative and generative approaches. An example of discriminative approach is logistic regression (Brümmer et al., 2007; Morrison, 2013), widely used in automatic speaker recognition. However, generative approaches have been recently proposed as a robust alternative to other methods (Villalba & Brümmer, 2014). Reasons are given below.

In this work we follow two generative approaches, where the logLR is computed from the ratio between the likelihood distributions, following equation (1).

These likelihood distributions are computed from the joint probability density $p(s,H)$ or more exactly from the model representing the joint probability density $\hat{p}(s,H|\theta)$ where for convenience we will replace $\hat{p}$ by p and θ represents the parameters of the model as a vector. In this work we implement these models using Maximum Likelihood (ML) and Bayesian Inference (BI).

3.1. Maximum Likelihood

Let $S = \{(s^{(i)},H_i)\}^N_{i=1}$ be a dataset drawn i.i.d. from the model distribution, that is, each sample is independent from the samples in the same class and from the samples in the other class (does not provide information about the other class). Thus, the likelihood function is defined as:

$$p(S|\theta) = \prod_{i=1}^{N} p(s_i|H_i,\theta)\cdot p(H_i) \tag{2}$$

One can show that under this condition we can optimize each term of the likelihood function: $p(x|H_p)$, $p(x|H_d)$, $p(H_p)$, $p(H_d)$; independently. For an example see Bishop (2006, section 4.2.2). For the task we address we are only interested in estimating the parameters of the likelihood function, because the prior probabilities are not the responsibility of the forensic evaluation process, and we can express our model thus:

$$p(S = \{s^{(1)},s^{(2)},..s^{(N)}\}|\theta) = \prod_{i=1}^{N} p(s_i|\theta) \tag{3}$$

Setting the derivative to zero and finding the maximum value we end up with a choice of the parameters. For a univariate Gaussian model, $\theta = (\mu_j,\sigma_j^2)$ with $j \in \{p,d\}$, as:

$$\mu_j = \frac{1}{N_j}\sum_{i\in N_j} s_j^{(i)} \tag{4}$$

$$\sigma_j^2 = \frac{1}{N_j - 1}\sum_{i\in N_j} (s_j^{(i)} - \mu_j)^2 \tag{5}$$

3.2. Bayesian Parameter Interference

The approach of Bayesian inference starts by assuming we want to assign the predictive distribution $p(s|S)$, that is a distribution over the random variable s whose form depends directly on the data set and not on a set of parameters. The key idea is marginalization over all the possible parameters of the underlying model. Hence:

$$p(s|S) = \int_{\forall\theta} p(s|\theta)\cdot p(\theta|S)d\theta \tag{6}$$

The latter is no more than an expectation of a given parametric function, in this case Gaussian, with a parameter distribution, $p(\theta|S)$ that depends on our data set S. For this point we should focus all our attention on the second factor of the product, that is, obtaining a parameter distribution from S. We can do this by applying Bayes' theorem and defining a prior knowledge over the parameters. Thus:

[3] As described, the following proposed suspect-anchoring and reference-anchoring schemes make a distinction about how the scores are generated, not a distinction in the kind of statistical model that it used.

Ramos, D., Maroñas-Molano, J. and Lozano-Diez, A. (2019). Bayesian strategies for likelihood ratio computation in forensic voice comparison with automatic systems. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 89–95). Málaga: Universidad de Málaga.

$$p(\theta|S) = \frac{p(S|\theta) \cdot p(\theta)}{p(S)} \quad (7)$$

Where $p(S)$ is the area under the numerator function in the RHS of the equation. From this point we shall make some assumptions for having an analytic closed form for $p(s|S)$. First is that our model $p(s|\theta)$ is Gaussian. This means, taking in account that data from class H only give information about that class and that samples are drawn iid, that we can factorize $p(S|\theta)$ as we have already done in Eq. 3. Our last assumption is that the prior over the parameters $p(\theta)$ is a Gaussian-gamma function and thus we can rewrite the expression as the conjugate prior $p(\theta|\phi)$ where hyperparameters ϕ govern the form of the prior knowledge about this function. The Gaussian-gamma function is a conjugate prior of the Gaussian function and therefore the posterior of the parameter is also Gaussian-gamma. The integral from Eq. 6 has an analytic form which is a Student's t. We can find the details about the inference process in Minka (1998) and Brümmer (2011), where different approaches of the problem yield the same result. Minka (1998) uses a non-informative parameter prior when the entropy of the parameter is maximum, and Brümmer (2011) uses a Gaussian gamma. By setting the parameters of the Gaussian-gamma to specific values we can have a Gaussian-gamma that tends to be non-informative. In our work we use the approach in Brümmer (2011), assuming maximum entropy for the parameters of the model.

When we have data sparsity, Bayesian inference (BI) (fully-Bayesian, as it is implemented here) is better than ML. When the number of data samples is big, the BI inference tends to the same result as ML. We see this effect reflected in the tails of the Student's t distribution. We observe that in the data space where we have enough samples we have a good likelihood fit, but with sparse data the tails of the Student's t are heavier. This is reflected in the likelihood ratio giving a preference for one class in the well represented data space and no preference in places without representation. Figure 1 and Figure 2 show BI vs ML inference of the parameters. First we show the likelihood distributions and then the function that transforms scores into log-likelihood-ratios (LLRs) for a given dataset. It can be seen that the LLRs obtained by the fully-Bayesian model (BI) are much more limited when the scores become extreme. This is a desired effect in forensic science, because very big LR values make little sense for automatic speaker recognition systems. However, due to data sparsity, extremely big values of the LR are allowed with ML, which is a quite undesirable effect.

Figure 1: Likelihood Distributions. Dashed line represents ML densities and continuous line represents BI densities.

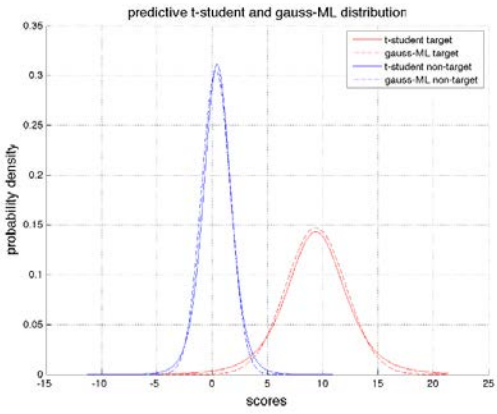
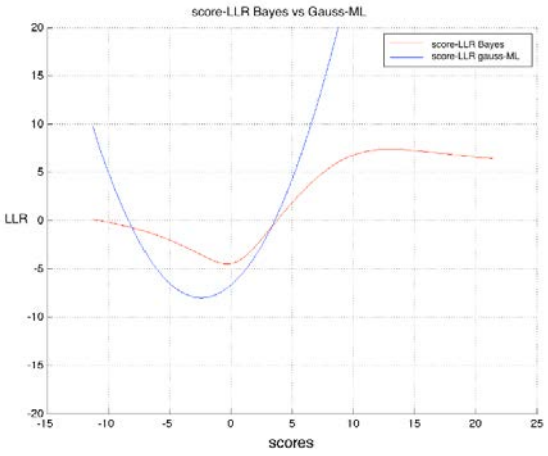


Figure 2: Function that transforms scores into log-likelihood-ratios (LLRs) for a given training dataset of scores.



4. EXPERIMENTS

4.1. Database and automatic speaker recognition system

For the experiments in this work, we used the data provided by NIST for the 2014 Speaker Recognition i-Vector challenge[4]. For this challenge, 600-dimensional i-Vectors were provided from conversational telephone speech data available for previous NIST Speaker Recognition Evaluations (SRE's), from 2004 to 2012. Different amounts of speech were used to compute the i-Vectors, following a log normal distribution with mean of 39.58 seconds. From these i-Vectors, scores were generated using PLDA (Matejka et al., 2011). To develop the PLDA system, we used i-Vectors from utterances with more than 30 seconds of speech in the development set, and the ground truth labels provided by NIST after the evaluation. This subset consists of 17424 i-Vectors, from 3769 different speaker identities. The evaluation data provided for this challenge comprises five i-Vectors for each target speaker model, and single i-Vectors representing test segments. The number of target speaker models was 1306, and the number of test i-Vectors, 9634, resulting in over 12 million trials.

[4] https://www.nist.gov/sites/default/files/documents/itl/iad/mig/sreivectorchallenge_2013-11-18_r0.pdf

Ramos, D., Maroñas-Molano, J. and Lozano-Diez, A. (2019). Bayesian strategies for likelihood ratio computation in forensic voice comparison with automatic systems. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 89–95). Málaga: Universidad de Málaga.

4.2. Experimental protocol

For the experiments in this work, we have used the scores generated in the i-Vector challenge in different ways depending on the proposed anchoring scheme, as described below:

(1) Suspect-anchored (SA):

- S_p is generated by drawing scores from a pool of scores including all possible combinations of two utterances from the suspect, without including the speech segment(s) of the suspect that might be present in the case. Different amounts of N_p scores are drawn, in order to simulate data sparsity in S_p.
- S_d is generated by comparing all utterances from the suspect with all utterances from other identities, excluding all utterances present in the case.

(2) Reference-anchored (RA):

- S_p is generated by drawing scores from a pool of scores $\Delta(q_p^{(i)},r)$, where r is the reference suspect speech in the case, and $\{q_p^{(i)}\}$ are the remaining utterances from the suspect, excluding q when it comes from the suspect. Different amounts of N_p scores are drawn, in order to simulate data sparsity in S_p.
- S_d is generated by comparing r, i.e. the reference suspect speech in the case, with a number of speech segments from other identities, excluding all utterances present in the case.

Notice that, since score sparsity rarely affects S_d, N_d is not constrained in order to focus on the sparsity effects in S_p. Finally, in order to use the same speech data for every value of N_p, only suspects with more than 10 utterances have been selected. All in all, to measure performance, a number of 18192 same-origin and 11560 different-origin LR values have been computed as a minimum, with even more values depending on the value of N_p and the anchoring scheme.

4.3. Results

Figure 3 shows the results in terms of C_{llr} of the ML and fully-Bayesian schemes in the suspect-anchored (SA) scheme; and Figure 4 shows the corresponding results for the reference-anchoring (RA) scheme. Results show that the fully-Bayesian approach clearly outperforms ML for all sizes of the training score set N_p. Also, the value of C_{llr} is much lower for this model, especially for lower values of N_p. This supports the hypothesis that the fully-Bayesian model allows the incorporation of uncertainty, due to score sparsity, to the LR in an effective way. This is especially relevant for the lowest values of N_p, because the ML approach yields much higher values of C_{llr}, which means much poorer performance. Moreover, $C_{llr} < 1$ always obtains for the fully-Bayesian model, which means that the approach is always informative for evidence evaluation. This is not the case of ML, since C_{llr} is much bigger than 1 for many values of N_p.

Figure 3: C_{llr} values for the ML and fully-Bayesian models for the suspect-anchored (SA) scheme, as a function of N_p.

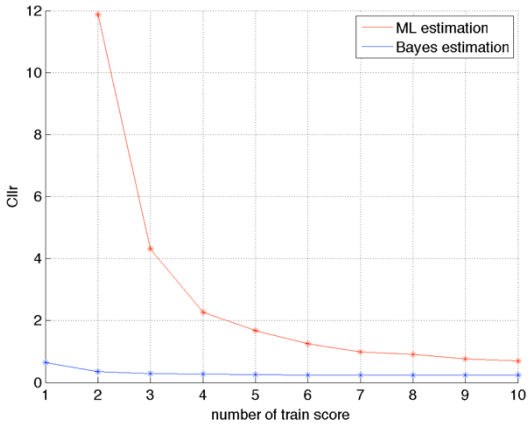
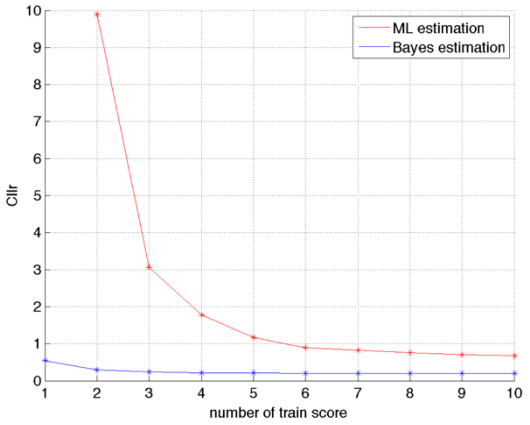


Figure 4: C_{llr} values for the ML and fully-Bayesian models for the reference-anchored (RA) scheme, as a function of N_p.



Moreover, the comparison of both figures shows that the RA approach outperforms SA. This was expected, since RA allows the training scores to resemble the conditions of the score in each case in a much adequate way.

5. CONCLUSIONS

In this work, the use of fully-Bayesian Gaussian models have proven to be adequate for forensic voice comparison using automatic systems. This is particularly true when compared to widely used ML models, that have shown to be very sensitive to sparsity in the training scores, a situation that often happens to training scores under H_p. On the other hand, fully-Bayesian methods effectively cope with the lack of data by incorporating the associated uncertainty, leading to much more moderate LR values, and drastically improving the value of C_{llr}. In fact, performance with fully-Bayesian models is always better than not reporting the *LR* (meaning $C_{llr} < 1$ always). Given the difficulty of the NIST i-Vector Challenge task, we can recommend fully-Bayesian methods to compute LRs in forensic practice, although more research is needed to compare the proposed approach with other models.

Ramos, D., Maroñas-Molano, J. and Lozano-Diez, A. (2019). Bayesian strategies for likelihood ratio computation in forensic voice comparison with automatic systems. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 89–95). Málaga: Universidad de Málaga.

One interesting issue in this article is related to the anchoring schemes. We have reported experiments where a suspect-anchored and reference-anchored yields adequate (i.e. informative) and robust performance, the latter outperforming the former. Although this result confirms the research hypotheses, recent discussion on this topic motivates further research (Alberink, de Jongh, & Rodriguez, 2014; Hepler et al., 2011).

6. REFERENCES

Alberink, I., de Jongh, A., & Rodriguez, C. (2014). Fingermark evidence evaluation based on automated fingerprint identification system matching scores: the effect of different types of conditioning on likelihood ratios. *Journal of Forensic Sciences, 59*(1), 70–81.

Bishop, C. M. (2006). *Pattern recognition and machine learning (Information Science and Statistics)*. Secaucus, NJ: Springer-Verlag.

Bolck, A., Ni, H., & Lopatka, M. (2015). Evaluating score-and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison. *Law, Probability and Risk, 14*(3), 243–266.

Brümmer, N. (2011). *Fully Bayesian score calibration assuming Gaussian distributions*. Available at https://sites.google.com/site/nikobrummer/.

Brümmer, N. et al. (2007). Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. *IEEE Transactions on Audio, Speech and Signal Processing, 15*(7), 2072–2084.

Doddington, G. et al. (1998). Sheeps, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In *Proceedings of ICSLP*.

Drygajlo, A. & Haraksim, R. (2016). Biometric evidence in forensic automatic speaker recognition. In M. Tistarelli & C. Champod (Eds.), *Handbook of biometrics for forensic science*. Dordrecht: Springer.

Drygajlo, A. et al. (2016). Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition. In *Proceedings of ENFSI*.

Gonzalez-Rodriguez, J. et al. (2007). Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Transactions on Audio, Speech and Language Processing, 15*(7), 2072–2084.

Hepler, A. B. et al. (2011). Score-based likelihood ratios for handwriting evidence. *Forensic Science International, 219*(3), 129–140.

Mandasari, M. I. et al. (2013). Quality measure functions for calibration of speaker recognition systems in various duration conditions. *IEEE Transactions on Audio, Speech, and Language Processing, 21*(11), 2425–2438.

Matejka, P. et al. (2011). Full-covariance UBM and heavy-tailed PLDA in i-Vector speaker verification. In *Proceedings of ICASSP*.

Ramos, D. et al. (2016). From biometric scores to forensic likelihood ratios. In M. Tistarelli & C. Champod (Eds.), *Handbook of biometrics for forensic science*. Dordrecht: Springer.

Ramos, D. et al. (2006). Suspect-adapted MAP estimation of within-source distributions in generative likelihood ratio estimation. In *Procedings of Odyssey*.

Villalba, J. & Brümmer, N. (2014). Bayesian calibration for forensic evidence reporting. In *Proceedings of Interspeech*.

Ramos, D., Maroñas-Molano, J. and Lozano-Diez, A. (2019). Bayesian strategies for likelihood ratio computation in forensic voice comparison with automatic systems. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 89–95). Málaga: Universidad de Málaga.

Subsidia: Tools and Resources for Speech Sciences

EMULANDO: Corpus de habla con acento no nativo auténtico y disimulado

José María Lahoz-Bengoechea[1], Juana Gil Fernández[2] y Clara Luna García García de León[3]

[1] Universidad Complutense de Madrid
[2] Instituto Cervantes de Lyon
[3] Universidad Nacional de Educación a Distancia
e-mail: jmlahoz@ucm.es, juanagilfernandez@gmail.com, claralunagarcia@gmail.com

Citation / Cómo citar este artículo: Lahoz-Bengoechea, J. M., Gil Fernández, J., & García García de León, C. L. (2019). EMULANDO: Corpus de habla con acento no nativo auténtico y disimulado. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 97–101). Málaga: Universidad de Málaga.

RESUMEN: Los criminales normalmente intentan camuflar su identidad recurriendo a algún tipo de distorsión de su habla normal (por ejemplo, haciéndose pasar por extranjeros). EMULANDO es un corpus de habla en español con acento extranjero real y fingido (inglés, francés y ruso). Este tipo de datos permite estudios y aplicaciones muy interesantes para el ámbito de la fonética forense. En el artículo se comentan algunas claves que pueden revelar tales tipos de disimulo: la desviación estándar de los formantes vocálicos, el timbre y la duración de las vocales de relleno, las características acústicas de algunas fricativas, o ciertos patrones entonativos. Además, el corpus puede utilizarse como material de entrenamiento para los sistemas biométricos automáticos.

Palabras clave: habla disimulada; acento extranjero; fonética forense; español.

ABSTRACT: Criminals tend to disguise their identity by distorting their normal speech in some way (for example, by pretending they are foreigners). EMULANDO is a corpus of real and feigned foreign-accented Spanish speech, including English, French, and Russian accents. These data may yield interesting studies and applications in the field of forensic phonetics. Some cues are discussed that can unveil such cases of disguise: the standard deviation of vowel formants, the timbre and duration of filler vowels, the acoustic characteristics of some fricatives, or intonational patterns. The corpus can also be used as training material for automatic biometric systems.

Palabras clave: disguised speech; foreign accent; forensic phonetics; Spanish.

1. INTRODUCCIÓN

El corpus que aquí se presenta es uno de los resultados del proyecto EMULANDO (*Estudios multilingües del acento no nativo disimulado*), que a su vez se integra en un proyecto mayor, DIANA (*Disimulo intencionado del acento nativo en el habla*). Dichos proyectos estudian el fingimiento del acento extranjero, una de las formas más habituales con las que los delincuentes modifican su manera normal de hablar para ocultar su lugar de origen, o incluso su personalidad (Masthoff, 1996). En concreto, el corpus de EMULANDO está constituido por muestras de habla en español con acento inglés, francés y ruso, algunas de las cuales son ejemplos producidos por verdaderos hablantes nativos de esas lenguas, mientras que otras han sido emitidas por hispanohablantes que fingían tener esos acentos.

Cuando un sujeto tiene intención de delinquir y sabe que su voz será examinada, en más de la mitad de los casos trata de alterarla por algún medio (Endres, Bambach y Flösser, 1971; Masthoff, 1996), lo que dificulta notablemente las tareas de discriminación o las de identificación de voces, tanto por parte de oyentes expertos como profanos. La estrategia de disimulo más frecuente es la modificación del registro vocal, de tal modo que se sustituye la fonación modal por el *falsetto*, la voz pulsada (*creak*) o el susurro. Este aspecto ha sido abordado previamente en otro proyecto, denominado CIVIL (Cualidad individual de la voz e identificación de locutor), descrito, por ejemplo, en San Segundo, Alves y Fernández Trinidad (2013). Por su parte, la imitación de un acento extranjero o dialectal es otro de los métodos más habitualmente utilizados por los malhechores, puesto que no siempre resulta fácil identificar cuándo un acento es fingido (Neuhauser, 2008). Otras estrategias también frecuentes incluyen la modificación de rasgos prosódicos como la f_0, los cambios en la velocidad de elocución, la imitación de rasgos patológicos o la alteración de las resonancias mediante el pinzamiento de la nariz o bien cubriendo la boca o el micro del teléfono (Masthoff, 1996). Adicionalmente, existen

procedimientos electrónicos de alteración de la voz, aunque estos suelen restringirse a casos de grandes organizaciones criminales, que cuentan con medios más sofisticados. Lo más común, y sobre todo en el caso de amenazas telefónicas, secuestros o extorsiones, es el disimulo no electrónico.

Los métodos de disimulo basados en la imitación de acentos extranjeros se han estudiado desde diversos puntos de vista: la capacidad de los hablantes de imitar acentos (Neuhauser, 2008), la de los oyentes de reconocerlos (Munro, Derwing y Burgess, 2010), la de determinar si se trata de un acento genuino o fingido (Neuhauser y Simpson, 2007), y la de identificar al locutor a pesar de su intención de camuflarse tras un acento impostado (Markham, 2007; Neuhauser, 2008; Tate, 1979). El disimulo del acento implica modificar determinados aspectos del sistema fonológico de la lengua en la que se habla, para hacerlos más similares a los de la lengua cuyo acento se está imitando. En este sentido, aunque algunas conclusiones generales de los estudios previos son extrapolables, resulta necesario desarrollar estudios particulares para cada combinación de lenguas.

El proyecto EMULANDO se centra en el estudio del español hablado con acento inglés, francés o ruso con el objeto de sistematizar las diferencias de pronunciación (segmentales y suprasegmentales) que que existen entre los casos de acento extranjero auténtico y aquellos que son fingidos. Es de esperar que existan incoherencias que delaten los casos de imitación. Dicho de otro modo, junto a algunos rasgos del acento que se adopta, pueden persistir otros propios de la lengua materna. Así, es plausible que ciertos atributos de la impostación sean difíciles de mantener durante mucho tiempo y se produzcan alternancias entre casos mejor y peor imitados. También cabe pensar que algunas propiedades del habla sean más difíciles de controlar o de modificar en función de la voluntad del hablante, quien puede que ni siquiera sepa acerca de la existencia o de la importancia de dichas propiedades.

Son precisamente estas características, las que escapan al control consciente de los sujetos, las que pueden resultar más valiosas en un contexto forense, cuando un perito debe determinar con qué probabilidad dos muestras de habla pertenecen (o no) a un mismo individuo. Por ejemplo, se ha descrito que los hablantes de alemán que imitan un acento francés mantienen determinadas características propias del alemán, como la inserción de oclusivas glotales ante palabras que empiezan por vocal, o la producción de clics velares en la distensión de las consonantes oclusivas (Simpson, 2007; Simpson y Neuhauser, 2009, 2010).

El corpus EMULANDO es, asimismo, un recurso valioso para profundizar en el conocimiento de los estereotipos que tienen los hablantes acerca de la manera de pronunciar que caracteriza determinados acentos extranjeros. Por otro lado, puede servir de base para estudiar la percepción del grado de verosimilitud de un acento y también para medir la discriminación de locutores cuando se comparan muestras en español con o sin impostación.

2. DISEÑO DEL CORPUS

2.1. Participantes

El corpus está formado por muestras de habla de un total de 60 locutores: 20 de ellos hablan en español con acento inglés, 20 con acento francés y 20 con acento ruso.

De los 20 locutores de cada acento, 8 son realmente hablantes nativos de inglés, francés o ruso (según corresponda), divididos a su vez en dos grupos, de 4 individuos cada uno, en función de su nivel de conocimientos de español como lengua extranjera. El grupo de nivel más bajo incluye sujetos con alguno de los primeros niveles (A1, A2 o B1) del Marco Común Europeo de Referencia para las lenguas —abreviado en adelante como MCER— (Council of Europe, 2001). En el grupo de nivel más alto se integran hablantes con alguno de los tres niveles restantes del MCER, a saber, B2, C1 o C2.

Por cada acento extranjero (inglés, francés o ruso), participaron además 12 sujetos nativos de español que fingieron hablar con ese acento: 4 de ellos no tenían ningún conocimiento previo de la lengua cuyo acento debían imitar, 4 tenían un nivel bajo (A1-B1 del MCER) y 4 tenían un nivel alto (B2-C2 del MCER).

2.2. Tareas y materiales

Cada sujeto participó en el experimento en dos sesiones diferentes, distanciadas entre sí al menos dos semanas, siguiendo la recomendación de recabar muestras separadas en el tiempo (Gil Fernández, Alves, y Hierro, 2012; Gil Fernández, Fernández Trinidad, Infante Ríos, y Lahoz-Bengoechea, 2017).

En la primera sesión, se pidió a los informantes que leyeran en voz alta dos textos en español fonéticamente equilibrados. El primero de ellos es el que proponen Bruyninckx, Harmegnies, Llisterri y Poch-Olivé (1994), mientras que el otro se corresponde con el texto fijo del corpus AHUMADA (Ortega-García, González-Rodríguez y Marrero-Aguiar, 2000). Ambos aparecen recogidos en San Segundo *et al*. (2013). Para obtener una muestra de control de cómo hablarían de manera habitual los hispanohablantes nativos, se les pidió (tan solo en el caso de esta tarea), que, antes de leer el texto fingiendo el acento extranjero que tuvieran asignado, lo hicieran también como lo harían normalmente en español. Los que no tenían el español como lengua materna simplemente recibieron la instrucción de leer los textos una vez. En este caso, lógicamente, el acento extranjero afloró en mayor o en menor medida según el individuo y su nivel de español. En el resto de tareas, todos los participantes hablaban con acento extranjero (bien real o bien imitado).

A continuación, los sujetos participaron en dos juegos de rol de unos cinco minutos de duración cada uno, lo que permitió obtener registros de una conversación semiespontánea. En el primer juego de

Lahoz-Bengoechea, J. M., Gil Fernández, J., & García García de León, C. L. (2019). EMULANDO: Corpus de habla con acento no nativo auténtico y disimulado. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 97–101). Málaga: Universidad de Málaga.

rol, el informante debía asumir el papel de un estudiante extranjero que quiere venir a España para cursar un grado universitario, y llama a la embajada para preguntar qué documentos necesita aportar o qué trámites debe seguir. El entrevistador, por su parte, tenía un guion con ciertas preguntas que debía hacerle al informante (Apéndice 1), y también disponía de varias hojas informativas con distintos tipos de datos para poder responder a las preguntas del sujeto.

Para la segunda conversación semiespontánea, el informante debía simular una llamada de extorsión anónima, explicando al investigador que lo había visto con su amante, y pidiéndole dinero en metálico para no publicar las fotos. Como en el caso anterior, tanto el participante como el entrevistador disponían de un pequeño guion con algunas pautas para orientar la conversación (Apéndice 2). De este modo, se consigue que algunos elementos del diálogo se repitan en todos los informantes, mientras que otros son improvisados.

En la segunda sesión, se repitieron las tareas correspondientes a los juegos de rol, para poder estudiar la estabilidad en el tiempo de los distintos rasgos fonéticos. Además, para obtener datos sobre contornos entonativos de distintos tipos, se pidió a los informantes que respondieran a un cuestionario guiado, similar al que presentan Prieto y Roseano (2010) en su apéndice. La tarea consiste en completar una serie de situaciones comunicativas que ponen al locutor en antecedentes para que así pueda emitir un enunciado lo más natural posible, con una interpretación pragmática adecuada a la situación propuesta. De esta manera se consigue elicitar distintas modalidades oracionales, estructuras informativas y valores de evidencialidad. Al igual que siempre, los hablantes no nativos de español impregnaron sus producciones con su acento característico; los nativos fingieron el acento que correspondiera en cada caso.

3. RESULTADOS

Los estudios realizados hasta el momento sobre el corpus EMULANDO arrojan resultados interesantes, que se resumirán a continuación.

En el caso de la imitación del acento inglés, se han encontrado varias divergencias con respecto al español hablado con acento inglés genuino. En el ámbito segmental, se observan diferencias significativas en la desviación estándar de los formantes vocálicos, que es mayor en el caso de la imitación (Gibson, Blecua Falgueras, y Cicres, 2017).

También hay diferencias en los valores medios de algunos de los parámetros acústicos que suelen utilizarse para describir las fricativas, y esto es así tanto en el caso de la /s/ como de la /θ/. El acento inglés disimulado, en estos casos, se puede distinguir del real a partir de la duración acústica de los segmentos, del número de cruces por cero en cada unidad de tiempo, del centro de gravedad y del coeficiente de asimetría, además del coeficiente de apuntamiento de la /θ/ y de la desviación estándar del espectro de la /s/ (Cicres y Fernández Trinidad, 2017).

Por otro lado, los patrones entonativos utilizados para las enumeraciones (tanto las que constan de una cantidad fija de elementos como las variables) son diferentes entre sí, y diferentes en inglés L1 que en español L1. Sin embargo, los imitadores del acento inglés mantienen una entonación mucho más parecida a la del español (Estebas Vilaplana, 2017).

En cuanto a la habilidad de los oyentes de distinguir entre un verdadero acento extranjero y uno disimulado, el estudio de Gibson *et al*. (2017) muestra que los sujetos que tienen el inglés como lengua materna son mejores que los hispanohablantes nativos a la hora de discriminar si el acento inglés es verdadero o imitado, al menos cuando deben basarse en las diferencias de timbre vocálico.

Por lo que respecta al ruso, Gil Fernández, Lahoz-Bengoechea, y Villa Villa (2017) estudian la vocal de relleno (que aparece habitualmente en hesitaciones, entre otras posibilidades). Con este objetivo, parten del corpus CIVIL (San Segundo *et al*., 2013) para el español L1 y de un corpus *ad hoc* para el ruso L1. En cada lengua, comparan estas vocales de relleno con las vocales léxicas acústicamente más próximas en cada caso (la /e/ en el español, y la /e/ y la /a/ en ruso). Una característica muy notoria es la mayor duración de la vocal de relleno, pero también existen diferencias significativas en cuanto al timbre. En español, los rellenos presentan valores de F2 y de F3 más elevados que los de la /e/ léxica. En ruso, los rellenos muestran, en cambio, valores de F2 y de F3 más bajos que la /e/ y también que la /a/, mientras que, en cuanto a F1, crean un grupo intermedio entre /e/ y /a/, significativamente distinto de ambos. También hay diferencias en la f_0 inherente de estas tres vocales rusas, siendo la vocal de relleno la que presenta una menor f_0.

A continuación, estos autores presentan en otro estudio una comparación del acento español nativo, el español hablado por rusos con un nivel C de español, y el español con acento ruso fingido por hablantes nativos de español con un nivel C de ruso (Lahoz-Bengoechea, Villa Villa, y Gil Fernández, 2017). El resultado es que los valores formánticos (F1 y F2) de la vocal de relleno en el acento disimulado no difieren significativamente de aquellos del verdadero acento ruso, mientras que ambos se apartan del acento español normal. No obstante, se observa una gran variación de un imitador a otro. En cambio, la duración de estas vocales, que no presenta diferencias entre el español nativo y el español hablado por rusos nativos (con nivel C), sí es significativamente menor en el caso del disimulo que en los otros dos grupos. Los autores proponen, como interpretación, que los hablantes son conscientes de que están imitando un acento y, por tanto, no se trata de rellenos realmente espontáneos. Por el contrario, llevan a cabo producciones muy controladas para acertar con el objetivo articulatorio (diferente timbre) y que suenen lo más parecidas al ruso como sea posible. Eso podría justificar la menor duración de tales realizaciones.

Lahoz-Bengoechea, J. M., Gil Fernández, J., & García García de León, C. L. (2019). EMULANDO: Corpus de habla con acento no nativo auténtico y disimulado. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 97–101). Málaga: Universidad de Málaga.

4. DISCUSIÓN Y CONCLUSIONES

Los análisis que se han llevado a cabo por el momento a partir del corpus EMULANDO muestran que existen abundantes y variadas diferencias fonéticas entre un verdadero acento extranjero y uno disimulado. Estos datos encierran una gran utilidad en el ámbito forense para ayudar a distinguir mejor entre ambos casos.

La relevancia de este corpus no se limita a la investigación básica (si bien esta es muy importante), sino que además brinda oportunidades de mejorar los sistemas biométricos automáticos. Al fin y al cabo, como reconocen Künzel, González-Rodríguez, y Ortega-García (2004), estos sistemas deben entrenarse a partir de grandes bases de datos, no solo de habla normal, sino con el mismo tipo de disimulo que se prentende desenmascarar. En este mismo sentido se manifiestan también otros autores (Perrot, Preteux, Vasseur, y Chollet, 2007; Zhang y Tan, 2008).

Además, el corpus permite ahondar en el papel tan importante que juegan los aspectos de detalle fonético, algo que cada vez se tiene más en cuenta en el análisis fonológico de las lenguas (véase, por ejemplo, Lahoz-Bengoechea, 2015).

En definitiva, las posibilidades de estudio sobre los distintos acentos que recoge el corpus y sobre los diferentes fenómenos fonéticos representados son todavía muy extensas, de modo que este recurso da cabida a muchos trabajos futuros.

5. REFERENCIAS

Bruyninckx, M., Harmegnies, B., Llisterri, J., y Poch-Olivé, D. (1994). Language-induced voice quality variability in bilinguals. *Journal of Phonetics*, *22*, 19-31.

Cicres, J., y Fernández Trinidad, M. (2017). Análisis de los sonidos fricativos en un corpus de acento no nativo disimulado. En V. Marrero Aguiar y E. Estebas Vilaplana (Eds.), *Tendencias actuales en fonética experimental: Cruce de disciplinas en el centenario del Manual de Pronunciación Española (Tomás Navarro Tomás)* (pp. 308-312). Madrid: UNED.

Council of Europe. (2001). *Common European framework for languages: Learning, teaching, assessment*. Strasbourg: Language Policy Division.

Endres, W., Bambach, W., y Flösser, G. (1971). Voice spectrograms as a function of age, voice disguise, and voice imitation. *Journal of the Acoustical Society of America*, *49*(6B), 1842-1848.

Estebas Vilaplana, E. (2017). Análisis de los rasgos prosódicos en el acento imitado: el caso de las enumeraciones. En V. Marrero Aguiar y E. Estebas Vilaplana (Eds.), *Tendencias actuales en fonética experimental: Cruce de disciplinas en el centenario del Manual de Pronunciación Española (Tomás Navarro Tomás)* (pp. 116-120). Madrid: UNED.

Gibson, M., Blecua Falgueras, B., y Cicres, J. (2017). Are American English better at distinguishing fake English accents in Spanish than native Spanish speakers? Presentado en Phonetik und Phonologie im deutschsprachigen Raum.

Gil Fernández, J., Alves, H., y Hierro, J. A. (2012). Proposition raisonnée de protocole de capture de voix connues à des fins judiciaires. *Revue Internationale de Criminologie et de Police Technique et Scientifique*, *65*(3), 319-344.

Gil Fernández, J., Fernández Trinidad, M., Infante Ríos, P., y Lahoz-Bengoechea, J. M. (2017). Obtaining speech samples for research and expertise in forensic phonetics. En L. Mariottini y F. Orletti (Eds.), *Forensic communication in theory and practice* (pp. 27-46). Cambridge: Cambridge Scholars Publishing.

Gil Fernández, J., Lahoz-Bengoechea, J. M., y Villa Villa, J. (2017). La vocal de relleno en español y en ruso: Caracterización acústica e implicaciones teóricas. *Estudios Filológicos*, *60*, 69-94.

Künzel, H. J., González-Rodríguez, J., y Ortega-García, J. (2004). Effect of voice disguise on the performance of a forensic automatic speaker recognition system. Presentado en Odissey 2004. The Speaker and Language Recognition Workshop.

Lahoz-Bengoechea, J. M. (2015). *Fonética y fonología de los fenómenos de refuerzo consonántico en el seno de unidades léxicas en español* (PhD dissertation). Universidad Complutense de Madrid.

Lahoz-Bengoechea, J. M., Villa Villa, J., y Gil Fernández, J. (2017). Fillers in disguised accented speech. Presentado en 13th Biennial Conference of the International Association of Forensic Linguists.

Markham, D. (2007). Listeners and disguised voices: the imitation and perception of dialectal accent. *The International Journal of Speech, Language and the Law*, *6*(2), 290-299.

Masthoff, H. (1996). A report on a voice disguise experiment. *Forensic Linguistics*, *3*(1), 160-167.

Munro, M. J., Derwing, T. M., y Burgess, C. S. (2010). Detection of nonnative speaker status from content-masked speech. *Speech Communication*, *52*(7-8), 626-637.

Neuhauser, S. (2008). Voice disguise using a foreign accent: Phonetic and linguistic variation. *The International Journal of Speech, Language and the Law*, *15*, 131-159.

Neuhauser, S., y Simpson, A. P. (2007). Imitated or authentic? Listeners' judgements of foreign accents. En *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 1805-1808).

Ortega-García, J., González-Rodríguez, J., y Marrero Aguiar, V. (2000). AHUMADA: A large speech corpus in Spanish for speaker characterization and identification. *Speech Communication*, *31*(2-3), 255-264.

Perrot, P., Preteux, C., Vasseur, S., y Chollet, G. (2007). Detection and recognition of voice disguise. En *Proceedings of the 16th Annual Conference of the International Association for Forensic Phonetics and Acoustics*.

Prieto, P., y Roseano, P. (Eds.). (2010). *Transcription of intonation of the Spanish language*. München:

Lincom Europa. Recuperado de http://prosodia.upf.edu/home/arxiu/publicacions/prieto/transcription_intonation_spanish.php

San Segundo, E., Alves, H., y Fernández Trinidad, M. (2013). CIVIL corpus: Voice quality for speaker forensic comparison. *Procedia: Social and Behavioral Sciences*, *95*, 587-593.

Simpson, A. P. (2007). Phonetic details of nonpulmonic stop release in German: inter-and intraindividual variation. En *Proceedings of the 16th Annual Conference of the International Association for Forensic Phonetics and Acoustics*.

Simpson, A. P., y Neuhauser, S. (2009). Enduring nature of epiphenomenal non-pulmonic sound production under disguise: A preliminary study. En *Proceedings of the Conference of the International Association for Forensic Phonetics and Acoustics*.

Simpson, A. P., y Neuhauser, S. (2010). The persistence of ephiphenomenal sound production in foreign accent disguise. En *Proceedings of the Conference of the International Association for Forensic Phonetics and Acoustics*.

Tate, D. A. (1979). Preliminary data on dialect in speech disguise. En *Current issues in the phonetic sciences: Proceedings of the IPS-77 Congress* (pp. 847-850).

Zhang, C., y Tan, T. (2008). Voice disguise and automatic speaker recognition. *Forensic Science International*, *175*(2-3), 118-122.

APÉNDICE 1

Para la conversación que simula una llamada a la embajada, los participantes recibieron las instrucciones siguientes.

Participante 1.—Eres un estudiante de 20 años que está planeando ir a vivir a España durante 4 años, para estudiar un Grado en Traducción e Interpretación en Granada, y has oído que para un periodo largo de estancia debes solicitar una serie de papeles. Te han hablado del NIE pero no tienes claro si debes pedirlo o no. Si es que no, quieres saber qué tipo de permiso tienes que pedir. Aún estás en tu país, y te gustaría hacer los trámites antes de ir a España. Llamas al servicio de atención de la embajada de España en tu país para obtener las siguientes informaciones:

- Saber si tienes que pedir el NIE.
- Si no, saber qué tipo de permiso debes solicitar.
- Saber si existe la posibilidad de hacer los trámites desde el país de origen.
- Qué documentación es necesaria.
- Dónde deben conseguirse los impresos.
- Cómo y dónde debe entregarse la documentación.
- Cuánto tiempo tarda.

Participante 2.—Eres un empleado del servicio de atención de la embajada de España en [Reino Unido / Francia / Rusia]. Recibes la llamada de una persona del país, pero que habla español, solicitando información, y tú has de dársela si la tienes y, si no, le has de decir amablemente que no dispones de esa información. [Siguen nueve páginas con datos extraídos de la web http://extranjeros.empleo.gob.es, así como de la web http://www.seap.minhap.gob.es].

APÉNDICE 2

Para la conversación que simula una llamada de extorsión, los participantes recibieron las instrucciones siguientes.

Participante 1.—Eres un delincuente que tienes en tu poder unas fotos comprometidas de un empresario (Miguel Fidalgo) con su amante. Llamas por teléfono a esta persona para comunicarle que tienes esas fotos (él aún no lo sabe) y para conseguir que te dé un dinero a cambio de no hacerlas públicas ni enseñárselas a su mujer.

Le puedes dar las siguientes informaciones:

- Las fotos fueron tomadas cuando el empresario y la amante estaban en un coche en la puerta de la casa de la amante, a altas horas de la madrugada.
- Son cinco fotos realizadas con un objetivo que permite ver claramente las escenas.
- El lugar exacto fue la calle de Ruiz Zapata, 6.
- La fecha y la hora de realización de las fotos: martes 9 de diciembre de 2015.
- La fecha de la llamada: 15 de diciembre de 2015.

Y le pides lo siguiente:

- 20 000 euros en metálico.
- En un sobre cerrado.
- Depositados el día 16 de diciembre a las 9 de la mañana.
- Depositados en una maleta en la consigna de la estación de Atocha.
- La llave de la consigna tiene que dejarla en un sobre pequeño a las 10 de la mañana del mismo día 16 en la papelera situada a la izquierda de la puerta principal de la estación.
- Si avisa a la policía, una copia de las fotos será entregada de inmediato a su mujer.

Participante 2.—Eres un empresario de una mediana empresa, Miguel Fidalgo, casado desde hace 15 años, con dos niños pequeños. Desde hace un año matienes otra relación. Para ti, 20 000 euros es muchísimo dinero y tu empresa, además, no está pasando un buen momento. En la conversación con el extorsionador, intentas obtener información sobre esa persona, por si es alguien conocido, y quieres insistir en tres puntos:

- Ver las fotos antes de pagar nada.
- Entender bien el proceso de entrega del dinero.
- Hacer hablar al extorsionador, por si puedes reconocer su voz.

Subsidia: Tools and Resources for Speech Sciences

16

Detecting neuromotor disease in speech articulation

Pedro Gómez[1], Daniel Palacios[1], Andrés Gómez[1], Cristina Carmona[2], Ana R. Londral[3], Victoria Rodellar[1], Víctor Nieto[1], Miguel A. Ferrer[2], Agustín Álvarez[1]

[1] Universidad Politécnica de Madrid
[2] Universidad de Las Palmas de Gran Canaria
[3] Universidade de Lisboa
e-mail: pedro@fi.upm.es

Citation / Cómo citar este artículo: Gómez, P., *et al.* (2019). Detecting neuromotor disease in speech articulation. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 103–108). Málaga: Universidad de Málaga.

ABSTRACT: Speech articulation may be an important resource to study neuromotor activity in relation with neurological diseases, such as Parkinson's, Alzheimer's or Amyotrophic Lateral Sclerosis. Through the present work a biomechanical model for the joint structure of the jaw and tongue is introduced. This model allows putting into relation formant trajectories with jaw-tongue kinematics in terms of dorso-ventral and rostro-caudal movements of the structure. The reconstruction of the absolute velocity of the system from inverse filtering estimates of the first two formants allows the introduction of the velocity distribution histograms as descriptors of articulation behavior. A study case using recordings separated in time from a patient suffering from Amyotrophic Lateral Sclerosis, compared to distributions from a normative control is presented. It may be seen that patient's distributions show larger contents for low ranges of velocity and smaller contents for high velocity ranges compared to the control distribution. Global comparisons of velocity distribution profiles between the control and patient's records using Information Theory measurements may be used to monitor illness conditions and progress.

Keywords: speech processing; speech articulation; speech kinematics; neuromotor dysarthria; amyotrophic lateral sclerosis.

RESUMEN: La articulación del habla puede ser un recurso importante para estudiar la actividad neuromotora en relación con enfermedades neurológicas, como el Parkinson, el Alzheimer o la Esclerosis Lateral Amiotrófica. Este trabajo presenta un modelo biomecánico de la estructura que forman conjuntamente la mandíbula y la lengua. Este modelo permite relacionar las trayectorias de los formantes con aspectos cinemáticos de la mandíbula y la lengua en términos de movimientos dorso-ventrales y rostro-caudales de dicha estructura. La reconstrucción de la velocidad absoluta del sistema a partir de un filtrado inverso de las estimaciones de los dos primeros formantes permite introducir histogramas de la distribución de la velocidad como descriptores del comportamiento articulatorio. Se presenta un estudio de caso a partir de grabaciones separadas en el tiempo de un paciente que sufre Esclerosis Lateral Amiotrófica, comparadas con las distribuciones de un sujeto control normativo. Se observa que las distribuciones del paciente muestran una mayor actividad en los rangos de velocidad bajos y una menor actividad en los rangos de velocidad altos, en comparación con la distribución control. Se propone hacer comparaciones globales de los perfiles de distribución de la velocidad entre las grabaciones control y las del paciente utilizando medidas basadas en la Teoría de la Información, y que este uso puede ayudar a monitorizar las condiciones y la evolución de la enfermedad.

Palabras clave: procesamiento del habla; articulación del habla; cinemática del habla; disartria neuromotora; esclerosis lateral amiotrófica.

1. INTRODUCTION

Speech articulation is a result of oro-naso-pharyngeal tract modifications produced by the movement of four main groups of muscles (Jürgens, 2002): jaw muscles, lingual extrinsic and intrinsic, oro-facial, and velo-pharyngeal.

As neuromotor units can produce only muscle contractions, when activated, each group of muscles is fitted to an agonist-antagonist pair (Kandel, Schwartz, & Jessell, 2000) to produce smooth movements in both directions. When the agonist group contracts and pulls in one direction, the antagonist group must relax and yield, and vice-versa. This delicate action-reaction

stimulus is regulated by an excitatory-inhibitory network of neurons.

The steady movement of muscles is a result of these agonist-antagonist actions, and has been very well described and modelled for the hand-writing function (Plamondon, Djioua & Mathieu, 2013). It has been shown that this model can also be applied to the study of speech articulation (Gómez-Vilda, Londral, Rodellar-Biarge, Ferrández-Vicente, & de Carvalho, 2015).

The intention of the present study is to show that the implications of the agonist-antagonist neuromotor function can be taken one further step ahead when dealing with speech fluency features (Singh, Bucks, & Cuerden, 2001), such as verbal rate, mean duration of pauses, standardized phonation time, and standardized pause rate, or articulation features, such as: speech rate (syllables / total duration time) or articulation rate (syllables / total locution time).

The fluency of speech production is subject to a temporal dynamic flow, which may be observed at different time scales. In the first level, the presence or absence of speech can be characterized by the presence of phonic groups, which are segments of time where a signal activity over the background noise level of the channel can be observed over a significant value. The intervals between phonic groups are considered as pauses. In the second level, phonic groups can be divided into phonated and non-phonated intervals, depending if there is vocal fold activity involved or not. In the third level, phonated intervals are divided into segments where formant activity is stable under a modulation frequency limit (in Hz/s) or not.

Having into account the ubiquity of speech recording and transmission on IP platforms, the present work is intended to search for a statistical description of the dynamic phenomena present in the three levels to describe articulation dynamics for its application in neurologic disease characterization from speech.

The paper is organized as follows. The basic model explaining articulation neuro-motor foundations is presented in section 2. Section 3 is devoted to introduce a study case used as an example for the characterization of pathologic speech (a case of amyotrophic lateral sclerosis), and presents the basic algorithmic procedures and the application supporting them. Section 4 presents the results of analysis and their statistical analysis. And section 5 summarizes the main conclusions of the study.

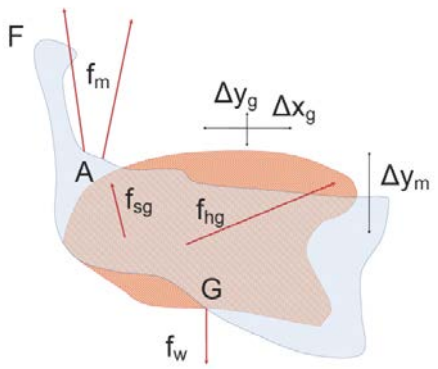
2. BACKGROUND MODEL

The process of articulation is a complex biomechanical task which implies the timely organized activation of different muscles under the control of the neuromotor cortex. The concurrent activation of respiration and phonation muscles must be accompanied by the joint action of at least the jaw, lingual (extrinsic and intrinsic) and facial, as well as the gloso-velo-pharyngeal muscles. Many neurologic pathologies result in the inadequate functioning of some of these biomechanical systems, and altered phono-acoustic correlates appear as a consequence, which may serve as indicators of the pathological condition of the speaker, and help in the characterization of the pathological process under a functional point of view.

The complexity of the whole articulation system requires a divide-and-conquer approach for its study. In the present work the study will be focused in the jaw-tongue biomechanical system (Gerard, Perrier & Payan, 2006), considered as a first-order mass-spring structure as described in Figure 1.

The jaw-tongue subsystem is presented as a third-class lever (Röhrle & Pullan, 2007), which is fixed to the cranial structure at point F (fulcrum), experiencing the force of gravity at G, and supported mainly by the action of the masseter pair at A. In this model several other acting muscles and substructures have been omitted for the sake of computational simplification. The model assumes that all forces acting on the jaw-tongue massive structure (other than gravity, which is referred to the centre of gravity) can be referred to a certain joint-mandible dynamic reference point (JMDRP), with implicit coordinates x and y in the sagittal plane, such that the axis x refers to movement in the dorsal-ventral direction (DV), and the axis y refers to movement in the caudal-rostral direction (CR). Lateral movements orthogonal to the sagittal plane are assumed small enough not to be considered (thus giving a system with two degrees of freedom). The kinematic variables relevant to the study are the displacements Δx and Δy relative to the JMDRP, which will be contributed mainly by the CR displacement of the jaw (Δy_m) and the DV and CR displacements of the tongue (Δx_g Δy_g). Important additional assumptions to this model are that the tongue system is the main surface opposite to the palate ceiling, acting as a solidary hydrostatic bulk, and that the relative displacements between these surfaces configure the main articulation cavity (Gerard *et al.*, 2006). On its turn, for the dynamic part, the main forces acting on the JMDRP (besides gravity acting on the center of gravity, as said) are the masseter contraction (f_m), the styloglossus action (f_{sg}) and the hyoglossus action (f_{hg}, both superior and inferior muscle bundles).

Figure 1: Biomechanical model of the jaw-tongue subsystem. The jaw bone is represented in light grey; the tongue structure is represented in light orange.



Under these assumptions, the resonance model involving the first two formants f_1 and f_2 can be put into relation with the positions of the jaw-tongue bulk relative to the JMDRP (Sanguineti, Laboissiere & Payan, 1997) by the dynamic system in (1):

$$\begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} \quad (1)$$

where a_{ij} are the transformation weights explaining the position-formant association, and t is the time. This relationship is known to be one-to-many, i.e. the same pair of formants $\{f_1, f_2\}$ may be associated to more than a single articulation position. This inconvenience may be handled by modelling the joint probability of all the possible articulation positions associated to a given formant pair (Dromey, Jang & Hollis, 2013).

Under certain invertibility assumptions, which will not be given here for the sake of brevity, the system in (1) may be written in opposite terms to help expressing the algorithmic methodology implied in the process of deriving kinematic variables from acoustical ones, as in (2):

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix} \quad (2)$$

where w_{ij} are the weights of the inverse system. The first time derivative of this system allows associating formant derivatives in time with the JDMRP kinematics, as in (3):

$$\begin{bmatrix} v_x(t) \\ v_y(t) \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \begin{bmatrix} \frac{df_1(t)}{dt} \\ \frac{df_2(t)}{dt} \end{bmatrix} \quad (3)$$

where it has been assumed that the system is linear and time-invariant, and v_x and v_y are the DV and CR velocities of the JMDRP.

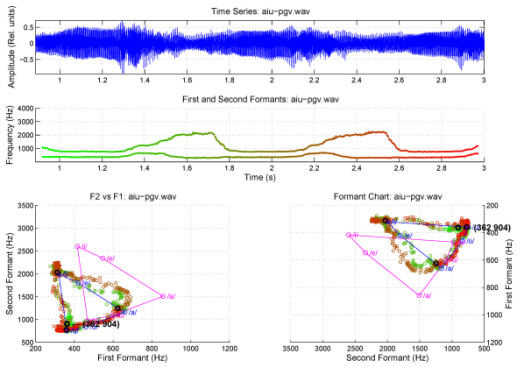
The values of the weights w_{ij} are related to the kinematics of the specific speaker, and can be considered as a biometrical mark of the person. It may be hypothesized that the DV velocity will be mostly related to changes in the second formant (back-front), and that the CR velocity will be related to the dynamics of the first formant (up-down).

An estimate of the absolute velocity of the JMDPR may be evaluated as in (4):

$$\left|v_{RP}(t)\right| = \sqrt{\left(w_{21}\frac{df_1(t)}{dt}\right)^2 + \left(w_{12}\frac{df_2(t)}{dt}\right)^2} \quad (4)$$

Therefore it will be hypothesized that w_{11} and w_{22} will be negligible compared to w_{12} and w_{21}. Reliable estimates for these scale factors may be obtained from diphthong articulations as for instance [aj] or [ja], involving changes in the positions of the JMDRP which are not affected by strong labialization. In a practical case, estimations of the scale coefficients may be obtained from sequences as the one shown in Figure 2, using the first two formant changes comprised between 1.46 s and 1.6 s, or 2.26 s and 2.4 s.

Figure 2: Formant structure of the monotonically repeated sequence /aiu/ uttered as [...ajijuwa...]. Upper template: time domain signal; middle template: first two formant patterns in time; lower template: two representations of formant positions over the vowel triangle.



The averaged estimations for both coefficients from the data in Figure 2 are respectively w_{12}=1.62.10^{-3} cm.s and w_{21}=1.47.10^{-3} cm.s.

3. MATERIALS AND METHODS

In what follows, results from a study case involving a patient of Amyotrophic Lateral Sclerosis (ALS) will be shown. Five recordings were taken from a 64-year old female patient from the Neurology Department of Hospital de Santa Maria, in Lisbon, Portugal, suffering from ALS, who had been diagnosed 1 year before. Recordings were taken at each neurological control, spaced 3 months (respectively: P1T0, P1T1, P1T2, P1T3 and P1T4). The recordings consisted in the utterance of a popular sentence from Fernando Pessoa: /*tudo vale a pena quando a alma não é pequena*/ (IPA: [tʊðʊɥɑlɐ pɐ̃næ kwændʊaaɫmɐ nɐ̃ɛ̃ pkɐ̃næ]) by the ALS patient during the evaluations. The recordings were taken at 44.1 kHz and 16 bits. The same recording was taken from a 36-year old healthy female control. The selection of ALS patients for this study was not coincidental. It is well known that these patients suffer from a continuous degradation of neuromotor functions which contribute to a progressive reduction of the vowel triangle (Yunusova, Weismer, Westbury, & Lindstrom, 2008). The main aim of the present study is to check the extent of the deterioration of articulation functions in the dynamic time domain as well. The basic methodological protocol consists in the following steps:

- Recordings are undersampled to 8 kHz.
- The vocal tract transfer function of the speech segment is evaluated by a 9-pole adaptive inverse LP filter (Deller, Proakis & Hansen, 1993) with a low-memory adaptive step to grasp fine time variations.
- The first two formants are estimated by evaluating the roots of the associated inverse polynomials. The formant estimation resolution used is 2 Hz, and an estimation is produced every 2 ms.

Gómez, P., *et al*. (2019). Detecting neuromotor disease in speech articulation. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 103–108). Málaga: Universidad de Málaga.

- The derivatives of the first two formants are used to estimate the absolute velocity of the JMDRP following Equation (4).
- The values of the absolute velocity are used to build a histogram as a function of the absolute velocity distribution.
- The histograms are used to estimate probability density functions by Kolmogorov-Smirnov approximations (Webb, 2003).
- Kullback-Leibler's Divergence (Webb, 2003) is estimated between each patient's recording distribution $p_{Pi}(v)$ vs that of the control subject $p_C(v)$ as by (5), where the absolute velocity of the JMDRP has been defined in a given interval, which for the present study is in the range R_v={0, 200 cm.s^{-1}}. The above described steps are programmed into a Graphical User Interface (BioMet®Ling: www.glottex.com), which is shown in Figure 3.

$$D_{KL}(p_{Pi}, p_C) = \int_{v \in R_v} p_C(v) \log \left| \frac{p_{Pi}(v)}{P_C(v)} \right| dv \quad (5)$$

4. RESULTS AND DISCUSSION

In what follows the absolute velocity profiles for the control subject (CF) and the first (P1T0) and last (P1T4) patient's utterances are given as a reference in Figure 4. These profiles correspond to low-frequency movements under 20 Hz.

It may be seen that the upper template (Figure 4a) shows a fast and well organized action pattern, where the strongest neuromotor spikes reach values between 20 and 40 cm.s^{-1}. The total utterance duration is around 2.7 s, with two pauses. The middle template (Figure 4b), corresponding to the first patient's evaluation shows a slower utterance lasting 5.3 s, where four pauses may be seen, and the strongest spikes are between 10 and 30 cm.s^{-1}. The lower template, corresponding to the last patient's evaluation (Figure 4c) needs almost 9 s to complete the same utterance, there are no pauses, and the largest spikes are between 8 and 20 cm.s^{-1}.

On its turn, Figure 5a shows the results of estimating the probability density functions of the absolute velocity profiles (low and high frequency) for CF and the five patient recordings, the comparison between CF and the first patient's one (Figure 5b), between CF and the last patient's one (Figure 5c) and between the patient's first and last ones (Figure 5d).

Figure 3: Graphical User Interface of BioMet®Ling

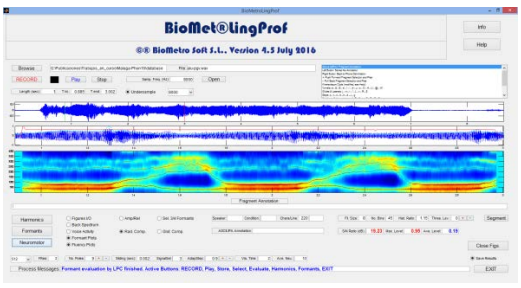


Figure 4a: Absolute velocity profile from the control subject.

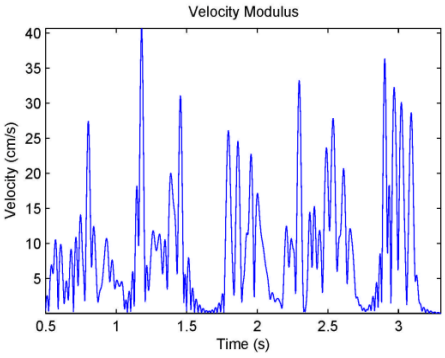


Figure 4b: Absolute velocity profile from the patient's first utterance.

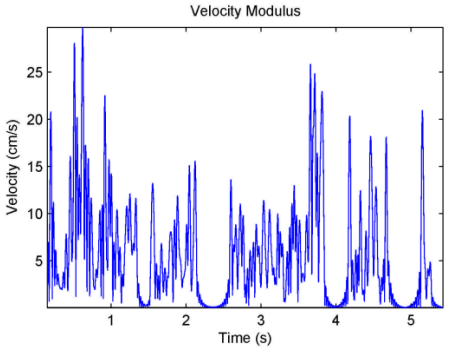
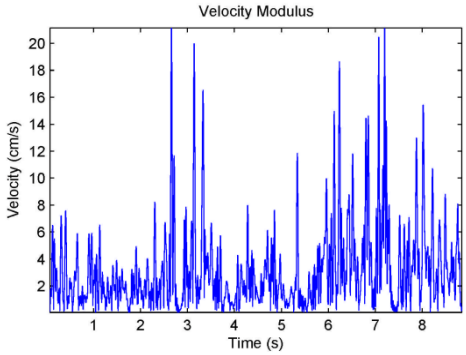


Figure 4c: Absolute velocity profile from the patient's last utterance.



It may be seen that the control subject probability density function shows velocity profiles over 80 cm.s^{-1}, which can be barely appreciated in the distributions from the ALS patient. The velocity range is larger than in the plots of Figure 4 because high and low frequency contents have been taken into account. As illness progresses, the distribution contents displace to lower velocity bins.

Figure 5b shows that the velocity distribution from the patient's first recording displays more contents for lower velocity bins than the one from the control subject, whereas for larger velocity bins the control subject shows more contents than the patient's one. This tendency is more evident when comparing the

Figure 5a: Absolute velocity probability functions for the control and patient's utterances.

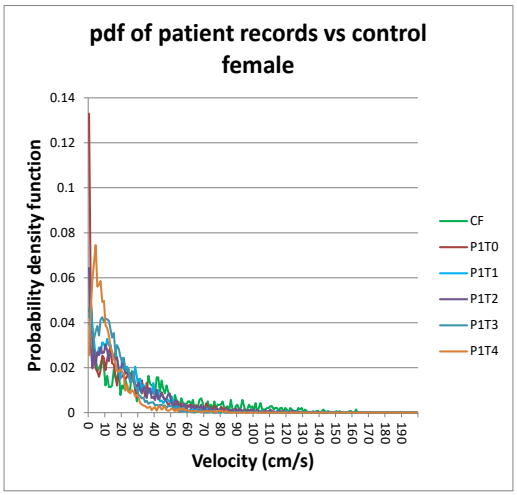
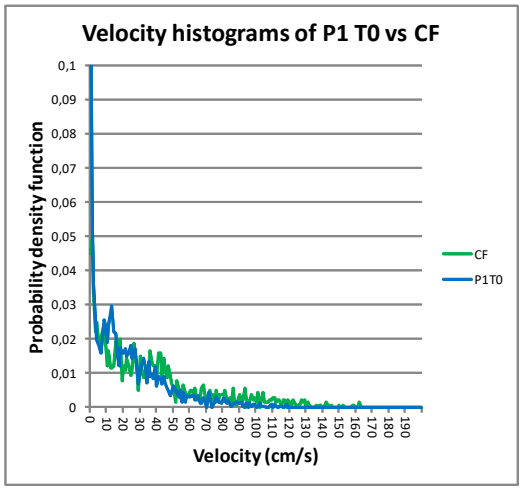


Figure 5b: Absolute velocity probability functions for the control and patient's first utterance.



patient's last distribution against the control subject (Figure 5c). Finally, in Figure 5d the patient's first and last distributions are compared, showing that the articulation dynamics of the patient has lost components in the high velocity range, which are transferred to the low velocity range. This is a clear indication of illness progression, which manifests itself as a sloth and less vivid articulation activity due to the degradation of jaw and tongue neuromotor units characteristic of ALS (known also as the motor neuron, or Lou Gherig disease).

Table 1 shows the results from estimating Kullback-Leibler's Divergence for the patient's five recordings vs the control subject.

Table 1: Kullback-Leibler's Divergence for the patient's five chronological recordings vs the control subject.

Recording	KLD
P1T0	0.42657477
P1T1	0.53332203
P1T2	0.51182122
P1T3	0.70738088
P1T4	0.93337742

Figure 5c: Absolute velocity probability functions for the control and patient's last utterance.

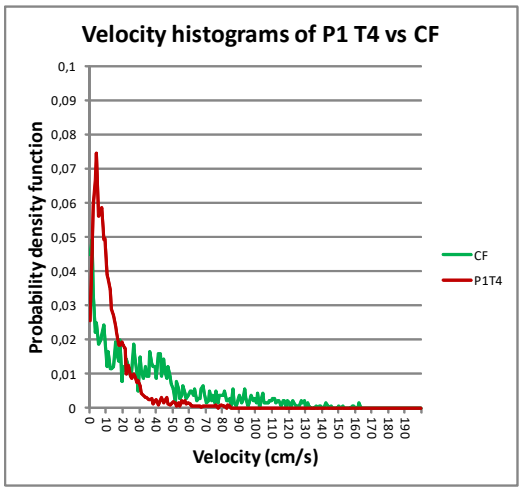
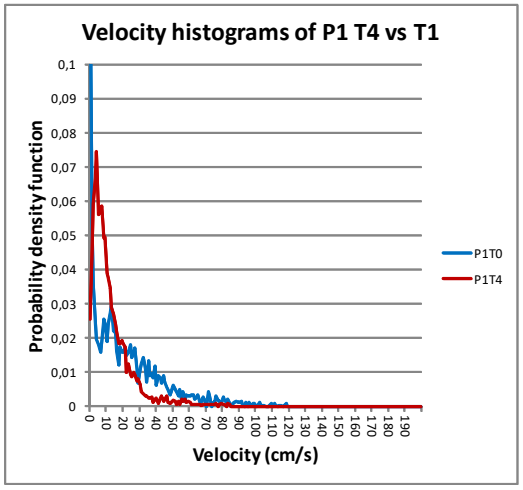


Figure 5d: Absolute velocity probability functions for the patient's first and last utterances.



It can be concluded again that the absolute divergence from the normative articulation profile shown by the control subject and that of the ALS patient increases as illness progresses except for recording P1T2, where a possible stagnation in the degradation of speech articulation may be circumstantially observed.

Generally speaking, it may be concluded that the articulation dynamic features have been captured by the absolute velocity probability distribution as derived from the velocity histogram in the range 0–200 cm.s^{-1}, which may become a comparison standard when information-theory derived statistics as Kullback-Leibler's Divergence are used. A very interesting property of the velocity histogram is that it may comprise different articulation dynamics in its structure, such as at the level of pauses and phonic groups (very low level velocity profiles), phonated and non-phonated intervals (low level velocity profiles), and formant modulation in syllabic segments (mid- to high level velocity profiles). The characterization of the velocity profiles accordingly to speech own dynamics could open its possible application to other neurodegenerative diseases which present fluency and

dynamic dysarthrias, such as Parkinson's or Alzheimer's as well.

5. CONCLUSIONS

Several conclusions can be derived from the present study, the following seem to be the most relevant ones among them:

- The articulation dynamics may be derived from the temporal evolution of the first two formant derivatives.
- A frequency resolution of 2 Hz and a time resolution of 2 ms give enough accuracy to characterize fast formant changes.
- The absolute velocity profile of the JMDRP is a rather semantic correlate, relating high and low motion with the normal behaviour of the main biomechanical system related to speech articulation.
- The application of the methodology to a case study of progressive speech deterioration produced by ALS shows the viability and applicability of the methodology.
- The use of statistical distance measurements derived from Information Theory may be a powerful means to produce objective estimates to track illness progress, opening new ways for distant patient monitoring by e-Health platforms.

These conclusions are to be confirmed on a larger number of cases, and the application of the methodology to other neurologic pathologies is foreseen in the next future.

6. REFERENCES

Deller Jr, J. R., Proakis, J. G., & Hansen, J. H. (1993). *Discrete time processing of speech signals*. Englewood Cliffs, NJ: Prentice Hall.

Dromey, C., Jang, G. O., & Hollis, K. (2013). Assessing correlations between lingual movements and formants. *Speech Communication*, *55*(2), 315–328.

Gerard, J. M., Perrier, P., & Payan, Y. (2006). 3D biomechanical tongue modeling to study speech production. In J. Harrington, & M. Tabain (Eds.), *Speech production: Models, phonetic processes, and techniques* (pp. 85–102). New York: Psychology Press.

Gómez-Vilda, P., Londral, A. R. M., Rodellar-Biarge, V., Ferrández-Vicente, J. M., & de Carvalho, M. (2015). Monitoring amyotrophic lateral sclerosis by biomechanical modeling of speech production. *Neurocomputing*, *151*, 130–138.

Jürgens, U. (2002). Neural pathways underlying vocal control. *Neuroscience & Biobehavioral Reviews*, *26*(2), 235–258.

Kandel, E. R., Schwartz, J. H. & Jessell, T. M. (2000). *Principles of neural science*. New York: McGraw-Hill.

Plamondon, R., Djioua, M., & Mathieu, P. A. (2013). Time-dependence between upper arm muscles activity during rapid movements: Observation of the proportional effects predicted by the kinematic theory. *Human Movement Science*, *32*(5), 1026–1039.

Röhrle, O. & Pullan, A. J. (2007). Three-dimensional finite element modelling of muscle forces during mastication. *Journal of Biomechanics*, *40*, 3363–3372.

Sanguineti, V., Laboissiere, R., & Payan, Y. (1997). A control model of human tongue movements in speech. *Biological Cybernetics*, *77*(1), 11–22.

Singh, S., Bucks, R. S., & Cuerden, J. M. (2001). Evaluation of an objective technique for analysing temporal variables in DAT spontaneous speech. *Aphasiology*, *15*(6), 571–583.

Webb, A. R. (2003). *Statistical Pattern Recognition*. Chichester, UK: Wiley.

Yunusova, Y., Weismer, G., Westbury, J. R., & Lindstrom, M. J. (2008). Articulatory movements during vowels in speakers with dysarthria and healthy controls. *Journal of Speech, Language, and Hearing Research*, *51*(3), 596–611.

Subsidia: Tools and Resources for Speech Sciences

17

Perceptual experiments in Praat: beyond the standards

Rubén Pérez Ramón

[1] Universidad del País Vasco
e-mail: rperez.ram@gmail.com

Citation / Cómo citar esta publicación: Perez-Ramon, R. (2019). Perceptual experiments in Praat: beyond the standards. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 109–115). Málaga: Universidad de Málaga.

ABSTRACT: Research in experimental phonetics requires, generally, the use of support tools that facilitate data collection as much as possible. Throughout history, these tools have been diverse in nature and have been supported by a number of platforms, spanning from pens to computers. Technology has reached the point of creating advanced tools that make the job of the researcher easier than what has ever been expected, but there is not, as of today, a platform to encompass a standard set of tools for evaluation in phonetics. In the present work, *Praatception* is introduced. *Praatception* is a platform which includes several tools for the research in perception. It works through the acoustic analysis tool *Praat*, so it is free and open-code.

Keywords: perception; experimentation; methodology; tool

RESUMEN: La investigación en fonética experimental requiere, en numerosas ocasiones, el uso de herramientas de apoyo a la investigación que faciliten en la medida de lo posible la recogida de datos. A lo largo de la historia, estas herramientas han adquirido diversas formas y se han sustentado en soportes de distinta índole. La evolución de la tecnología ha conseguido crear herramientas avanzadas que facilitan la labor del investigador, pero no existe aún una plataforma que aúne un conjunto de herramientas que estandaricen, tanto como se pueda, la metodología empleada en la evaluación fonética. En el presente trabajo se presenta la herramienta *Praatception*, un conjunto de aplicaciones diseñadas a facilitar la investigación en fonética perceptiva. *Praatception* funciona a través de la herramienta de análisis acústico *Praat*, por lo que es gratuita y de código abierto.

Palabras clave: percepción; experimentación; metodología; herramienta

1. INTRODUCTION

1.1. Perception and the research in phonetics

Traditionally, phonetics as a field of study has been considered from three different angles that, being inevitably related, focus in different aspects: articulatory phonetics, which studies the way the sounds of speech are produced; acoustic phonetics, in charge of the speech signal; and perceptual phonetics, the one referred to how speech is received by the listener. All three of them require very specific techniques for analysis and evaluation. Thanks to the advancements of technology and the contributions of the scientific community, these techniques are growing both in number and quality by the day.

There is an important difference between the nature of articulatory and acoustic phonetics and that of perceptual phonetics. The first two are *physical* manifestations of speech, this is, measurable events. They are limited, therefore, by the easiness of those events to be measured. There is an obvious evolution from the direct palatography technique (Abercrombie, 1957; Marchal, 1988) to the new ways of the 3D palatography (Legou, 2008), which allow the researcher to extract data with a precision quite unthinkable not such a long time ago.

The limitations of perceptual phonetics are different. Ideally, perception could also be classified as a physical event, being that the sounds are perceived by the ear and transmitted to the brain, where they are decoded and interpreted. The problem is that very little is known about what happens in the brain of a listener, and, even though improvements are steadily being made in that area (Grant, 2016), we are still too far from the point of completely understanding it. In order to know what happens in the ears and brains of the listeners there is no more option than asking them directly. This situation is double-edged: on one hand, the impossibility of acquiring and interpreting more sophisticated data makes the collection easier; on the other, a great amount of data is required to achieve consistency. It is crucial, therefore, that the process of data collection becomes as optimal as possible.

In the present work we introduce *Praatception*, a tool for research in perception whose aim is to ease the work of the researcher. *Praatception* is a plug-in for *Praat*, a free software widely used for speech analysis. The presented tool includes several pre-defined interfaces for perceptual experimenting, so the researcher can choose the paradigm that best suits its experimental requires.

1.2. Perceptual tasks

The way perception is measured has changed considerably through the years, both because of the technical advantages and the development of new methodologies. One aspect of perception that has been studied for a long time is intelligibility. As early as 1956, Egan presented a study in which message transmission in noise was studied (see Egan, 1956). Even though intelligibility was not addressed by that name, this study represents one early work in which this dimension was tested. In 1982, Nelson (see Nelson, 1982) defines *Intelligibility* as "the apprehension of the message in the sense intended by the speaker". Later, in Smith (1985), the concept of intelligibility is disambiguated in three different facets: intelligibility, comprehensibility and interpretability. Each of these terms is associated with a different meaning: utterance recognition, utterance meaning and meaning behind the utterance respectively. In other words, if intelligibility is the ability of a listener to recognize utterances, comprehensibility is the reported difficulty to understand those utterances, while interpretability would be the ability of the individual to fit that utterance into his or her knowledge of the world.

Generally, intelligibility is measured as the percentage of correct answers (i.e. the total amount of recognised utterances. Some authors, nevertheless, prefer to measure only what they call *key words* (Barefoot, 1993) or even the ability to paraphrase (Brodkey, 1972). In any case, intelligibility arises as an objective measure that can be scored, while comprehensibility is a subjective method of evaluation.

Liberman (1958) presents one early example of categorisation and discrimination tasks. They synthetically generate a 14-step continuum between the three voiced plosives and asked participants to label each of the stimuli as either [b], [d] or [g]. Their results show that, at certain points in the continuum, the perception shifts from one category to the other. This task, that they refer to as "sound labelling", is what will later be known as a categorisation task. McQueen (1996) defines these kind of tasks as those in which the participant hears a continuum of ambiguous sounds between two unambiguous endpoints. His or her instructions are to decide to which one of the two endpoints the presented stimuli belong.

Categorisation is related to discrimination. It is expected that the point of the continuum in which the perception shifts from one category to the other is also the place where two contiguous steps are more easily discriminated by the listener. When categorisation and discrimination performance are correlated, it is said that the continuum is perceived categorically (see MacMillan, Kaplan, & Creelman, 1977). A discrimination task, therefore, is usually presented alongside the categorisation task.

Discrimination can be measured through different paradigms. The simplest one is an AX experiment, in which two stimuli are presented simultaneously and the listeners are asked to judge whether they are identical or not. In Liberman (1958), on the other hand, an ABX task is presented, in which the participant is asked to decide if the X stimulus is more similar to A or B. Other types of discrimination tasks are reviewed in Gerrits & Schouten (2004) and their relation with categorical perception is analysed.

The term *categorisation* has usually been presented as a synonym of *identification* (see McQueen, 1996), but for the purpose of the present work, identification will be described as a mixture of intelligibility and categorisation. In Gerrits & Schouten (2004), an identification task is presented as one in which listeners are asked to choose from an open-set, i.e. a whole inventory. It is, therefore, a task in which the participant can freely choose among every possible answer, as in the intelligibility task. The advantage of an identification task compared to an intelligibility task is that some problems such as typos or non-expected answers are avoided by providing options to choose.

1.3. Tools for perceptual testing and advantages of *Praatception*

There are several software tools that allow the user to perform perceptual experimenting. One of the firsts to include interactive interfaces was *PsyScope* (Cohen, Macwhinney, Flatt, & Provost, 1993). It is a tool for designing and running psychology experiments with the advantage that it does not require the user to know about coding. It is limited, nevertheless, to Macintosh computers.

Other platforms, such as *E-prime* (Dickinson, 2011), offer more advanced support and a wide set of customizable tools for tasks beyond perception such as eye-tracking or EEG. As others, nevertheless, this tool is not free, which may be a disadvantage to the researcher.

Paradigm (López-Bascuas, 1999) is a computer program dedicated specifically to perceptual experimenting. It allows the user to choose different experimental paradigms among identification and discrimination tasks. When it was first launched, Paradigm was limited to Windows systems but, to our knowledge, it is currently discontinued.

Platforms such as *Matlab* or *Python* have also been used to generate custom interfaces for perceptual experimenting (see García Lecumberri, Barra Chicote, Pérez Ramón, Yamagishi, & Cooke, M., 2014, for an example of a Matlab-implemented experiment). They require the user to have advanced knowledge in coding in order to provide a successful interface; moreover, Matlab is not free of charge.

Pérez Ramón, R. (2019). Perceptual experiments in *Praat*: Beyond the standards. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 109–115). Málaga: Universidad de Málaga.

Folerpa (Aguete Cajiao, 2016) is a free web-based tool for perceptual experimentation. After a simple registration process, the user can create different test modules. It offers a high degree of customization, and several experimental paradigms can be arranged. The fact that the tool is online makes it available worldwide and supported by most of the computers. On the other hand, the availability of the tool is limited by the internet access of the researcher. Furthermore, the process of loading the sound files can become arduous for big experiments or when the quality of the available network is not powerful enough.

Praat offers a native ready-to-use interface for perceptual experimenting. The researcher can easily arrange categorisation, discrimination or identification tasks and yield likert scales for goodness-of-fit judgement. Sometimes, nevertheless, this interface falls short regarding questions such as format, personalization, etc. The very creators or *Praat* suggest in the manual that the user should take a different approach and use other platforms or the *Demo* window, a not so known feature included in their software.

This is, precisely, the basis of *Praatception*: the *Demo* window. Through this feature the user can define completely customizable windows that can be linked between them as required, generating a workflow. See Boersma (2016) for more information regarding the use of the *Demo* window.

The advantages of *Praatception* with respect to the aforementioned methods presented here are three: first, it is an open-code, free tool; Second, it works in every system that can support *Praat* which, to this day, are OSX, Windows and Unix; And, finally, it offers a high degree of customization. The researcher can generate a completely customized experiment and load it into a USB stick to set it up in whatever computer she or he needs to.

2. PRAATCEPTION: FIRST STEPS

As stated previously, *Praatception* is an extension designed to run within the *Praat* environment. It is installed as any other plug-in (see Boersma, 2016, for more detailed instructions regarding this topic). Once it is installed, the user will be able to define new experiments that will be saved in *Praat*'s standard interface so they can be run any number of times just by clicking a button.

When the experiment is loaded, the user is required to enter an ID that will serve as the unique identification name of the participant. It is recommended that the researcher assigns the IDs of the participants, as *Praatception* will not control if they are repeated or not. The ID field accepts both numeric and alphabetic characters, so the researcher can specify features (e.g. IDs could be "001", "002", etc. or "01f", "01m" to differentiate female and male participants).

2.1. Stimuli

Most of the tasks included in the package require stimuli in the form of sounds (i.e. "*.wav*" files). These files will be in a single folder with a unique name.

2.2. The loading file

Praatception is meant to be as easy to use as possible, but it also needs to fit the requirements of several different experiments. In the hope that every researcher finds a way to accommodate the tool to the necessities of their studies, some personalization work is needed ahead of the experiment launching.

The loading file is a simple text (*.txt*) file that the researcher needs to provide *Praatception*, and it is meant to facilitate the post-processing of the data. Suppose an experiment in which the word *house* is read by five different speakers, and listeners are required to choose whether they are native speakers or not. Each of the files is named *house_A.wav*, *house_B.wav*, *house_C.wav*, *house_D.wav* and *house_E.wav*. For the results file, it would be good to have marked which of the five files was read with a native or a foreign accent. This information can be added to the loading file as this:

```
house_A.wav|native
house_B.wav|native
house_C.wav|foreign
house_D.wav|foreign
house_E.wav|native
```

This way, in the results file, each stimulus will be linked to the variables defined by the researcher. This can be done with as much variables as wanted, so different variables can be automatically added to the results file. The variables only need to be included in the same way. For example, the variables *nativeness*, *sex* and *age* could be stacked:

```
house_A.wav|native|male|23
house_B.wav|native|female|15
house_C.wav|foreign|male|16
house_D.wav|foreign|female|23
house_E.wav|native|female|42
```

At a certain point, *Praatception* will ask the user the name of the provided variables separated by commas (in this case: `nativeness, gender, age`), so the results table is correctly constructed (see section 4 for more detailed information about the results file).

A special case is the loading file for a discrimination task. In this case, not one stimuli but two need to be loaded. The way to specify this is by including the "+" sign between them:

```
house_A.wav+house_B.wav|different
house_B.wav+house_B.wav|same
house_B.wav+house_C.wav|different
house_A.wav+house_C.wav|different
house_C.wav+house_C.wav|same
```

The loading file is also the place to specify the text that should be included with the presentation of each stimulus. This can be done by adding the desired text after the variables between hash symbols. A loading file

like the one below will present the word *house* on screen when the sound *house_A.wav* is played:

```
house_A.wav|native|male|23#house#
house_B.wav|native|female|15
house_C.wav|foreign|male|16
house_D.wav|foreign|female|23
house_E.wav|native|female|42
```

Finally, the loading file also offers the possibility to include the expected answer, if there is one. If the researcher requires it to be specified, this can be done by including it between dollar symbols at the end of each line:

```
house_A.wav|native|male|23#house#
house_B.wav|native|female|15
house_C.wav|foreign|male|16$foreign$
house_D.wav|foreign|female|23$foreign$
house_E.wav|native|female|42
```

If the answer of the participant is the same as the one defined this way, the results file will present the value 1 under the column labelled *ok*; if, on the contrary, the answer is not the same, the value stored in the results file will be 0. If no correct answer is defined, the answer will be collected as NA.

Note that neither the variables, the on-screen word nor the correct answer are required; they are only a way to facilitate the data post-processing for certain tasks. It is completely optional for the researcher to include them or not. The loading file is the only stage of the process in which the user has to modify external files. Everything else will be customized through a pop-up formulary that appears when the plug-in is summoned.

2.3. Randomization strategies

At the moment, three different options are included as randomization strategies: *no randomization*, *basic randomization* and *no repeats randomization*. The first option will present the stimuli as they are ordered in the folder (generally in alphabetical order); the second one will randomize the sounds without taking care of anything else; finally, the *no repeats* strategy allows the user to prevent the same variable to appear twice in a row.

The user can specify which variable not to repeat consecutively. For example, in the aforementioned loading files, the variable *age* could be specified, so the stimuli will all be randomized but those stimuli with the same age (i.e. *house_A.wav* and *house_D.wav*) will not appear consecutively during the experiment.

2.4. Final pre-requisites

The researcher will also be asked by *Praatception* whether some specific text should appear permanently throughout the experiment or not (e.g. "*Please, type what you just heard*" can appear during an intelligibility task). Also, each of the interfaces will require different specifications (e.g. the intelligibility task may require information about whether the answer should ignore capital letters or non-alphabetic symbols). These will be detailed in the next section.

3. EXPERIMENTAL PARADIGMS

In this section, the currently available interfaces will be described. Note that the figures provided are mock-ups and not the actual window. Two reasons motivated the non-inclusion of the actual interfaces: on one hand, the space limitations of a publication, that require optimisation; secondly, the fact that an actual screenshot would represent only one of the many options that *Praatception* offers, contrarily to the schematic representations presented here.

3.1. Intelligibility

The intelligibility task interface included in *Praatception* is shown in Figure 1. It can be combined with a likert scale for comprehensibility checking, but it is not required. *Praatception* can manage some other options, such as the possibility to ignore capitalization or remove specific symbols from the analysis.

(1) (optional): Permanent text (intelligibility). This will be displayed throughout the whole experiment. It is defined when the experiment is generated.
(2) (optional): Permanent text (comprehensibility). Instructions for the comprehensibility task, if it is included.
(3) and (4) (required if the comprehensibility task is enabled): Text for the ends of the comprehensibility scale.
(5) (required if the comprehensibility task is enabled): Number of steps in the likert scale. The user can choose 5, 7 or 9 points.
(6) (optional): *Repeat* button.
(7) (required): Text box where the user inserts the heard utterance.
(8) (required): The present stimulus vs the total number of stimuli (e.g. 4/100).

3.2. Categorisation

The categorisation task included is structured in the form of a two-alternative forced choice (2AFC) task (see Figure 2). The participant will hear one sound and

Figure 1: Mock-up of the interface for an intelligibility and comprehensibility task.

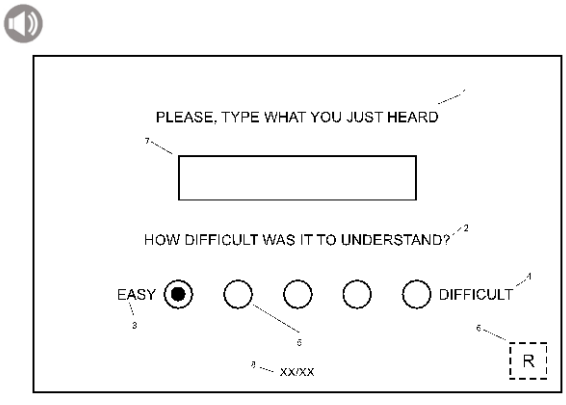


Figure 2: Mock-up of the interface for a categorisation task.

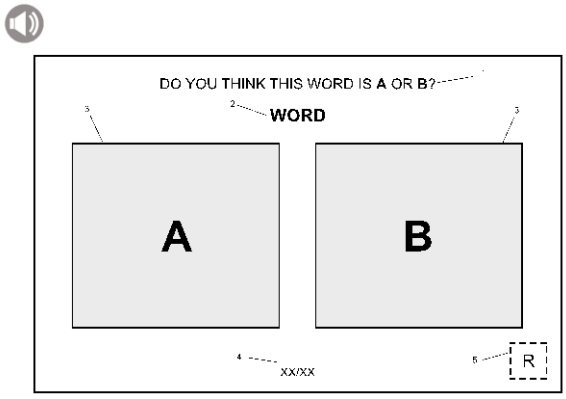
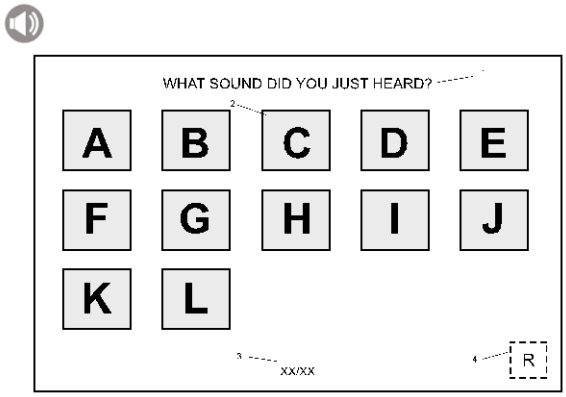


Figure 3: Mock-up of the interface for an identification task.



choose between one of the two presented possible answers.

(1) (optional): Permanent text. This will be displayed throughout the whole experiment. It is defined when the experiment is generated.
(2) (optional): The sound/word/sentence reproduced orthographically transcribed.
(3) (required): The two options presented for the user to choose one of them by clicking.
(4) (required): The present stimulus vs the total number of stimuli (e.g. 4/100).
(5) (optional): *Repeat* button.

3.3. Identification

In the identification task (see Figure 3), the participant will be presented with a bigger set of options (the consonants in a phonological system, a big set of words, etc.). His task will be to choose the one that most suits the proposed utterance.

(1) (optional): Permanent text. This will be displayed throughout the whole experiment. It is defined when the experiment is generated.
(2) (required): Buttons for options. In the pre-processing of the experiment, the researcher can specify any number of options between 3 and 15 and label them accordingly with the purpose of the experiment.
(3) (required): The present stimulus vs the total number of stimuli (e.g. 4/100).
(4) (optional): *Repeat* button.

3.4. Discrimination

The interface for the discrimination task is similar to the one for the categorisation task (see section 3.2). The difference for the interface of this task is that it is not possible to include the utterance orthographically transcribed on-screen. Note that the loading file is also different (see section 2.2).

The discrimination task is presented separately from the one for the categorisation task, even though they are both similar. The reason for this is that some researchers may find it useful to adjust specific settings, such as inter-stimuli time, instead of having to prepare the duplets manually. Also, and even though it is a feature not yet implemented, the analysis of the results will become more optimal if *Praatception* is able to identify them as a discrimination task, and not a categorisation one.

At the moment, the discrimination task included in *Praatception* is AX, i.e., two stimuli are presented consecutively and the listener is asked to choose whether they sound the same or different. For the moment, other discrimination tasks, such as ABX or AABA can be introduced with the categorisation interface, but the stimuli need to be prepared beforehand and the inter-stimulus time controlled by the researcher.

4. THE RESULTS FILE

Once the experiment is finished, the researcher will find the results in a tab-separated *.txt* file with headers in the same folder where the experiment file is. To ensure that the result files are not overwritten by repeated IDs, each results file will be named following the next pattern:

```
ID_W_M_D_Hor_Min_Sec_Y.txt
```

Where `ID` is the selected ID, `W` is the day of the week, `M-D-Y` are the current Month, Day and Year respectively and `Hor`, `Min` and `Sec` are the current Hour, Minute and Second. This way, the results files will never be overwritten.

Each of the columns of the results file (i.e. each of the tab-separated fields) will store the information of the experiment, the participant and the variables. Let us suppose a categorisation experiment designed to check the perception of a five steps continuum between a [ba] and a [da] syllables. The files are named consecutively from *ba_1.wav* to *ba_5.wav*. The researcher wants to store both ends of the continuum as two separate variables, *cont_str* and *cont_end*, and also wants to store the step of the continuum as a separate variable (1-5). Finally, as steps 1 and 5 are undoubtedly [ba] and [da] respectively, the user wants to check the % of correct answers in those specific points. The loading file will look like this:

```
ba_1.wav|b|d|1$b$
ba_2.wav|b|d|2
ba_3.wav|b|d|3
ba_4.wav|b|d|4
ba_5.wav|b|d|5$d$
```

Once the experiment is finished, the results file will be structured as shown in Table 1.

Every defined variable is presented in columns 3 to 7. Note that, in the case of those stimuli for which a correct answer has not been defined, the column *ok* presents the value NA. The reaction time is also stored in the last column.

The results for each participant will be stored separately, but, as they all have the same format, they can be merged afterwards in the analysis software of choice of the researcher.

5. FUTURE WORK

We have presented a tool aimed to facilitate research in perceptual phonetics. *Praatception* is currently in its development phase. The experimental interfaces showed in this chapter have already been tested and used in specific experiments, but they still lack some personalisation adjustments. The most immediate target is, therefore, to adapt the tasks to a customizable, user-friendly system that enables it to be used by any researcher.

It is also expected to include an automatic plot-generating system that, in a simple way, allows the researcher to make a quick analysis of the results. This could be useful to detect anomalies in the experiment that can potentially be solved so the results are not completely disregarded in an afterwards, deeper analysis. Please note that this is not meant to be a substitute of proper analysis/plotting tools, but just a glimpse of how the experiment is advancing.

Of course, *Praatception* is created with the spirit of sharing knowledge and make it grow. The scripts are available for the researcher to modify them as they best suit his or her requirements. Also, the e-mail of the researcher is open to suggestions, either to generate new experimental interfaces or to include others that may already exist. The final aim of the tool is to help the scientific community to find useful, standard methods for the research in perceptual phonetics; the more input the tool receives, therefore, the better.

Table 1: Example of the results file generated after a categorisation task.

ID	*file*	*cont_st*	*cont_ed*	*step*	*answer*	*ok*	*rt*
001	ba_1	b	d	1	b	1	0.145
001	ba_2	b	d	2	b	NA	0.265
001	ba_3	b	d	3	b	NA	1.56
001	ba_4	b	d	4	d	NA	0.43
001	ba_5	b	d	5	b	0	6.68

6. REFERENCES

Abercrombie, D. (1957). Direct palatography. *STUF - Language Typology and Universals, 10*(1–4). doi:10.1524/stuf.1957.10.14.21

Aguete Cajiao, A., Fernández Rei, E., & Osorio Peláez, C. (2016). FOLERPA: a tool for building and conducting perceptual experiments. In E. Fernández Rei, L. de Castro Moutinho, & R. Lídia Coimbra (Eds.), *Dialectologia. Special Issue VI*. 245–275.

Barefoot, S. M., Bochner, J. H., Johnson, B. A., & Eigen, B. A. (1993). Rating deaf speakers' comprehensibility. *American Journal of Speech-Language Pathology, 2*(3), 31. doi:10.1044/1058-0360.0203.31

Boersma, P. & Weenink, D. (2016). Praat: doing phonetics by computer [Computer program]. Version 6.0.22, retrieved 15 November 2016 from http://www.praat.org/

Brodkey, D. (1972). Dictation as a measure of mutual intelligibility: A pilot study. *Language Learning, 22*(2), 203–217. doi:10.1111/j.1467-1770.1972.tb00083.x

Cohen, J., Macwhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, and Computers, 25*(2), 257–271. doi:10.3758/bf03204507

Dickinson, J. (2011). The impact of 'violating the heterosexual norm' on reading speed and accuracy. *Psychology, 02*(05), 456–459. doi:10.4236/psych.2011.25071

Egan, J. P. (1956). On the transmission and confirmation of messages in noise. *The Journal of the Acoustical Society of America, 28*(1), 161–161. doi:10.1121/1.1918118

García Lecumberri, M. L., Barra Chicote, R., Pérez Ramón, R., Yamagishi, J., & Cooke, M. (2014). Generating segmental foreign accent. *Proceedings of Interspeech*.

Gerrits, E. & Schouten, M. E. (2004). Categorical perception depends on the discrimination task. *Perception and Psychophysics, 66*(3), 363–376. doi:10.3758/bf03194885

Grant, A., Benons, R., Johns, A., Hobson, M., & Nichols, D. (2016). Foreign accent perception and processing with EEG. *IMPULSE - The Premier Undegraduate Neuroscience Journal*.

Legou, T., Marchal, A., Meynadier, Y., & André, C. (2008). 3D palatography. *International Seminar on Speech Production*, 369–372.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1958). The discrimination of speech sounds within and across phoneme boundaries. *Erratum. Journal of Experimental Psychology, 55*(4), 396–396. doi:10.1037/h0038297

López-Bascuas, L. E., Marín, C. C., & García, F. J. (1999). A software tool for auditory and speech perception experimentation. *Behavior Research*

Methods, Instruments, and Computers, *31*(2), 334–340. doi:10.3758/bf03207729

Lutz, M. (n.d.). Learning Python [Computer software].

Marchal, A. (1988). *La palatographie*. Paris: Centre National de la Recherche Scientifique.

Macmillan, N. A., Kaplan, H. L., & Creelman, C. D. (1977). The psychophysics of categorical perception. *Psychological Review*, *84*(5), 452–471. doi:10.1037/0033-295x.84.5.452

MATLAB [Computer software]. (n.d.).

Mcqueen, J. (1996). Phonetic categorisation. *Language and Cognitive Processes*, *11*(6), 655–664. doi:10.1080/016909696387060

Nelson, C. (1982). Intelligibility and non-native varieties of English. In B. B. Kachru (Ed.), *The other tongue: English across cultures*. Oxford: Pergamon Press.

Smith, L. E., & Nelson, C. L. (1985). International intelligibility of English: Directions and resources. *World Englishes*, *4*(3), 333–342. doi:10.1111/j.1467-971x.1985.tb00423.x

Pérez Ramón, R. (2019). Perceptual experiments in *Praat*: Beyond the standards. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 109–115). Málaga: Universidad de Málaga.

Subsidia: Tools and Resources for Speech Sciences

18

VILE-P: un corpus para el estudio prosódico de la variación inter e intralocutor

Joaquim Llisterri[1], María J. Machuca[1] y Antonio Ríos[1]

[1] Universitat Autònoma de Barcelona
e-mail: Joaquim.Llisterri@uab.cat, Mariajesus.Machuca@uab.cat, Antonio.Rios@uab.cat

Citation / Cómo citar este artículo: Llisterri, J., Machuca, M. J., & Ríos, A. (2019). VILE-P: un corpus para el studio prosodico de la variación inter e intralocutor. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 117–123). Málaga: Universidad de Málaga.

RESUMEN: El corpus VILE-P constituye un recurso que permite realizar estudios acústicos sobre la variación fonética inter e intralocutor en español. Contiene datos de 30 hablantes —seleccionados entre los que forman parte del corpus Ahumada—, correspondientes a la lectura de un texto fonéticamente equilibrado y a un minuto de habla espontánea para cada locutor, recogidos durante tres sesiones de grabación separadas en el tiempo. El etiquetado, revisado manualmente por expertos, se ha llevado a cabo considerando 16 niveles de análisis, de los cuales tres hacen referencia a la información segmental y 13 a la suprasegmental. El acceso al corpus está abierto a los investigadores que deseen utilizarlo sin fines lucrativos ni comerciales.

Palabras clave: corpus oral; prosodia; variación fonética; español.

ABSTRACT: VILE-P is a corpus intended for acoustic studies about phonetic variation between and within speakers in Spanish. It comprises data on 30 speakers—a subset of those in the Ahumada corpus—, and includes the reading of a phonetically balanced text and one minute of spontaneous speech per speaker, collected three times across different recording sessions apart in time. Labeling, manually reviewed by experts, considers 16 levels of analysis (3 segmental and 13 suprasegmental). The corpus is open access for researchers who wish to use it for non-commercial purposes.

Keywords: oral corpus; prosody; phonetic variation; Spanish.

1. INTRODUCCIÓN

Las investigaciones más tradicionales en torno a la identificación o al reconocimiento automáticos del locutor se han centrado casi exclusivamente en rasgos que reflejan las características acústicas asociadas a la voz, definidos como de 'nivel bajo', y que se estudian, además, en intervalos temporales relativamente breves; no obstante, en épocas más recientes se ha hecho patente la importancia de los denominados rasgos de 'nivel alto', que dan cuenta de la información fonética segmental, prosódica y léxica presente en fragmentos más amplios de la señal sonora (Shriberg, 2007).

En la fonética judicial se emplean habitualmente ambos tipos de rasgos cuando se trata, por ejemplo, de comparar la voz de una persona cuya identidad se conoce —es decir, una muestra indubitada— con la grabación de una voz que podría corresponder a esta persona o bien pertenecer a otra, muestra que se conoce como dubitada. En cambio, en las técnicas de tipo automático que se han desarrollado para llevar a cabo esta misma tarea se han tenido en cuenta, por lo general, únicamente los rasgos de nivel bajo. Shriberg (2007) describe algunas de las ventajas que para el reconocimiento automático del locutor pueden ofrecer los rasgos de nivel alto, como

> the possibility of increased robustness to channel variation, since features such as lexical usage or temporal patterns do not change with changes in acoustic conditions (p. 242).

Por su parte, Adami (2007) postula que la prosodia proporciona información específica acerca del hablante, y propone que las diferencias prosódicas entre locutores pueden determinarse analizando las características de la entonación, del acento o de los patrones rítmicos que resultan de las variaciones de la frecuencia fundamental y de los contornos de energía. Tanto Shriberg como Adami coinciden en señalar que los sistemas convencionales de reconocimiento automático del locutor no incorporan completamente la diversidad de niveles de información presente en el habla, carencia que, según estos autores, se compensaría con el estudio de los elementos prosódicos —como plantea Adami— o de los rasgos de

nivel alto —como sugiere Shriberg—. En este sentido, en el proyecto VILE-P se abordó, como se explica a continuación, el análisis de los elementos suprasegmentales con el fin de dotar a los especialistas en fonética judicial de nuevos datos acústicos que permitan comparar, con un mayor grado de certeza, voces de locutores dubitados e indubitados.

1.1. Los objetivos del proyecto VILE-P

Para complementar estudios anteriores y paliar, en parte, las carencias de información sobre los correlatos temporales de algunas unidades fonéticas de naturaleza suprasegmental en español, el principal objetivo del proyecto VILE-P (Estudio acústico y perceptivo de la variación prosódica inter e intralocutor en español, FFI2010-21690-C02-02) fue establecer el papel de los elementos prosódicos —y, en particular, los de carácter temporal— para definir la individualidad de un hablante, tanto desde el punto de vista acústico como desde el perceptivo[1]. Además, se pretendía determinar el umbral diferencial de la velocidad de elocución y establecer algunos de los correlatos perceptivos del ritmo en español, contribuyendo así a obtener datos con los que no se contaba en su momento y, en el caso del ritmo, a esclarecer algunos aspectos de los debates surgidos durante los últimos años acerca de la tipología rítmica de las lenguas (Marrero, López Bascuas y Martín Fernández, 2015). VILE-P se llevó a cabo conjuntamente con el proyecto CIVIL (Cualidad individual de voz e identificación de locutor, FFI2010-21690-CO2-01) (Alves, Gil, Pérez Sanz y San Segundo, 2014), como parte de una iniciativa más amplia, financiada por el Ministerio de Ciencia e Innovación entre los años 2011 y 2014, centrada en estudiar el papel que desempeñan los elementos suprasegmentales del habla en la caracterización del locutor en el ámbito de la fonética judicial.

1.2. La selección del corpus

El corpus utilizado para VILE-P fue el mismo que se empleó en un proyecto anterior, VILE-I (Estudio acústico de la variación inter e intralocutor en español, BFF2001-2551), en el que se segmentó y se etiquetó la señal de habla para analizar acústicamente los elementos segmentales (Battaner *et al.*, 2007; Machuca, Ríos y Llisterri, 2014). Teniendo en cuenta la necesidad de reutilizar los recursos lingüísticos ya creados, al iniciar VILE-I se llevó a cabo una revisión de los corpus orales disponibles en español, valorando su utilidad en relación con los objetivos del proyecto: el estudio de la variación fonética inter e intralocutor con fines judiciales. En una fase inicial se consideraron las bases de datos existentes en ese momento, el año 2001: Ahumada (Ortega, González Rodríguez y Marrero, 2000), Albayzín (Moreno *et al.*, 1993), EUROM (Chan *et al.*, 1995), MulText (Campione y Véronis, 1998) y SpeechDat (Draxler, van den Heuvel y Tropf, 1998).

De esta revisión se desprendió que, en general, los datos recogidos en lo que se refiere al nivel segmental eran ampliamente suficientes; sin embargo, los elementos suprasegmentales se encontraban representados en menor medida; EUROM y MulText proporcionaban materiales adecuados para el estudio de la variación entonativa en el nivel oracional, pero se limitaban exclusivamente a tareas de lectura. En cambio, Ahumada, además de la lectura, ofrecía también fragmentos de habla espontánea de un minuto de duración, así como muestras de variación temporal, con textos leídos a tres velocidades de habla diferentes; asimismo, el hecho de que para cada hablante se hubieran realizado tres sesiones de grabación separadas en el tiempo permitía el análisis de la variación intralocutor. Por tales motivos, se eligió este corpus como punto de partida de los trabajos previstos a lo largo del proyecto VILE-I y, posteriormente, se enriqueció con nuevos niveles de etiquetado en el marco de VILE-P, como se explica en el apartado 4.

2. LOCUTORES

De los 104 locutores masculinos incluidos en Ahumada (Ortega *et al.*, 2000), se seleccionaron 30 hablantes para crear el corpus empleado en VILE-I y en VILE-P, partiendo de dos criterios: la presencia de rasgos que denotaran su procedencia geográfica y la espontaneidad de su producción. Se tomaron en consideración únicamente aquellos locutores en los que era menos evidente la influencia dialectal —básicamente, los que mostraban las características propias de las variedades norte y centropeninsulares del español— y, de estos, los que entre las tareas diseñadas para obtener habla espontánea no optaron por describir el entorno en el que se encontraban en el momento de la grabación —ya que con esta tarea no se conseguía un grado de espontaneidad suficiente para los objetivos del proyecto—, sino que o bien narraron una historia o bien explicaron sus gustos y preferencias en lo que se refiere al deporte, al cine, a los juegos, etc.

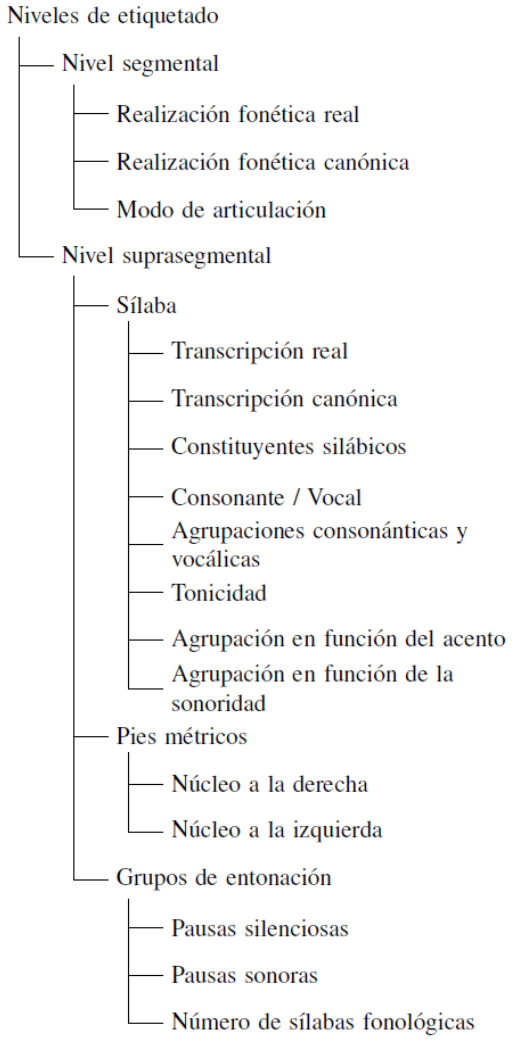
3. TAREAS Y SESIONES DE GRABACIÓN

Como uno de los objetivos de VILE-P era estudiar la variabilidad que se produce en las realizaciones tanto de un mismo locutor como al comparar distintos locutores, se seleccionaron las grabaciones de las tres sesiones en las que los hablantes leían a una velocidad de elocución normal un texto fonéticamente equilibrado, así como las correspondientes a las tres sesiones en las que se recogía habla espontánea. El texto leído, de una duración de aproximadamente un minuto, contiene 179 palabras y 712 segmentos, por lo que se han etiquetado 21 360 segmentos considerando los 30 informantes y las tres sesiones de grabación. La producción espontánea se limita también a un minuto

[1] En este proyecto participaron, además de los autores del presente trabajo, Elena Battaner (Universidad Rey Juan Carlos), Luis Enrique López Bascuas (Universidad Complutense de Madrid), Victoria Marrero (Universidad Nacional de Educación a Distancia), José Luis Martín Fernández (Universidad Complutense de Madrid) y Montserrat Riera (Universitat Autònoma de Barcelona). Toda la información relativa a VILE-P se encuentra recogida en VILE (2014).

Llisterri, J., Machuca, M. J., & Ríos, A. (2019). VILE-P: un corpus para el studio prosodico de la variación inter e intralocutor. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 117–123). Málaga: Universidad de Málaga.

Figura 1: Niveles de etiquetado de la señal de habla en el corpus VILE-P.



ypara este estilo de habla se cuenta, en total, con 80 379 segmentos. El intervalo entre las distintas sesiones de grabación oscilaba entre un mínimo de 11 días y un máximo de 40.

4. TRANSCRIPCIÓN Y ETIQUETADO

Los niveles de etiquetado del corpus VILE-P se ordenan jerárquicamente, de forma que se inicia la segmentación partiendo de la realización fonética de cada sonido vocálico o consonántico y se acaba con el etiquetado de los grupos de entonación. Se han establecido 17 niveles, pero de ellos solo 16 están relacionados con el etiquetado fonético. Concretamente, el último nivel hace referencia a la duración total de la señal de habla que contiene el fichero, ya que es un valor que se necesita para calcular otros parámetros de interés, como pueden ser la velocidad de articulación y la velocidad de elocución de un hablante. Por lo tanto, la anotación se estructura en 16 niveles, de los que tres hacen referencia a la información segmental y 13 a la suprasegmental, tal como puede observarse en la Figura 1. En la Figura 2 se muestran, en un ejemplo tomado del corpus, los niveles de etiquetadoque se describen a continuación.

a. Elementos segmentales

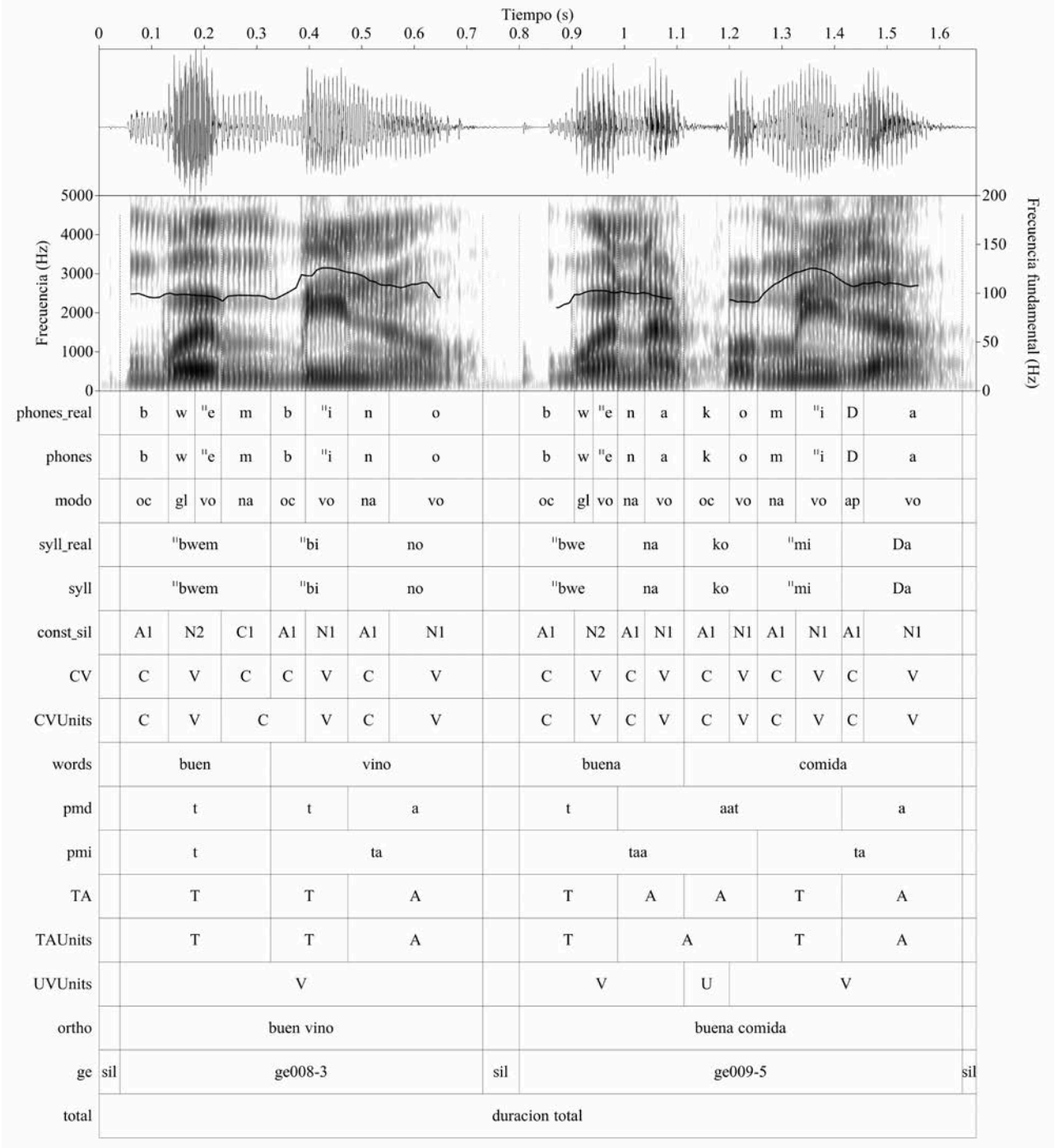
Para el análisis de los elementos segmentales se parte del nivel denominado phones_real, que contiene la segmentación y la transcripción fonética de cada nunciado tal como lo pronunciaron realmente los locutores. Tanto la segmentación de la señal como como el etiquetado se obtuvieron de forma semiautomática mediante el programa EasyAlign (Goldman y Schwab, 2014) a partir de la transcripción ortográfica del contenido de la grabación y, posteriormente, se revisaron y modificaron manualmente. El alfabeto utilizado para la transcripción fonética fue SAMPA (Wells, 2005) en su adaptación al español realizada por Llisterri y Mariño (1993); el inventario empleado se recoge en la Tabla 1.

El nivel definido como phones contiene la segmentación y la transcripción fonética canónica de la señal de habla. Se entiende por 'forma canónica' la que sería propia del estándar culto centropeninsular del español, tal como se describe, por ejemplo, en las obras de Navarro Tomás (1918) o de Quilis (1993). La segmentación que aparece en este caso es idéntica a la del nivel phones_real; sin embargo, la transcripción varía en algunas ocasiones a causa de las diferencias entre la realización de los hablantes —reflejada en la transcripción fonética real— y la canónica. Los principales fenómenos que se representan de manera distinta en uno y otro nivel corresponden a coalescencias, a elisiones o epéntesis, a procesos de debilitamiento o de refuerzo, a sonorizaciones o ensordecimientos y a aspiraciones que puede realizar el

Tabla 1: Inventario fonético utilizado en el etiquetado segmental del corpus VILE-P. Los símbolos a la izquierda de cada columna corresponden a la transcripción en SAMPA y, los de la derecha, a su equivalente en el Alfabeto Fonético Internacional.

Oclusivas		Fricativas		Aproximantes		Africada	
p	[p]	f	[f]	B	[β̞]	tS	[t͡ʃ]
t	[t]	T	[θ]	D	[ð̞]		
k	[k]	s	[s]	G	[ɣ̞]		
b	[b]	z	[z]				
d	[d]	jj	[ʝ]				
g	[g]	x	[x]				
		h	[h]				
Nasales		**Laterales**		**Róticas**			
m	[m]	l	[l]	r	[r]		
n	[n]	L	[ʎ]	4	[ɾ]		
J	[ɲ]						
N	[ŋ]						
Vocales		**Paravocales**					
i	[i]	j	[i̯]				
e	[e]						
a	[a]						
o	[o]						
u	[u]	w	[u̯]				

Figura 2: Etiquetado del fragmento *buen vino, buena comida* en el corpus VILE-P.



locutor, tal como se muestra en los ejemplos de la Tabla 2.

Finalmente, el nivel denominado modo corresponde al modo de articulación de los sonidos segmentados en el primer nivel.

El inventario de etiquetas comprende, en este caso, los siete modos consonánticos del español (oclusivo, fricativo, aproximante, africado, nasal, lateral y rótico), además de la clasificación como vocal o paravocal para diferenciar el carácter silábico o no silábico de los elementos vocálicos (Tabla 1). Conviene precisar que en la Figura 2 aparece la etiqueta gl, correspondiente a *glide*, término empleado en el sistema interno de etiquetado para referirse a las paravocales; del mismo modo, vibrante se utilizó para etiquetar las consonantes róticas.

b. Elementos suprasegmentales

Toda la información recogida en el resto de niveles de etiquetado del corpus corresponde al ámbito

suprasegmental, de manera que se consideran las características relacionadas con la sílaba, con el pie métrico, con las pausas y con los grupos de entonación. El nivel denominado syll_real contiene la segmentación y la transcripción en sílabas correspondientes a la pronunciación real de los locutores, mientras que en el nivel syll se representa la división silábica canónica. Las principales diferencias entre ambos niveles tienen que ver con los procesos de reajuste silábico que se dan, sobre todo, en el habla espontánea; por ejemplo, *es al zoo* se transcribe en su forma canónica como [e.sal.ˈθo.o], aunque en el corpus aparece como [es.al.ˈθo].

En el siguiente nivel, const_sil, se etiquetan los constituyentes silábicos a partir de la información contenida en los niveles phones_real y syll_real. Se trata, por tanto, de una representación que refleja la realización de los hablantes. Se propone aquí una segmentación de las sílabas en tres tipos de constituyentes —ataque, núcleo y coda— que se etiquetan con su inicial (A, N y C) acompañada del número de elementos que lo conforman. Así, por ejemplo, un ataque de dos elementos se representa como A2 (*flor*), un núcleo de tres elementos se etiqueta como N3 (*buey*) y una coda de un elemento aparece como C1 (*sal*). El objetivo de este etiquetado es doble: por un lado, se indica la posición de la consonante dentro de la sílaba y, por otro, se establece si existe un grupo consonántico o un grupo vocálico.

En el nivel CV se parte de la segmentación de los sonidos realizada previamente en phones_real y se clasifican los segmentos como consonánticos (C) o como vocálicos (V), para después, en el siguiente nivel, CVUnits, agrupar las secuencias de consonantes o las de vocales, etiquetándolas también como C o como V. Además, se considera la tonicidad de las sílabas en el nivel TA y la agrupación de los elementos tónicos y átonos en el nivel TAUnits. En el nivel UVUnits se realiza la agrupación de los sonidos sordos y la de los sonoros: cada secuencia de sonidos sordos consecutivos se etiqueta como U y cada conjunto de sonidos sonoros sucesivos, como V. En la Tabla 3 se ofrecen algunos ejemplos de estos niveles de etiquetado.

Tabla 2: Diferencias entre la transcripción de la forma canónica y la transcripción de la pronunciación real.

Fenómeno	Canónica	Real
Coalescencia	[me e]	[me]
Elisión	[akoˌsˈt̪aðo]	[akoˌsˈt̪ao]
Epéntesis	[ˈu̯erto]	[ˈɣu̯erto]
Debilitamiento	[akoˌsˈt̪aðo]	[aɣˌosˈt̪aðo]
Sonorización	[iˈɾemos a]	[iˈɾemoˌs a]
Aspiración	[los ke]	[loh ke]

Tabla 3: Etiquetado en los niveles relativos a las secuencias de consonantes y vocales, a la tonicidad y a la sonoridad.

Nivel	Transcripción	Etiquetado
CV	[floɾ]	CCVC
	[bu̯ei̯]	CVVV
CVUnits	[floɾ]	CVC
	[bu̯ei̯]	CV
TA	[en.la.ˈka.ma.ɾa]	AATAA
TAUnits	[en.la.ˈka.ma.ɾa]	ATA
UVUnits	[en.la.ˈka.ma.ɾa]	VUV

En cuanto a los niveles que hacen referencia a los pies métricos, en el nivel pmd la segmentación se lleva a cabo considerando que los pies métricos tienen su núcleo a la derecha, mientras que en el nivel pmi se ofrece la segmentación propia de los pies métricos con el núcleo a la izquierda. Cada pie métrico se etiqueta mediante una cadena de símbolos a y t, que se refieren a las sílabas átonas (a) y a la tónica (t) que contiene. Las sílabas extramétricas se etiquetan únicamente con la cadena de sílabas átonas correspondiente. Así, el sintagma *buena comida* se representaría como [bu̯e nakomi(ða)] en el nivel pmd y como [bu̯enako mi(ða)] en el pmi.

En el nivel word se segmentan cada una de las palabras del fichero, entendiendo como palabra una unidad lingüística, ya sea una forma dotada de significado léxico (sustantivo, adjetivo, verbo y adverbio) o una palabra funcional, que solo aporta información gramatical.

Por último, se han etiquetado los grupos de entonación (ge) y se ha incorporado su correspondiente representación ortográfica (ortho). En este caso, se han tenido en cuenta cinco fenómenos que constituyen los criterios a partir de los que se establece la existencia de una frontera prosódica: pausas silenciosas, pausas sonoras, alargamientos, voz rota (en inglés, *creaky voice*) e inflexiones tonales. Al mismo tiempo que se dividen los grupos de entonación, se etiquetan también las pausas sonoras (sil son) y las pausas silenciosas (sil). Cuando al inicio de un grupo entonativo se encuentra una consonante oclusiva sorda, se ha optado por emplear la etiqueta sil ocl para señalar que en ese silencio está incluida la fase de oclusión de la consonante. A este respecto, cabe precisar que se ha adoptado un umbral con un valor de 90 ms para los casos en los que una pausa silenciosa va seguida de una oclusiva sorda: si la ausencia de energía sonora supera los 90 ms, se considera que existe una pausa; en cambio, si no sobrepasa este valor, se considera que el silencio corresponde únicamente a una oclusión.

Para separar los grupos de entonación se ha diferenciado entre pausa sonora y alargamiento. Se han anotado como pausas sonoras aquellos casos en los que se produce un alargamiento de un murmullo nasal o un alargamiento de una vocal (generalmente /e/ o /a/) sin que se dé una continuidad entre el material fónico que precede o que sigue a ese alargamiento como, por ejemplo, en | sil | sil son (*eeeh*) | *por la tarde nos fuimos a Mérida a comer*. Cuando se considera que existe una continuidad respecto a lo que previamente estaba diciendo el locutor, el alargamiento se incluye dentro del grupo de entonación en el que aparece y no se segmenta un nuevo grupo, pues no se trataría de una pausa sonora: | sil | *estuvimos haciendooo* | sil |.

Sin embargo, en los fragmentos en los que ese alargamiento va acompañado de una inflexión tonal o de voz rota, se marca un límite prosódico.

En algunos segmentos se detectan irregularidades en las vibraciones de los pliegues vocales (es decir, voz rota), que aparecen en momentos en los que el locutor rectifica, organiza su discurso o planifica la continuación del mismo. La voz rota únicamente se ha tenido en cuenta para delimitar grupos entonativos si se presenta acompañada de una inflexión tonal o de un alargamiento. También se han establecido divisiones entre grupos de entonación cuando se percibe auditivamente un cambio tonal que se refleja, acústicamente, en una inflexión de la curva melódica. En algunas ocasiones, parece percibirse una variación tonal que, sin embargo, no se visualiza en los movimientos de la frecuencia fundamental; en estos casos, se ha optado por definir un único grupo de entonación. Asimismo, las repeticiones y rectificaciones, si no comportan un cambio tonal o si no se cumplen algunos de los criterios mencionados previamente, se incluyen dentro del mismo grupo de entonación que la secuencia que viene a continuación.

5. CONCLUSIONES

El etiquetado del corpus VILE-P puede calificarse de exhaustivo, no solo porque contiene un número importante de niveles de etiquetado (16 niveles), sino también porque la segmentación y la transcripción automáticas de la señal se han revisado manualmente en todos los niveles, de modo que reflejan la producción real de los sonidos del habla por parte del locutor, a la vez que se ofrecen también las formas canónicas esperables en la norma culta centropeninsular del español. En este sentido, VILE-P es un corpus que permite realizar estudios lingüísticos en los que se pueden relacionar los procesos fonológicos con los fonéticos en dos estilos bien diferenciados, como son la lectura y el habla espontánea. También es un corpus que contiene información fonética relevante para el reconocimiento, la identificación y la verificación automáticas de locutores, así como para la práctica judicial, ya que se han considerado diferentes hablantes en distintas situaciones comunicativas y en tres sesiones de grabación separadas en el tiempo, lo que facilita el estudio de la variación inter e intralocutor. Por otro lado, el hecho de que todas las unidades estén segmentadas y etiquetadas puede resultar útil para desarrollar nuevas herramientas de etiquetado automático o para entrenar sistemas de reconocimiento incorporando un conocimiento fonético que contribuya a mejorar su rendimiento.

El corpus VILE-P estará próximamente a disposición de los investigadores que deseen utilizarlo sin fines lucrativos ni comerciales en el *Dipòsit Digital de Documents* (http://ddd.uab.cat) de la Universitat Autònoma de Barcelona.

6. REFERENCIAS

Adami, A. G. (2007). Modeling prosodic differences for speaker recognition. *Speech Communication*, *49*(4), 277–291. https://doi.org/10.1016/j.specom.2007.02.005

Alves, H., Gil, J., Pérez Sanz, C. y San Segundo, E. (2014). La cualidad individual de la voz y la identificación del locutor: el proyecto CIVIL. En Y. Congosto, M. L. Montero y A. Salvador (Eds.), *Fonética experimental, educación superior e investigación* (Vol. 1, pp. 591–612). Madrid: Arco/Libros.

Battaner, E., Carbó, C., Gil, J., Llisterri, J., Machuca, M. J., Madrigal, N., . . . y Ríos, A. (2007). VILE: Estudio acústico de la variación inter e intralocutor en español. En M. González González, E. Fernández Rei y B. González Rei (Eds.), *III Congreso Internacional de Fonética Experimental, Actas do congreso organizado pola Dirección Xeral de Creación e Difusión Cultural, a Universidade de Santiago de Compostela e a Real Academia Galega, Santiago, 24-26 de outubro de 2005* (pp. 157–167). Santiago de Compostela: Xunta de Galicia.

Campione, E. y Véronis, J. (1998). A multilingual prosodic database. En *ICSLP 1998. Proceedings of the 5th International Conference on Spoken Language Processing* (pp. 3163–3166). Sydney, Australia. November 30-December 4, 1998.

Chan, D., Fourcin, A., Gibbon, D., Granström, B., Huckvale, M., Kokkinakis, G., . . . y Zeiliger, J. (1995). EUROM - A spoken language resource for the EU - The SAM Projects. En *Eurospeech 1995. Proceedings of the 4th European Conference on Speech Communication and Technology* (Vol. 1, pp. 867–870). Madrid, Spain. 18-21 September, 1995.

Draxler, C., van den Heuvel, H. y Tropf, H. (1998). SpeechDat experiences in creating large multilingual speech databases for teleservices. En *LREC 1998. Proceedings of the 1st International Conference on Language Resources and Evaluation* (Vol. 1, pp. 361–366). Granada, Spain. 28-30 May, 1998.

Goldman, J.-P. y Schwab, S. (2014). EasyAlign Spanish: an (semi-)automatic segmentation tool under Praat. En Y. Congosto, M. L. Montero y A. Salvador (Eds.), *Fonética experimental, educación superior e investigación* (Vol. 1, pp. 629–640). Madrid: Arco/Libros.

Llisterri, J. y Mariño, J. B. (1993). *Spanish adaptation of SAMPA and automatic phonetic transcription* (Technical Report N.º SAM-A/UPC/001/V1). ESPRIT Project 6819 SAM-A, Speech Technology Assessment in Multilingual Applications. Consultado en http://liceu.uab.cat/~joaquim/publicacions/SAMPA_Spanish_93.pdf

Machuca, M. J., Ríos, A. y Llisterri, J. (2014). Conocimiento fonético y fonética judicial. En A.

Hidalgo, C. Hernández y F. J. Cantero (Eds.), *La fonética como ámbito interdisciplinar. Estudios de fonopragmática, fonética aplicada y otras interfaces. Quaderns de Filologia: Estudis Lingüístics XIX* (pp. 95–111). València: Universitat de València.

Marrero, V., López Bascuas, L. E. y Martín Fernández, J. L. (2015). El ritmo lingüístico en la caracterización del locutor. Percepción y aplicaciones judiciales. En A. Cabedo (Ed.), *Perspectivas actuales en el análisis fónico del habla. Tradición y avances en la fonética experimental* [Anejo núm. 7 de *Normas. Revista de Estudios Lingüísticos Hispánicos*] (pp. 335–346). València: Departamento de Filología Española, Universitat de València.

Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J. B. y Nadeu, C. (1993). Albayzín speech database: Design of the phonetic corpus. En *Eurospeech 1993. Proceedings of the 3rd European Conference on Speech Communication and Technology* (Vol. 1, pp. 175–178). Berlin, Germany. 21-23 September, 1993.

Navarro Tomás, T. (1918). *Manual de pronunciación española*. Madrid: Centro de Estudios Históricos, Junta para Ampliación de Estudios e Investigaciones Científicas.

Ortega, J., González Rodríguez, J. y Marrero, V. (2000). AHUMADA: A large speech corpus in Spanish for speaker characterization and identification. *Speech Communication, 31*(2-3), 254–264.

Quilis, A. (1993). *Tratado de fonología y fonética españolas*. Madrid: Gredos.

Shriberg, E. E. (2007). Higher level features in speaker recognition. En C. Muller (Ed.), *Speaker classification I. Fundamentals, features, and methods* (pp. 241–259). Berlin - Heidelberg - New York: Springer.

VILE. (2014). VILE, Estudio acústico y perceptivo de la variación inter e intralocutor en español [Página web]. Consultado en http://liceu.uab.cat/~joaquim/VILE.html

Wells, J. C. (2005). SAMPA Computer Readable Phonetic Alphabet [Documento en línea]. Consultado en http://www.phon.ucl.ac.uk/home/sampa/index.html

Llisterri, J., Machuca, M. J., & Ríos, A. (2019). VILE-P: un corpus para el studio prosodico de la variación inter e intralocutor. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 117–123). Málaga: Universidad de Málaga.

Subsidia: Tools and Resources for Speech Sciences

Génesis y aspectos fundamentales de ProDis

Ana Maria Fernández Planas[1], Paolo Roseano[1], Wendy Elvira-García[1] y Simone Balocco[1]

[1] Universitat de Barcelona
e-mail: anamariafernandezp@ub.edu

Citation / Cómo citar esta publicación: Fernández Planas, A. M., Roseano, P., Elvira-García, W., & Balocco, S. (2019). Génesis y aspectos fundamentales de ProDis. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 125–132). Málaga: Universidad de Málaga.

RESUMEN: El objetivo de este trabajo consiste en presentar ProDis, una herramienta informática para el análisis dialectométrico de la entonación creada por el equipo del *Laboratori de Fonètica* de la *Universitat de Barcelona*. El trabajo presenta un breve estado de la cuestión sobre los sistemas de dialectometría disponibles en la actualidad y se hace especial hincapié en los sistemas capaces de tratar datos prosódicos numéricos y en la necesidad del Atlas Multimedia de la Entonación del Espacio Románico (AMPER) de crear una herramienta propia que sirva para gestionar los datos recogidos mediante el corpus fijo del proyecto en cuestión. Asimismo, se detallan los sistemas anteriores que han perimido el nacimiento de la herramienta. Posteriormente, se realiza una presentación de ProDis en la que se abordan el método que usa para calcular las distancias prosódicas, las salidas del programa y la información que se puede obtener de dicho análisis.

Palabras clave: dialectometría; entonación; lenguas romances.

ABSTRACT: The aim of this paper is to present ProDis, a software for the dialectometric analysis of intonation. In the first part of the article we present a state of the art of currently available dialectometrical tools. Since the prosodic data that are collected within the *Atlas Multimedia de la Entonación del Espacio Románico* (AMPER) are numeric, we pay special attention to the tools that can dialectometrize numeric prosodic data. After resuming the features of the existing tools that, to a certain extent, can carry out the dialectometric analyses of numeric data, we present more in detail the characteristics of ProDis, the prosodic dialectometrical tool created at the Phonetics Laboratory of the University of Barcelona. The aspects of ProDis that are described are the method used to calculate prosodic distances, the outputs of the program, and the information that can be obtained thanks to the prosodic dialectometry.

Keywords: dialectometry; intonation; Romance languages.

1. INTRODUCCIÓN

Los atlas lingüísticos clásicos constituyen una fuente maravillosa de datos fonéticos, morfológicos, sintácticos y, sobre todo, léxicos (recuérdese por ejemplo el ALiR ―Atlas Linguistique Roman―, el ALPI ―Atlas Lingüístico de la Península Ibérica―, el ALE ―Atlas Linguistique de l'Europe―, o el ALAC ―Atlas Lingüístico de América Central―, por ejemplo, entre otros muchos). El estudio de sus datos fue el punto de partida para el establecimiento de isoglosas y fronteras dialectales que clasificaban los puntos de encuesta en grupos a partir de rasgos considerados muy relevantes de forma cualitativa por los investigadores.

Desde los años 70 y 80 del siglo XX se ha producido un paso natural en el desarrollo dialectal de la mano de la llamada dialectometría, que pretende establecer agrupaciones entre la masa de datos empíricos obtenidos en grandes bases de datos a partir de criterios cuantitativos y de procedimientos estadísticos objetivos. El término dialectometría se debe a Séguy (1973), uno de los padres de dicha disciplina junto con Guiter, aunque fue Goebl quien le dio un impulso definitivo. Goebl (1981) define la dialectometría como una alianza metodológica entre la geolingüística y la taxonomía numérica como disciplina matemática. Exactamente, el autor lo expone de forma sintética de la siguiente manera: dialectometría = geografía lingüística + taxonomía numérica (Goebl, 1981, p. 349).

Ciertamente, lo que los estudios dialectométricos pretenden es utilizar una enorme cantidad de datos que se han generado a través de los estudios dialectológicos y los atlas lingüísticos para establecer agrupaciones entre la masa de datos empíricos disponibles y obtener una distribución en el espacio virtual de los datos (Fernández Planas, Roseano, Martínez Celdrán y Romera Barrios, 2011, p. 145), o también en forma de agrupaciones reflejadas en dendrogramas. Sus resultados permiten una rápida asociación entre los

elementos considerados a partir de su cercanía o su lejanía —es decir, de sus semejanzas o de sus diferencias— y posibilitan condensar una gran cantidad de información cuantitativa en un espacio relativamente reducido.

La dialectometría no pretende eliminar el estudio dialectológico tradicional, sino que busca completarlo y erigirse como una herramienta esencialmente útil cuando se manejan cantidades enormes de datos a partir de grandes bases. Sin embargo, ofrece ciertas ventajas respecto a la dialectología tradicional: (1) permite gestionar sin gran esfuerzo por parte del investigador grandes cantidades de datos, de donde se puede inferir que permite llegar a conclusiones estadísticamente fiables, en el sentido de que van más allá de las meras intuiciones de los investigadores; (2) no hay apriorismos en el tratamiento de los datos, ya que el estudio se centra en valoraciones cuantitativas y no cualitativas. Este hecho supone un cambio radical respecto a la dialectología tradicional, más bien cualitativa, y constituye el fundamento de la reticencia de algunos autores; y (3) la forma de presentación de los datos (análisis de clúster en dendrogramas y escalamiento multidimensional, básicamente) es totalmente visual y favorece una comprensión bastante rápida de los hechos.

Por lo que respecta a las lenguas romances, el método se ha aplicado principalmente a las áreas lingüísticas del ladino (Goebl, 1993; Bauer, 2005), el italiano (Bauer, 2003), el francés (Séguy, 1973; Verlinde, 1988; Goebl, 1987; Goeble, 2003), el gallego (Álvarez Blanco, Dubert, y Sousa, 2006; Sousa, 2006; Saramago, 2002), el bable (D'Andrés Díaz, Álvarez-Balbuena García, y Suárez Fernández, 2003) o el catalán (Clua, 2004, Polanco, 1992). Fuera de la Romania se utiliza también en estudios dialectológicos de lenguas como el holandés (Heeringa y Nerbonne, 2001), el inglés (Goebl y Schiltz, 1997), o el euskara (Aurrekoetxea, 1992). Normalmente se ha trabajado con datos fonético-segmentales, morfológicos o léxicos, como hemos dicho. El estudio de los aspectos prosódicos de diferentes variedades se ha trabajado muchísimo menos.

En el seno de macroproyecto AMPER, Atlas Multimedia de Prosodia del Espacio Románico (Contini, 1992; Contini *et al.*, 2002; Romano y Contini, 2001; Contini, Lai y Romano, 2002; Romano, 2003; Fernández Planas, 2005), tras tener muy avanzada una enorme base de datos prosódicos acústicos, se impone trabajar en la comparación y clasificación de las variedades románicas. De hecho, que en el marco de AMPER se llegara a utilizar la dialectometría era también un desarrollo natural y un paso esperable. Así pues, el llamado "corpus fijo" en el proyecto constituye un terreno óptimo para el estudio dialectométrico. AMPER trabaja con habla cercana a habla de laboratorio en el llamado corpus fijo que cuenta con frases enunciativas e interrogativas absolutas que presentan la misma estructura SVO y número de sílabas en todas las lenguas (más o menos), con todas las combinaciones acentuales posibles en todas las posiciones de la frase salvo en el verbo, con dos hablantes por punto de encuesta (como mínimo) que repiten cada uno de ellos tres veces cada frase.

2. LA DIALECTOMETRÍA Y SUS DATOS

En realidad, cuando hablamos de dialectometría no estamos refiriéndonos a una única técnica, sino a un paraguas metodológico que incluye técnicas distintas que trabajan con el mismo objetivo (la aplicación de técnicas estadísticas a grandes bases de datos para averiguar cómo se agrupan a partir de las distancias que mantienen los elementos entre sí a partir de sus características) pero con diferentes algoritmos y con diferentes tipos de datos.

Desde sus inicios, igual que la dialectología clásica, la dialectometría ha usado datos fonético-fonológicos segmentales, morfológicos, léxicos o sintácticos que se almacenan en bases de datos alfabéticos. Desde el punto de vista estadístico, eso implica que se necesitan algoritmos capaces de establecer distancias cuantitativas a partir de variables nominales, principalmente Levenshtein (Kessler, 1995). El uso de datos prosódicos provenientes de análisis acústicos, más recientemente, ha planteado una cuestión metodológica crucial: conviene operar con variables numéricas, lo cual implica el hecho de necesitar otro tipo de métricas para trabajar. Si los datos prosódicos se transcriben con símbolos alfabéticos, tanto en un nivel más superficial y cercano a las melodías acústicas, como en un nivel mucho más profundo o fonológico, de acuerdo con los postulados para el sistema, conseguimos una cadena alfabética que vuelve a necesitar algoritmos que trabajen con datos alfabéticos o nominales.

Existen herramientas disponibles para trabajar en dialectometría tanto con datos numéricos como con datos alfabéticos. Gabmap (Nerbonne, Colen, Gooskens, Kleiweg, y Leinonen, 2011) puede trabajar con ambas, pero con muchas limitaciones porque el programa no permite vectores y reduce cada variable a un único valor. VisualDialectometry (Goebl y Haimerl, 2004) o DiaTech (Aurrekoetxea, Fernández-Aguirre, Rubio, Ruiz, y Sánchez, 2013) operan con datos nominales pero también ofrecen restricciones severas en el tratamiento de datos prosódicos porque, por ejemplo, no aceptan caracteres que se utilizan en el etiquetaje con los sistemas ToBI ("%" o "*", por ejemplo). Además, los algoritmos no son suficientemente robustos como para hacer frente a diferencias en variabilidad en la longitud de las etiquetas prosódicas.

3. ¿POR QUÉ CREAR UNA HERRAMIENTA NUEVA PARA EL TRATAMIENTO PROSÓDICO DE LAS DISTANCIAS ENTRE LOS DATOS?

Necesitamos una herramienta que trabaje con los datos prosódicos obtenidos en el marco AMPER y que refleje las especificidades de dicho tipo de datos.

La prosodia, por una parte, se manifiesta, fundamentalmente, en tres parámetros: f_0, duración e intensidad, de forma numérica; respectivamente, en Hz o semitonos, en segundos (o milésimas de segundo) y en

Fernández Planas, A. M., Roseano, P., Elvira-García, W., & Balocco, S. (2019). Génesis y aspectos fundamentales de ProDis. In J. M. Lahoz-Bengoechea, R. Pérez Ramón, & J. Villa Villa (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 125–132). Málaga: Universidad de Málaga.

decibelios. Por otra parte, la prosodia vehicula información sobre la modalidad oracional, el acento léxico, la estructura sintáctica o la manifestación del foco. Finalmente, la prosodia expresa diferencias diafásicas, diastráticas o diatópicas (en terminología coseriana). En este último terreno, AMPER (Martínez-Celdrán y Fernández Planas, 2003–2016), junto con el IARI (Prieto, Borràs-Comes y Roseano, 2010–2014), ha ido conformando una enorme base de datos prosódicos acústicos que es susceptible de ser sometida a estudios dialectométricos a partir de algoritmos que trabajen con datos numéricos.

En el seno de este proyecto hubo una propuesta de herramienta dialectométrica, Stat-Distances (Rilliard y Lai, 2008), que no se acabó de desarrollar del todo y, a pesar de proporcionar resultados interesantes, no ofrecía datos imprescindibles como la matriz de distancias o la explicitación de los algoritmos que usaba. Nuestro primer contacto con la metodología dialectométrica fue de la mano de este programa y nos sirvió para empezar a constatar que nuestros resultados no siempre eran totalmente coincidentes con los obtenidos por la dialectología tradicional, lo cual es plausible ya que la dialectología clásica nunca había tratado este tipo de datos. En seguida Stat-Distances dejó de estar disponible para los investigadores que trabajamos en el proyecto. Entre los diferentes grupos implicados se han sucedido otras propuestas. A saber: un script en R (Martínez Calvo y Fernández Rei, 2015) y el programa que se presenta en este trabajo, que es la versión mejorada de unas rutinas previas que llamamos Calcu-Dista (Roseano, Elvira-García, y Fernández Planas, en revisión; Elvira-García, 2014). La especificidad de los datos prosódicos y la falta de una herramienta útil para operar con ellos nos llevó a proponer nuestra herramienta, a la que llamamos ProDis, procedente de la expresión *Prosodic Distances* (Elvira-García, Balocco, Fernández Planas, Roseano, y Martínez Celdrán, 2015; Fernández Planas, 2016a, 2016b), que es la versión mejorada de Calcu-Dista (Roseano, Fernández Planas, Elvira-García, Cerdà Massó, y Martínez Celdrán, 2015) y también se inspira en Stat-Distances. ProDis funciona, como es habitual en los trabajos en el seno de AMPER, en el entorno MatLab.

Así pues, nuestra herramienta constituye nuestra contribución para construir un programa potente capaz de trabajar en dialectometría a partir de los datos prosódicos numéricos que obtenemos en nuestros análisis en el seno de AMPER. En concreto, responde, en esta primera fase, a las exigencias técnicas que habían surgido dentro del marco del proyecto. En concreto, se necesitaba que nuestra herramienta fuera "amigable" y flexible, que fuera capaz de trabajar con datos numéricos, de solucionar los problemas debidos a las diferencias en número de sílabas entre frases en distintas lenguas en el marco AMPER (por ejemplo, *O pássaro gosta de Renato*, 10 sílabas, *vs. La guitarra se toca con paciencia*, 11 sílabas), de considerar de cada frase 3 repeticiones, de ponderar f_0 por duración, por intensidad o por ambos parámetros a la vez y, finalmente, de trabajar con nuevas lenguas adicionalmente.

4. LA GÉNESIS DE PRODIS: DE CALCU-DISTA A PRODIS

ProDis, como antes Calcu-Dista y antes todavía Stat-Distances, utiliza la fórmula (1), que se inspira en la que propuso Hermes (1998) y que resulta ser una fórmula sencilla para calcular distancias entre datos acústicos numéricos.

$$RMS = \sqrt{\frac{\sum_{i=1}^{N}(f_0x_i - f_0y_i)^2}{N}} \quad (1)$$

El antepasado más cercano de ProDis, conocido como Calcu-Dista, era, más que una herramienta, una rutina para el cálculo de distancias prosódicas a partir de los datos numéricos de las melodías en semitonos. Esa rutina se fundamentaba en tres programas bien conocidos: Praat v. 5.4.01 (Boersma y Weenink, 2014), Excel (Microsoft Office 2007) y SPSS Statistics 20 (IBM). Los programas en cuestión tomaban con punto de partida los datos acústicos previamente procesados por tres programas creados en el seno del Laboratori de Fonètica de la UB y circunscritos al ámbito AMPER: AMPER-Reno, AMPER-Extra y AMPER-Eti (Roseano, 2012).

En primer lugar, un script de Praat creado *ad hoc* extraía, a partir de los archivos txt de cada repetición de las frases proporcionados por AMPER-2006 (López Bobo, Muñiz Cachón, y Díaz Gómez, 2007), los valores de f_0 en semitonos y los colocaba en una matriz de datos comparando cada repetición de una frase en un mismo hablante y entre hablantes distintos considerando tres valores por vocal.

En segundo lugar, un análisis en Excel sobre la salida de Praat aplicaba la fórmula de las distancias escogida. Se escogió como índice de la distancia entonativa entre dos frases, que podemos llamar x e y, la media cuadrática de la diferencia entre los valores de f_0 de la frase x y de la frase y en cada uno de los puntos de medición. Para los dos conjuntos x e y de valores de f_0 $\{f_0x_1, f_0x_2, ..., f_0x_N\}$ y $\{f_0y_1, f_0y_2, ..., f_0y_N\}$, donde N es el número de puntos de medición de f_0 en cada una de las dos frases, mientras que f_0x_i y f_0y_i son los valores de f_0 en semitonos en cada uno esos puntos.

Esta fórmula proporcionaba la distancia entre dos frases con la misma estructura (por ejemplo, entre dos declarativas SVO con sujeto llano, verbo llano y objeto esdrújulo) de dos puntos de encuesta. Para determinar la distancia general entre todas las frases de dos puntos de encuesta, puesto que la distribución de las distancias no es normal, de acuerdo con De Castro Moutinho, Coimbra, Rilliard, y Romano (2011, p. 44) se escogió la mediana de las RMS calculadas por cada pareja de frases x e y. A partir de las medianas de las distancias entre cada par de puntos de encuesta se pudo construir la matriz de distancias correspondiente.

En tercer lugar, la matriz de distancias constituía, a su vez, la base para la fase final del proceso de análisis,

Fernández Planas, A. M., Roseano, P., Elvira-García, W., & Balocco, S. (2019). Génesis y aspectos fundamentales de ProDis. In J. M. Lahoz-Bengoechea, R. Pérez Ramón, & J. Villa Villa (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 125–132). Málaga: Universidad de Málaga

que se efectuaba con SPSS y consistía en lo siguiente. En primer lugar, en un análisis de conglomerados clúster, técnica multivariante cuya finalidad es clasificar los puntos de encuesta en grupos a partir de la semejanza entre sus características entonativas donde como método de comparación se utiliza la media de las distancias entre los grupos tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos. El resultado se expresa en forma de dendrograma que permite ver, en forma de árbol invertido, cómo se agrupan los datos hasta el nivel que se considera oportuno. Y en segundo lugar, un análisis de tipo escalamiento multimdimensional (EMD o MDS) que representa bidimensional o tridimensionalmente de forma gráfica las distancias entre los sujetos o puntos de encuesta de la manera más objetiva posible en un espacio virtual. Este método estadístico pretende construir un espacio métrico con el menor número de dimensiones posibles, de tal manera que permite representar las proximidades o preferencias entre los objetos con el mayor grado de fidelidad. Desde un conjunto de objetos se establecen sus propiedades numéricas a partir de las cuales se elaboran las tablas de proximidad (o de similitud) y, finalmente, se trasladan estas proximidades a un espacio, un mapa de objetos (Matas Crespo, 2006). En realidad, ambos tipos de gráficos —dendrogramas y espacios MDS— proporcionan la misma información, y así se puede comprobar en el apartado de resultados. La ventaja de ambas formas de representación es la de permitir captar la distribución y la agrupación de los datos sin necesidad de tener que recurrir a una matriz de distancias numérica de proporciones enormes.

Como medida utilizamos el intervalo de distancia euclidiana (2).

$$(2) \qquad d(x,y) = \|x - y\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

El funcionamiento de Calcu-Dista, como era esperable, ha sido validado estadísticamente mediante la comparación con resultados obtenidos con métodos comparables (Fernández Planas *et al.*, 2015).

5. CARACTERÍSTICAS FUNDAMENTALES DE PRODIS

La idea de partida era: (1) cómo poder establecer una matriz de distancias por informantes y por puntos de encuesta (reuniendo en un bloque los distintos informantes del mismo punto de encuesta) a partir de una matriz inmensa de datos obtenidos en AMPER-2006 en txt de datos en semitonos; y (2) cómo ver reflejadas las matrices de distancias (o de proximidades) de forma gráfica en forma de dendrogramas y de distribución en espacios virtuales.

ProDis realiza la media y la mediana de correlación por informantes y por punto de encuesta. A continuación, a partir de estos datos realiza un análisis de clúster que permite clasificar los informantes o los puntos de encuesta en diferentes grupos, tanto en forma de dendrograma como de distribución en un espacio virtual, según su semejanza o su proximidad.

Concretamente, a partir de los datos en semitonos de la f_0 de las frases, computa las correlaciones entre ellos, analiza la media y la mediana de las correlaciones para cada hablante, construye la matriz de correlaciones entre hablantes y entre puntos de encuesta, lleva a cabo el análisis de clúster entre hablantes y entre puntos de encuesta y prepara las diferentes salidas gráficas (los dendrogramas y EMD), pero también los gráficos que permiten establecer la validación estadística de los datos como los mapas de correlación, de desviación estándar, los gráficos de silueta o los gráficos de Shepard.

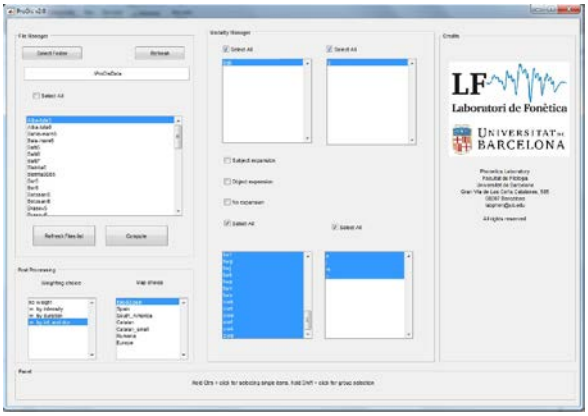
En la Figura 1 aparece una imagen de la interfaz de la herramienta.

Para el cálculo de las correlaciones se utilizan las de Pearson con cuatro métricas diferentes para cada análisis: (1) sin ponderar; (2) ponderadas por la intensidad (a la manera de Hermes, 1998), ponderadas por la duración, y ponderadas por la intensidad y la duración. Existen diferentes métodos para mediar las distancias en dialectometría y la correlación de Pearson (la que hemos estimado mejor) es una de ellas. Se presentó una revisión crítica de ellos en Elvira-García, Balocco, Roseano, y Fernández Planas (2016) y también en Hermes (1998). Serían, entre otros, el algoritmo de Levenstein (Kessler, 1995), la distancia euclidiana (Nerbonne *et al.*, 2011; Roseano *et al.*, 2015); Mahalanobis (Wouters y Macon, 1998), correlación de Spearman (Hermes, 1998), correlación de Pearson (Heeringa y Gooskens, 2003), tau de Kendall (Hermes, 1998).

Para las correlaciones, el programa compara cada repetición de una frase de un hablante de un punto de encuesta con las otras dos repeticiones del mismo informante de la misma frase y con las tres repeticiones de la frase con la misma estructura sintáctica de otro informante. Ello permite obtener una matriz de correlaciones como la que aparece en la Figura 2. Los valores, lógicamente, van de −1 a 1.

Los mismos datos los podemos ver de forma completa y más fácilmente aprehensible en un mapa de correlaciones como el que aparece en la Figura 3. Cabe destacar que en el gráfico en cuestión no aparecen valores numéricos, sino gradaciones de colores que

Figura 1: Interfaz de ProDis.



Fernández Planas, A. M., Roseano, P., Elvira-García, W., & Balocco, S. (2019). Génesis y aspectos fundamentales de ProDis. In J. M. Lahoz-Bengoechea, R. Pérez Ramón, & J. Villa Villa (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 125–132). Málaga: Universidad de Málaga.

Figura 2: Ejemplo de matriz de correlaciones.

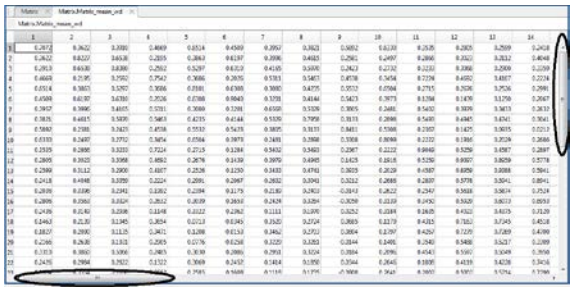


Figura 3: Ejemplo de mapa de correlaciones.

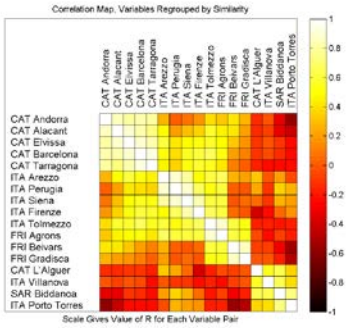
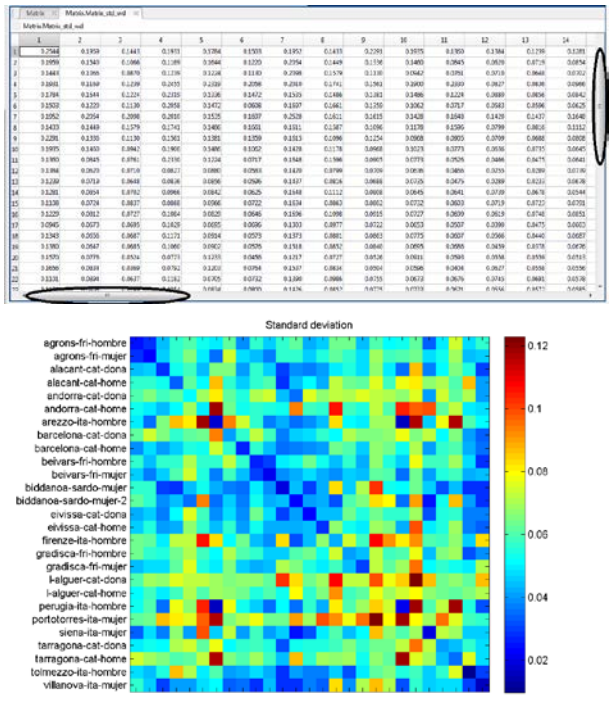


Figura 4: Matriz y mapa de desviación estándar.



indican grados de correlación distintos. De esa manera, la matriz resulta de comprensión más inmediata.

La herramienta ofrece también la matriz y el mapa de desviación estándar por informante y por punto de encuesta (Figura 4), lo cual es muy interesante para comprobar la coherencia intrasujetos y entre sujetos, por un lado, y también intrapuntos de encuesta y entre puntos de encuesta, por otro lado.

Se aprecia fácilmente cómo los puntos más tendentes a rojo (por ejemplo los valores de la voz femenina de L'Alguer en catalán, en el mapa de ejemplo) permiten adivinar que el informante (o el punto) en cuestión es menos homogéneo en los patrones melódicos que utiliza.

ProDis proporciona también gráficos de silueta que, de alguna manera, indican los grupos que matemáticamente sería óptimo establecer en los dendrogramas que agrupan progresivamente los elementos hasta unirlos todos, aunque es la decisión del investigador la que establece el límite de agrupaciones. Para las agrupaciones se toma como referencia el elemento más lejano. Podemos ver un ejemplo de dendrograma en la Figura 5.

El dendrograma funciona con una técnica multivariante que crea clústering aglomerativo jerárquico usando un método de agrupación completo, grupos basados en el elemento más lejano.

El gráfico de Shepard, que se refiere a los gráficos en EMD, nos demuestra cuando tiende a una línea, que los gráficos son fiables y válidos.

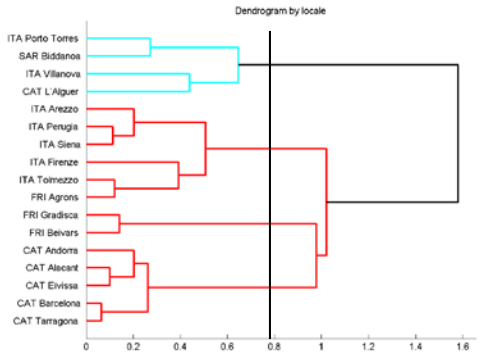
En la Figura 6 vemos dos formas de visualizar los gráficos EMC en ProDis: a) en dos dimensiones; b) en tres dimensiones. Aunque el gráfico en tres dimensiones es más recomendable porque el valor de Stress tiende a ser más bajo que en el gráfico en dos dimensiones, en ocasiones puede ser más difícil de interpretar.

Finalmente, mediante ProDis podemos obtener mapas geográficos donde se representen en colores coincidentes con los dendrogramas y con los EMD las localizaciones de los informantes o de los puntos de encuesta. Véase un ejemplo en la Figura 7.

6. CONCLUSIÓN. MEJORAS DE PRODIS RESPECTO A STAT-DISTANCES Y LÍNEAS DE FUTURO

ProDis satisface con creces la idea de partida que señalábamos en el inicio del aparatado anterior, ya que por un lado permite establecer una matriz de distancias por informantes y por puntos de encuesta a partir de los datos numéricos obtenidos en AMPER-2006 y, por otra parte, transforma las matrices de distancias en gráficos de más fácil interpretación. Además, mejora y supera en prestaciones las que ofrecía Stat-Distances porque: (1) considera repertorios de datos no coincidentes en el

Figura 5: Ejemplo de dendrograma con una línea que establece el límite de agrupaciones que se sería óptimo tener en cuenta.



Fernández Planas, A. M., Roseano, P., Elvira-García, W., & Balocco, S. (2019). Génesis y aspectos fundamentales de ProDis. In J. M. Lahoz-Bengoechea, R. Pérez Ramón, & J. Villa Villa (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 125–132). Málaga: Universidad de Málaga

Figura 6: Gráficos EMC en dos dimensiones (arriba) y en tres dimensiones (abajo).

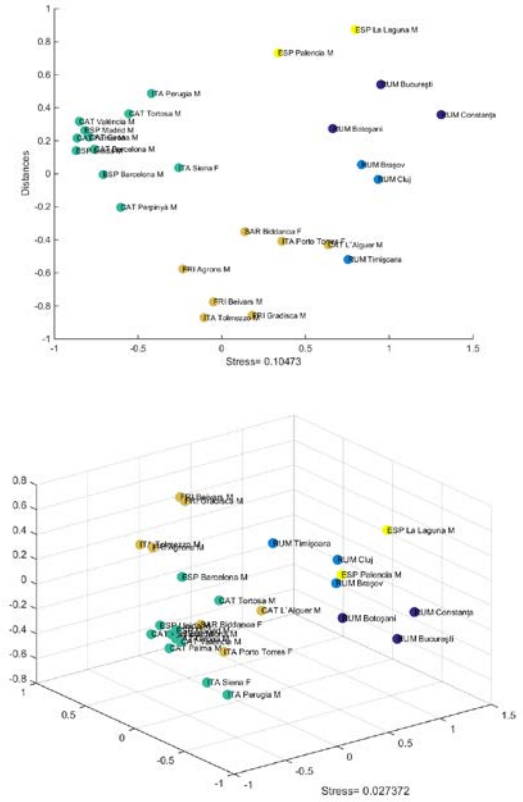
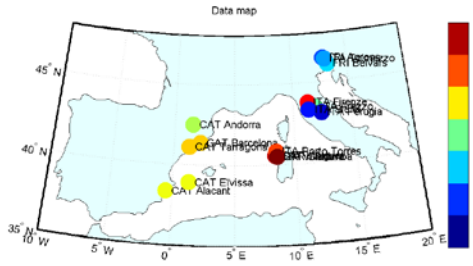


Figura 7: Ejemplo de mapa geográfico con la localización de los puntos de encuesta considerados.



número de sílabas; (2) puede ponderar por duración o por intensidad; (3) permite obtener matrices de confusión y de desviación estándar; (4) es capaz de comparar repertorios con diferente números de frases y diferente composición estructural; (5) puede exportar sus resultados a Excel, SPSS o R; (6) ofrece, entre sus resultados, gráficos de correlación, dendrogramas y MDS; (7) trabaja con semitonos, es decir, con datos normalizados; (8) es relativamente fácil de usar; (9) permite al usuario cierta flexibilidad.

Entre las líneas de desarrollo futuro inmediato, nos proponemos acelerar el software e incorporar EtiToBI (Elvira-García *et al.*, 2016) para llevar a cabo análisis dialectométricos con datos nominales (es decir, a partir de los etiquetajes de los datos en el sistema ToBI).

7. REFERENCIAS

Álvarez Blanco, R., Dubert, F., y Sousa, X. (2006). Aplicación da análise dialectométrica aos datos do Atlas Lingüístico Galego. En R. Álvarez Blanco, F. Dubert, y X. Sousa (Eds.), *Lingua e territorio* (pp. 461–493). Santiago de Compostela: Instituto da Lingua Galega – Consello da Cultura Galega.

Aurrekoetxea, G. (1992). Naffaroako Euskara: azterketa dialektometrikoa. *Utzaro*, *5*, 59–109.

Aurrekoetxea, G., Fernández-Aguirre, K., Rubio, J., Ruiz, B., y Sánchez, J. (2013). DiaTech: A new tool for dialectology. *Literary and Linguistic Computing*, *28*(1), 23–30.

Bauer, R. (2003). Sguardo dialettometrico apoyo alcune zone di transizione dell'Italia norte-orientale (lombardo vs. Trentino vs. Veneto). Parallel X. Sguardi reciprocidad. Vicende linguistiche e cultural dell'area italofona e germanófona. En R. Bombi y F. Fusco (Eds.), *Atti del Decimo Incontro italo-austriaco di linguisti* (pp. 93–119). Udine, Italia: Forum Editrice.

Bauer, R. (2005). La classificazione dialettometrica dei basiletti altoitaliani e Ladino rappresentati nell'Atlante linguistico del ladino dolomitico e dei dialetti limitrofi (ALD - I). En C. Guardiano *et al.* (Eds.), *Lingue, istituzioni, territorio. Riflessioni teoriche, proposte metodologiche ed esperienze di politica linguistica* (347–365). Roma, Italia: Bulzoni.

Boersma, P. y Weenink, D. (2014). Praat: doing phonetics by computer (Versión 5.4.01). Disponible en http://www.praat.org/.

Clua, E. (2004). El método dialectométrico: aplicación del análisis multivariante a la clasificación de las variedades del catalán. En M. P. Perea (Ed.), *Dialectología y recursos informáticos* (pp. 59–88). Barcelona: Universitat de Barcelona.

Contini, M. (1992). Vers une géoprosodie romane. *Actas del Nazioarteko Dialektologia Biltzarra Agiriak, Bilbao 1991* (pp. 83–109). Bilbao: Publicacions de la Real Academia de la Lengua Vasca.

Contini, M., Lai, J. P., y Romano, A. (2002). La géolinguistique à Grenoble: de l'AliR à AMPER. *Revue Belge de Philologie et d'Histoire*, *80*(3), 931–941.

Contini, M., Lai, J. P., Romano, A., y Roullet, S. (1998). Vers un atlas prosodique parlant des variétés romanes. En J.-C. Bouvier *et al.* (Eds.), *Mélanges offerts à X. Ravier* (pp. 73–84). Toulouse: Université de Toulouse – Le Mirail.

Contini, M., Lai, J. P., Romano, A., Roullet, S., De Castro Moutinho, L., Coimbra, R. L., Pereira Bendiha, U., y Secca Ruivo, S. (2002). Un projet d'atlas multimédia prosodique de l'espace roman. *Proceedings of the International Conference Speech Prosody 2002* (pp. 227–230). Aix-en-Provence: Laboratoire Parole et Langage.

D'Andrés Díaz, R., Álvarez-Balbuena García, F., y Suárez Fernández, X. M. (2007). Proxecto ETLEN para o estudio dialectográfico e dialectométrico da zona Eo-Navia, Asturias: fundamentos teóricos. Actas VII Congreso Internacional de Estudos Galegos: mulleres en Galicia: Galicia e os outros pobas da península (pp. 749–759). A Coruña: Ediciós do Castro.

De Castro Moutinho, L., Coimbra, R. L., Rilliard, A., y Romano, A. (2011). Mesure de la variation prosodique diatopique en portugais européen. *Estudios de Fonética Experimental*, *20*, 33–55.

Elvira-García, W. (2014). Calcu-Dista scripts package. Praat script. Disponible en http://stel.ub.edu/labfon/en/praat-scripts

Elvira-García, W., Balocco, S., Fernández Planas, A. M., Roseano, P.. Martínez Celdran, E. (2015). Presentació d'una aplicació informàtica per a l'anàlisi dialectomètrica de dades prosòdiques en el marc de l'Atles Multimèdia de la Prosòdia de l'Espai Romànic. Presentado en el VII Workshop sobre prosodia del catalán, Universitat de Barcelona, 22/06/2015.

Elvira-García, W., Balocco, S., Roseano, P., y Fernández Planas, A. M. (2016). Comparació de mesures de distància prosòdica entre varietats dialectals. Presentado en el VIII Workshop sobre prosodia del catalán, Universitat Pompeu Fabra, 04/07/2016.

Elvira-García, W., Roseano, P., Fernández Planas A. M. y Martínez Celdrán E. (2016). A tool for automatic transcription of intonation: Eti-ToBI a ToBI transcriber for Spanish and Catalan. *Language Resources and Evaluation*, *50*(4), 767–792.

Fernández Planas, A. M. (2005). Datos generales del proyecto AMPER en España. *Estudios de Fonética Experimental*, *14*, 13–27.

Fernández Planas, A. M. (2016a). Aspectos de ProDis, una nueva herramienta para el análisis dialectométrico prosódico. Presentado en el Workshop «Approaches to Sociolinguistic Aspects of Romanian and Spanish Intonation», Alexandru Ioan Cuza University of Iasi (Rumanía), 21/10/2016.

Fernández Planas, A. M. (2016b). Características generales de ProDis (herramienta para analizar distancias prosódicas). Presentado en el Servei de Tractament de la Parla i el So (STPS) de la Universitat Autònoma de Barcelona, 18/11/2016.

Fernández Planas, A. M., Dorta, J., Roseano, P., Díaz, X., Elvira-García, W., Martín Gómez, J. A., Martínez Celdrán, E. (2015). Distancia y proximidad prosódica entre algunas variedades del español: un estudio dialectométrico a partir de datos acústicos. *Revista de Lingüística Teórica y Aplicada*, *53*(2), 13–45.

Fernández Planas, A. M., Roseano, P., Martínez Celdrán, E., y Romera Barrios, L. (2011). Aproximación al análisis dialectométrico de la entonación en algunos puntos del dominio lingüístico catalán. *Estudios de Fonética Experimental*, *20*, 141–178.

Goebl, H. (1981). Eléments d'analyse dialectométrique (avec application à l'AIS). *Revue de Linguistique Romane*, *45*, 349–420.

Goebl, H. (1987). Encore un coup d'oeil dialectométrique sur las Tableaux phonétiques de patois suisses romands (TPPSR). *Vox Romanica*, *46*, 91–125.

Goebl, H. (1993). Dialectometry: A short overview of the principles and practice of quantitative classification of linguistic atlas data. En R. Köhler y B. B. Rieger (Eds.), *Contributions to quantitative linguistics* (pp. 277–315). Dordrecht: Springer.

Goebl, H. (2003). Regards dialectométriques sur les données de l'Atlas linguistique de la France (ALF): relations quantitatives et structures de profondeur. *Estudis Romànics*, *25*, 60–117.

Goebl, H. y Haimerl, E. (2004). Visual Dialectometry. http://www.dialectometry.com/dmdocs/index.html [28/11/2016].

Goebl, H. y Schiltz, G. (1997). Dialectometrical compilation of CLAE 1 and CLAE 2. Isoglosses and dialect integration. En W. Viereck, H. Ramisch, H. Händler, y C. Marx (Eds.), *The computer developed linguistic Atlas of England*, *Vol. 2* (pp. 13–21). Tübingen: Niemeyer.

Heeringa, W. y Gooskens, C. (2003). Norwegian dialects examined perceptually and acoustically. *Computers and the Humanities*, *37*(3), 293–315.

Heeringa, W. y Nerbonne, J. (2001). Dialect areas and dialect continua. *Language Variation and Change*, *13*, 375–400.

Hermes, D. J. (1998). Measuring the perceptual similarity of pitch contours. *Journal of Speech Language and Hearing Research*, *41*(1), 73–82.

Kessler, B. (1995). Computational dialectology in Irish Gaelic. En *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 60–66).

López Bobo, M. J., Muñiz Cachón, C., Díaz Gómez, L., Corral Blanco, N., Brezmes Alonso, D., y Alvarellos Pedrero, M. (2007). Análisis y representación de la entonación. Replanteamiento metodológico en el marco del proyecto AMPER. En J. Dorta (Ed.), *La prosodia en el ámbito lingüístico románico* (pp. 17–34). Santa Cruz de Tenerife: La Página Ediciones.

Martínez Calvo, A., y Fernández Rei, E. (2015). Unha ferramenta informática para a análise dialectométrica da prosodia. *Estudios de Fonética Experimental*, *24*, 289–303.

Martínez Celdrán, E. y Fernández Planas, A. M. (Coords.) (2003–2016). *Atlas Multimèdia de la Prosòdia de l'Espai Romànic*. http://stel.ub.edu/labfon/amper/cast/index_ampercat.html

Matas Crespo, J. (2006). La técnica del Escalamiento Multidimensional en el vocalismo: un análisis comparativo (Tesis Doctoral). Universitat de Barcelona.

Fernández Planas, A. M., Roseano, P., Elvira-García, W., & Balocco, S. (2019). Génesis y aspectos fundamentales de ProDis. In J. M. Lahoz-Bengoechea, R. Pérez Ramón, & J. Villa Villa (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 125–132). Málaga: Universidad de Málaga

Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P., y Leinonen, T. (2011). Gabmap: A web application for dialectology. *Dialectologia*. Special issue II, 65–89.

Polanco, L. (1992). Lengua y dialecto: una aplicación dialectométrica a la lengua catalana. *Miscelánea, 3*, 5–28.

Prieto, P., Borràs-Comes, J., y Roseano, P. (Coords.) (2010–2014). Interactive Atlas of Romance Intonation. http://prosodia.upf.edu/iari/.

Rilliard, A. y Lai, J. P. (2008). Outils pour le calcul et la comparaison prosodique dans le cadre du projet AMPER: L'exemple des variétés Occitane et Sarde. En A. Turculeţ (Ed.), *La variation diatopique de l'intonation dans le domaine roumain et roman* (pp. 217–229). Iaşi, Rumanía: Editura Universităţii Al. I. Cuza.

Romano, A. (2003). Un projet d'Atlas multimédia prosodique de l'espace roman (AMPER). En F. Sánchez Miret (Ed.), *Atti del XXIII CILFR, Vol. 1* (pp. 279–294). Tübingen: Niemeyer.

Romano, A. y Contini, M. (2001). Un progetto di Atlante geoprosodico multimediale delle varietà linguistiche romanze. En E. Magno Caldognettoy P. Cosi (Eds.), *Multimodalità e Multimedialità nella Comunicazione. Atti delle XI Giornate di Studio del "Gruppo di Fonetica Sperimentale" dell'Associazione Italiana di Acustica* (pp. 121–126). Padova: Unipress.

Roseano, P. (2012). La prosòdia del friulà en el marc de l'Atles Multimèdia de Prosòdia de l'Espai Romànic (Tesis Doctoral). Universitat de Barcelona.

Roseano, P., Elvira-García, W., y Fernández Planas, A. M. (en revisión). Calcu-Dista: A Tool for Dialectometric Analysis of Intonational Variation. En I. Feldhausen, M. M. Vanrell y U. Reich (Eds.), *Empirical Methods in Romance Prosody Research.* Language Science Press.

Roseano, P., Fernández Planas, A. M., Elvira-García, W., Cerdà Massó, R., y Martínez Celdrán, E. (2015). Contacto lingüístico y transferencia prosódica: El caso del alguerés. *Dialectologia et Geolinguistica*, *23*(1), 95–123.

Saramago, J. (2002). Diferenciação lexical interpontual nos territórios galego e português (Estudo dialectométrico aplicado a materiais galegos do ALGa). En R. Álvarez, F. Dubert García, y X. Sousa Fernández (Eds.), *Dialectoloxía e léxico* (pp. 41–68). Santiago de Compostela: Instituto da Lingua Galega – Consello da Cultura Galega.

Séguy, J. (1973). La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de Linguistique Romane*, *37*, 1–24.

Sousa, X. (2006). Análise dialectométrica das variedades xeolingüística galegas. En M. C. Rolão Bernardo y H. Mateus Montenegro (Eds.), *Actas do I Encontro de Estudos Dialectológicos* (pp. 345–362). Ponta Delgada, Portugal: Instituto Cultural de Ponta Delgada.

Verlinde, S. (1988). La dialectométrie et la détection des zones dialectales: L'architecture dialectale de l'Est de la Belgique romane. *Revue de Linguistique Romane*, *51*, 151–172.

Wouters, J., y Macon, M. W. (1998). A perceptual evaluation of distance measures for concatenative speech synthesis. En *Proceedings of the Fifth International Conference on Spoken Language Processing*.

Fernández Planas, A. M., Roseano, P., Elvira-García, W., & Balocco, S. (2019). Génesis y aspectos fundamentales de ProDis. In J. M. Lahoz-Bengoechea, R. Pérez Ramón, & J. Villa Villa (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 125–132). Málaga: Universidad de Málaga.

Subsidia: Tools and Resources for Speech Sciences

FonetiToBI, una herramienta para la anotación prosódica automática de corpus

Wendy Elvira-García[1], Juan María Garrido[1]

[1] Universidad de Barcelona
e-mail: wendyelvira@ub.edu, juanmaria.garrido@ub.edu

Citation / Cómo citar esta publicación: Elvira-García, W. & Garrido, J. M.. (2019). FonetiToBI, una herramienta para la anotación prosódica automática de corpus. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 133–142). Málaga: Universidad de Málaga.

RESUMEN: El objetivo de esta comunicación es presentar FonetiToBI, una herramienta basada en Praat para la anotación prosódica automática de enunciados en español y catalán en el marco del modelo métrico autosegmental. Se describen brevemente sus funcionalidades y su estructura, y se explican de forma general las reglas, basadas en conocimiento lingüístico-fonético, que se incluyen en los módulos que la componen. Se presenta también la evaluación realizada con enunciados del español y del catalán, los resultados de la cual muestran que se trata de una herramienta con un grado fiabilidad suficiente para su uso en tareas de anotación automática de corpus.

Palabras clave: Anotación prosódica, corpus, modelo autosegmental, Sp_ToBI, Cat_ToBI

ABSTRACT: The paper aims at describing FonetiToBI, a Praat-based tool for the automatic prosodic annotation of Spanish and Catalan utterances within the Autosegmental Metrical model. It describes briefly the structure of the script and its functionalities. Moreover, an overview of the knowledge-based linguistic-phonetic rules making up its modules is also presented. Finally, the evaluation of the tool carried out with utterances of Spanish and Catalan is detailed. The results of the evaluation show that the tool is reliable enough to be used in tasks of automatic annotation of corpora.

Keywords: Prosodic annotation, corpora, Autosegmental-Metrical Model, Sp_ToBI, Cat_ToBI

1. INTRODUCCIÓN

El objetivo de este trabajo es presentar FonetiToBI, una herramienta diseñada para la anotación prosódica automática de habla para el español y el catalán en el marco del Modelo Métrico Autosegmental (AM) (Pierrehumbert, 1980).

En los últimos años, se ha producido un auge en el desarrollo de herramientas para la anotación automática de corpus de habla, tanto a nivel segmental como suprasegmental. En el caso de la anotación segmental, herramientas como MAUS (Schiel, 1999), EasyAlign (Goldman, 2011) o SPPAS (Bigi, 2012) ofrecen resultados aceptables, aunque en diferente grado según la lengua y la herramienta. En el caso de la anotación suprasegmental, existen también desde hace años herramientas de segmentación automática de las unidades prosódicas (SPPAS, Bigi, 2012; SegProso, Garrido, 2013) y de la entonación (MoMel, Hirst y Espesser, 1993; MelAn, Garrido, 2010). En el marco del Modelo AM, existen también algunas herramientas, como AuToBI (Rosenberg, 2010), que permite la anotación automática de la entonación de enunciados del inglés, o Eti_ToBI (Elvira-García, Roseano y Fernández Planas, 2015; Elvira-García, Roseano, Fernández Planas, y Martínez Celdrán, 2016), diseñado para la anotación de enunciados en español y catalán.

Por otro lado, también está habiendo un genuino interés en definir sistemas de transcripción prosódica más objetivos y basados en los fenómenos fonéticos (Roseano y Fernández Planas, 2015; Hualde y Prieto, 2016).

La herramienta presentada aquí permite la anotación automática completa de los enunciados en el marco de las convenciones ToBI (*Tones and Break Indices*). Es el resultado de la integración de dos herramientas existentes previamente, aunque modificadas para el desarrollo de esta aplicación, SegProso y EtiToBI, la primera orientada a la anotación automática de unidades entonativas, y la segunda a la anotación de los eventos tonales. Las dos tienen en común que han sido desarrolladas, a diferencia de otras herramientas, implementando el conocimiento fonético y lingüístico necesario en forma de reglas.

Figura 1: Estructura y esquema de funcionamiento de FonetiToBI.

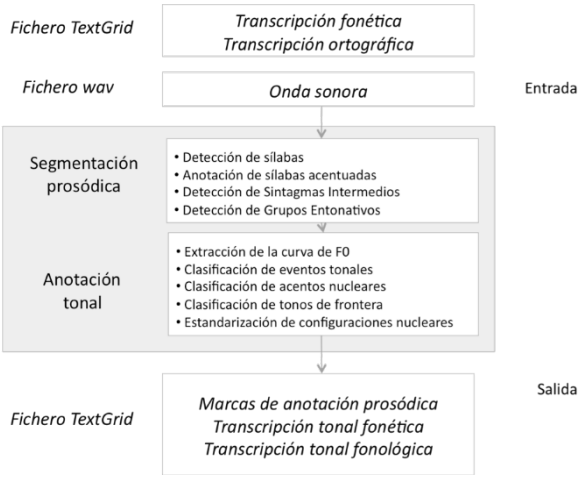
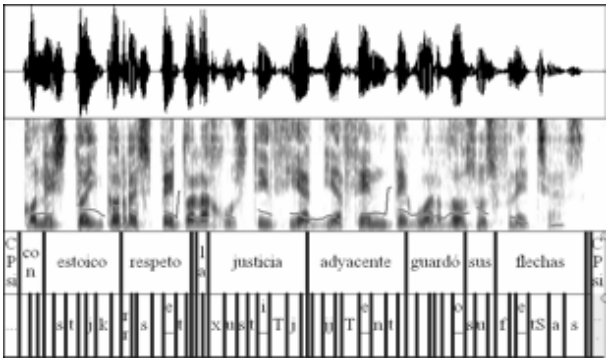


Figura 2: Ejemplo de TextGrid de entrada de FonetiToBI.



2. DESCRIPCIÓN DE LA HERRAMIENTA

FonetiToBI es una herramienta basada en Praat (Boersma, 2001) que permite obtener automáticamente una anotación ToBI completa (tonos y límites prosódicos) a partir de una transcripción ortográfica y fonética de los enunciados de entrada. La Figura 1 presenta su estructura y flujo de funcionamiento.

FonetiToBI está compuesto de dos módulos principales: el primero es una versión modificada de SegProso, y se encarga de identificar los límites prosódicos (*Break Indices*, BI); el segundo, EtiToBI, se encarga de la anotación de los tonos.

2.1. Entrada

FonetiToBI necesita como entrada dos ficheros: un fichero wav con la onda sonora, y un fichero TextGrid de Praat con la transcripción ortográfica y fonética del enunciado, alineada temporalmente con la señal de habla. El TextGrid debe contener, pues, al menos dos capas (*tiers*), una que contenga la transcripción ortográfica, palabra a palabra, y otra con la transcripción fonética, alófono a alófono, ambas alineadas temporalmente con la señal de habla. La Figura 2 presenta un ejemplo de TextGrid de entrada.

FonetiToBI admite como entrada transcripciones fonéticas realizadas utilizando tanto el alfabeto IPA (IPA, en línea) como SAMPA (Wells, 1995). Esto permite, por ejemplo, emplear como entrada ficheros TextGrid generados automáticamente con alguna herramienta de segmentación automática, como SPPAS, que generan la transcripción fonética con símbolos SAMPA.

Otros parámetros de entrada que deben especificarse al inicio son la versión de ToBI que se empleará para la anotación tonal (Sp_ToBI, para la anotación de enunciados en español, Prieto y Hualde, 2015; o Cat_ToBI, para la anotación del catalán, Prieto y Cabré, 2013), o si se llevará a cabo una revisión manual de la anotación al final del proceso.

2.2. Segmentación prosódica

El módulo de segmentación prosódica se encarga de generar una capa de segmentación prosódica (*Break Index Tier*), que contiene las marcas correspondientes a los diferentes tipos de límites prosódicos contemplados en sistema ToBI para el español: grupo clítico (0), palabra (1), sintagma intermedio (3) y grupo o frase entonativa (4). La identificación de estos límites en el enunciado de entrada se lleva a cabo, a partir de la transcripción fonética y ortográfica del enunciado, en tres fases: la primera, de identificación de las sílabas tónicas; la segunda, de identificación de los grupos entonativos; y finalmente, la identificación de los límites de sintagma intermedio. Después, en la última fase de procesado, la salida de estos tres pasos se integra y modifica para obtener una representación de la anotación prosódica en un formato compatible con las convenciones ToBI. Aunque no la única, esta ha sido la principal modificación llevada a cabo en SegProso para su integración en FonetiToBI.

2.2.1. Detección de las sílabas tónicas

La detección de las sílabas tónicas es un paso previo necesario para el etiquetado ToBI que se lleva a cabo en la segunda fase, pero se requiere también para la identificación de los límites de palabra fonológica (los etiquetados con '1' según las convenciones ToBI). Para su identificación, es necesario llevar a cabo previamente una agrupación de los alófonos en sílabas, que también es necesaria para la anotación de los tonos.

La detección de los límites silábicos se lleva a cabo a partir del análisis de la segmentación en palabras y alófonos proporcionada en el TextGrid de entrada, mediante la aplicación de una serie de reglas implementadas en un script de Praat cuyo flujo de funcionamiento puede resumirse de la siguiente manera:

- Se localizan en primer lugar los límites de palabra en la capa que contiene la segmentación ortográfica. Se asume que estos límites actúan como una barrera para la agrupación de los alófonos en sílabas, por lo que no se contemplan los fenómenos de resilabificación entre palabras.
- Se buscan los alófonos que representan núcleos silábicos en la capa de transcripción fonética. El procedimiento encargado de esta tarea comprueba si los alófonos se encuentran en la lista implementada de 'alófonos nucleares' (que contenía inicialmente solo las vocales del español y el catalán, pero que ha

sido ampliada posteriormente con vocales de otras lenguas, como el portugués de Brasil o el francés; Silva y Garrido, 2016).

- Se buscan los límites de la sílaba correspondiente a cada núcleo silábico detectado. Las reglas de silabificación que se aplican analizan los alófonos que aparecen en la transcripción entre el núcleo de la sílaba actual y el siguiente, y establecen el límite de sílaba al final del alófono identificado como final de la coda. Como ya se ha explicado, el límite de palabra se considera una barrera para la silabificación, por lo que cuando se alcanza un límite de palabra se asigna automáticamente un límite de sílaba en esa posición.

Finalmente, una vez establecidos los límites de la sílaba, si el alófono identificado como núcleo silábico lleva una marca de acento, la sílaba se anota como tónica. Es muy importante, por tanto, que la transcripción fonética de entrada incluya marcas de acento para el correcto funcionamiento de la herramienta.

2.2.2. Detección de sintagmas intermedios

Los sintagmas intermedios (*intermediate phrases*, ip) son, en el modelo AM, aquellas unidades entonativas cuyo límite no se marca con una pausa, sino únicamente con un tono de frontera y un alargamiento de la sílaba final de unidad. Su detección automática en la onda sonora de los enunciados debe orientarse, por tanto, a la identificación de estos dos indicios acústicos (movimientos tonales indicativos de la presencia de un tono de límite y alargamientos silábicos). El límite final de estas unidades se marca en la capa de segmentación prosódica con un '3'.

Las reglas implementadas en FonetiToBI para la detección de los límites de ip están orientadas únicamente a la detección de posibles tonos de frontera al final de las palabras tónicas. Requiere, por tanto, la segmentación en palabras proporcionada en el TextGrid de entrada y la identificación de las palabras tónicas llevada a cabo en la fase anterior. Dichas reglas intentan detectar dos tipos de indicios acústicos relacionados con la presencia de tonos de frontera: por un lado, la presencia en la curva de f_0 de determinados movimientos ascendentes indicativos de la existencia de un tono de frontera intermedia; y por otro, la presencia de reajustes de f_0, indicadores del inicio de una nueva unidad entonativa.

- Las reglas de identificación de tonos de frontera intermedia tratan de localizar movimientos ascendentes de f_0 con una pendiente lo suficientemente pronunciada como para ser interpretadas perceptivamente como tonos de frontera. Básicamente, estas reglas comparan el valor de f_0 en el centro de cada sílaba tónica con el valor de f_0 en la última sílaba de la palabra, o, si la sílaba tónica es la última, al final de la misma. Si la diferencia entre los dos valores supera un umbral (actualmente establecido en el 5 % del valor de f_0 en el centro de la sílaba tónica), se considera que existe un límite de frase intermedia al final de esa palabra. La Figura 3 presenta un ejemplo de límite detectado con este tipo de reglas.

Figura 3: Onda Sonora, espectrograma, curva de f_0 y anotación correspondiente al enunciado del catalán *'Alt, fort, i amb expressió salvatge'*, pronunciado por un hablante femenino. El intervalo seleccionado fue marcado como ip por las reglas de identificación de tonos de frontera.

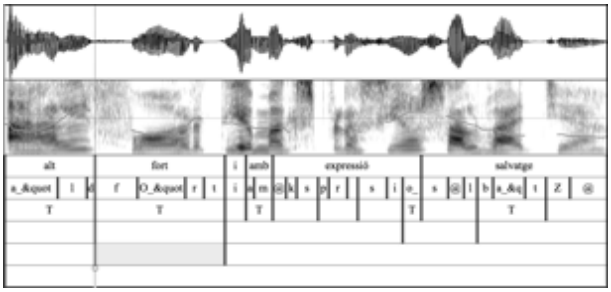
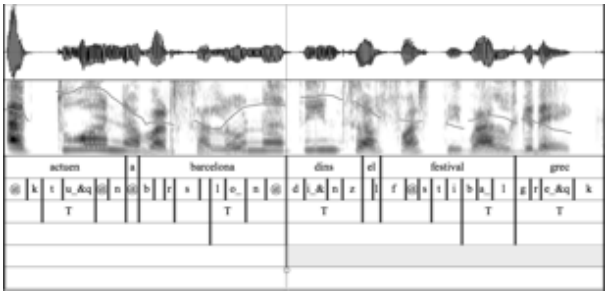


Figura 4: Onda Sonora, espectrograma, curva de f_0 y anotación correspondiente al enunciado del catalán *'Actuen a Barcelona dins el festival Grec'*, pronunciado por un hablante masculino. El intervalo seleccionado fue marcado como ip por las reglas de reajuste de f_0 .



- Las reglas de reajuste de f_0 tratan de identificar diferencias significativas de f_0 entre sílabas tónicas consecutivas: se miden los valores de f_0 en el centro de ambas sílabas, y si el valor de la segunda resulta ser significativamente mayor que el de la primera (la diferencia entre ambas debe ser al menos un 5 % del valor de f_0 en la primera tónica), se establece un límite de ip entre las dos palabras. La Figura 4 presenta un ejemplo de límite correspondiente a este segundo tipo.

2.2.3. Detección de grupos entonativos

Los grupos entonativos (*IntonationalPhrases*, IP) son unidades prosódicas marcadas por un tono de límite y una pausa. Se marcan mediante un '4' en la capa de segmentación prosódica.

Las reglas encargadas en FonetiToBI de la detección de límites de IP asumen una equivalencia plena entre IP y grupo fónico, un supuesto que, como es bien sabido, es cierto en la mayoría de los casos, pero no en todos. Dicho de otra manera, las reglas asumen que existe un límite de IP siempre que se produce una pausa silenciosa (y por tanto se cierra un grupo fónico), y no comprueban si antes de la pausa existe efectivamente un tono de frontera.

El procedimiento de detección de estos límites en FonetiToBI es, pues, sencillo: simplemente se localizan en la capa de la segmentación fonética de entrada los

segmentos etiquetados como silencio, y se asigna una marca de límite al final de la palabra que lo precede. Es muy importante, por tanto, que la anotación de los silencios sea lo más correcta posible en el TextGrid de entrada.

2.2.4. Conversión a formato ToBI

En su versión anterior, SegProso generaba salida diferentes capas (*tiers)*, cada una con la anotación correspondiente a una unidad prosódica diferente. En las convenciones ToBI, sin embargo, esta información se muestra en un solo tier (*Break Index tier*). La última fase del proceso de segmentación prosódica consiste entonces en convertir la salida de SegProso, en diferentes *tiers*, a un formato compatible con las convenciones ToBI, en un solo *tier*. De la salida original de SegProso, se mantiene también el *tier* correspondiente a la segmentación silábica, con indicación de las sílabas tónicas (etiqueta 'T')

2.3. Anotación tonal

El módulo de anotación tonal se encarga de generar una capa de tonos (*Tones*), que contiene las marcas correspondientes a los movimientos tonales de la curva de f_0 tal y como se explicitan en las convenciones para acentos tonales y tonos de frontera de los sistemas de transcripción Cat_ToBI (Prieto y Cabré, 2013) y Sp_ToBI (Hualde y Prieto, 2016).

La transcripción prosódica tiene dos niveles. El primero, que llamaremos transcripción fonética estrecha, es una transliteración de los movimientos tonales de f_0 en términos AM, es decir en L y H. Esta transcripción sigue las convenciones de transcripción prosódica fonética estrecha expuestas en Martínez Celdrán y Fernández Planas (2003) y Roseano y Fernández Planas (2013). En este nivel, el procedimiento de anotación empleado puede generar en algún caso etiquetas no totalmente acordes con las convenciones estándar ToBI, como son los tonos tritonales. El segundo nivel es una transcripción fonética ancha en la que los movimientos de la transcripción fonética estrecha se simplifican para dejar solo aquellos susceptibles de ser tonos fonológicos, es decir, los recogidos en las versiones actuales de Sp_ToBI y Cat_ToBI.

La clasificación delos movimientos tonales se lleva a cabo a partir de:

- la segmentación silábica;
- las marcas de segmentación prosódica (*Break Indices*);
- las marcas de acento léxico.

Esta información es proporcionada por el módulo de segmentación descrito en el apartado 2.2.

2.3.1. Extracción de la curva de f_0

Además de los datos provenientes de la salida del módulo de segmentación, para poder transcribir la entonación el script necesita los datos de f_0. Estos datos se extraen del sonido proporcionado en el input usando el método de autocorrelación propuesto por Boersma (1993).

Los valores mínimo y máximo de f_0 utilizados como parámetros por el método de extracción de f_0 (*pitch floor* y *pitch ceiling*) se calculan para cada IP también automáticamente. Para su obtención se aplica el método en dos pasos expuesto en Hirst (2011), que se basa a su vez en las investigaciones presentadas en De Looze (2010). Este método extrae un primer objeto 'Pitch' con un rango amplio, para después buscar el f_0 mínimo y máximo en ese objeto y crear un objeto nuevo con el valor mínimo (que se establece multiplicando por 0.75 el primer cuartil del rango anterior) y el máximo (obtenido multiplicando por 1.5 el tercer cuartil del rango).

El objeto 'Pitch' resultante es el que se usa para detectar los movimientos tonales y clasificarlos, es decir para realizar la transcripción tonal. Por lo tanto, si dicha curva contiene errores de detección de f_0, la transcripción resultante será errónea. Por este motivo, el transcriptor funciona mejor con grabaciones de calidad alta, un alto porcentaje de segmentos sonoros y pocos sonidos o elementos que puedan causar errores o problemas de detección en el contorno de f_0, como serían fricativas sibilantes o barras de explosión.

2.3.2. Clasificación de eventos tonales

La teoría AM prevé que los movimientos tonales estén asociados a algún elemento prominente en el enunciado (sílabas tónicas o fronteras prosódicas). Por lo tanto, el módulo usa la información contenida en el TextGrid para determinar la posibilidad de prominencia, y, a partir de ahí, analiza los movimientos de f_0 para encontrar eventos tonales.

El script considera que ha habido un movimiento tonal entre dos momentos del enunciado (por ejemplo, entre dos sílabas contiguas o entre el inicio de la sílaba tónica y el final) cuando la diferencia entre esos dos momentos sobrepasa los 1.5 semitonos. El uso de este umbral está avalado por trabajos que muestran que las diferencias menores de 1.5 semitonos no se usan con significado lingüístico (Pamies, Fernández Planas, Martínez Celdrán, Ortega-Escandell, & Amorós Céspedes, 2002; Rietvelt y Gussenhoven, 1985). Si existe un movimiento que sobrepasa los 1.5 semitonos alrededor de una sílaba prominente, el script detecta que ha habido un evento tonal y procede a clasificarlo.

Los eventos tonales en ToBI se clasifican en:

- acentos prenucleares, es decir, que no sea el último acento de grupo entonativo;
- acentos nucleares, el último de grupo entonativo;
- tonos de frontera.

Dado que los eventos tonales posibles en cada uno de estos casos son diferentes, las fórmulas que se tienen que aplicar para clasificarlos también lo son.

2.3.3. Acentos prenucleares

En el caso de los acentos prenucleares el script busca tres tipos de movimientos (diferencias) distintos:

- diferencias entre el centro de la sílaba pretónica, tónica y postónicas;
- diferencias entre el inicio y final de la tónica;
- diferencias entre el pico y el valle más próximos.

Como se ha explicado, las diferencias entre estos puntos se calculan en semitonos. Para ello, se sigue la fórmula propuesta por Noteboom (1997).

Si no hay ningún movimiento significativo (que sobrepase el umbral de 1.5 semitonos) entre ninguno de estos puntos, el script asigna un acento prenuclear monotonal. Este puede ser L* o H*. Si el acento es el primero de la frase, el nivel (L o H) se determinará calculando el rango del hablante en ese IP. Si la frecuencia de la sílaba que se está analizando es superior al 66 % del rango, el acento tonal (*pitch accent)* se etiqueta como H*; si no alcanza ese nivel, como L*. En el caso de que no sea el primer acento tonal de la IP, el script busca el acento anterior para determinar el tono dependiendo de esta etiqueta. Por ejemplo, si no hay movimiento tonal tras un acento L+H* o H*, el script etiquetará el tono como H*, pero si el tono anterior era L* o H+L* se etiquetará como L*.

En el caso de que haya un movimiento tonal el script aplica una serie de reglas (condiciones *if-then*) que determinan la transcripción del movimiento. En la actualidad el script contiene 64 reglas y 26 subreglas que permiten transcribir en español y en catalán, incluyendo los dialectos que han sido descritos para para el catalán en Prieto y Cabré (2013) y para el español en Prieto y Roseano (2010).

El funcionamiento del script se entiende mejor mediante un ejemplo. Imaginemos un movimiento tonal como el que aparece en la Figura 5, en la que cada recuadro representa una sílaba y el recuadro sombreado representa la sílaba tónica.

El script encontraría un movimiento significativo entre la pretónica y la tónica y adjudicaría el tono L*+H. Como la diferencia en la tónica también es significativa, sustituiría ese tono por L+H*. Después buscaría un movimiento entre tónica y postónica, y encontraría un descenso; por lo tanto, como los dos movimientos son significativos, adjudicaría un nuevo tono, con lo que quedaría un tono tritonal L+H*+L. En el caso de que aparezca un movimiento tritonal (como este), el script comprueba cuál de los dos movimientos que lo conforman es de menor rango y lo coloca entre paréntesis. Por lo tanto la transcripción fonética estrecha del tono sería L+(H*+L). Pero esta etiqueta no está recogida en el sistema Sp_ToBI[1], por lo que en la transcripción fonética ancha el movimiento de menor rango (el descenso después de la tónica) se simplificaría y pasaría a ser L+H*. Sin embargo, queda un último grupo de reglas que aplicar, las que calculan la diferencia entre pico y valle y la posición del pico. Al aplicar estas reglas, como el pico está en la postónica, la etiqueta volvería a sustituirse y se transcribiría como

Figura 5: Esquema de un movimiento tonal modelo en el pretonema. Cada recuadro representa una sílaba, y la tónica aparece sombreada.

L+<H*. En resumen, el script detectaría mediante una serie de reglas que el movimiento es un ascenso en la sílaba tónica con el pico en la postónica y lo transcribiría en consecuencia: en su transcripción fonética estrecha L+H*+L y en la transcripción fonética ancha L+<H*.

En el caso de los acentos descendentes, las reglas más complejas tienen que ver con el descuento de la declinación. Muchos de los movimientos descendentes que se pueden observar en una curva de f_0 no están causados por la existencia de una diana tonal baja en la curva, sino por la declinación. El módulo incluye un conjunto de reglas para evitar estos "falsos" tonos descendentes en la transcripción fonética ancha. En el prenúcleo, el problema radica en que algunos movimientos descendentes son en realidad sílabas tónicas desacentuadas. Para resolver este problema, cuando el módulo encuentra un acento descendente, exige que se cumpla además una de las dos condiciones siguientes: (1) que haya una diana tonal alta en la pretónica, es decir, que la pretónica no sea también descendente; (2) que haya un *plateau* –una 'meseta' alta– en las sílabas anteriores. Este supuesto se puede comprobar calculando si desde la sílaba anterior a la sílaba en la que se encuentra la tónica analizada no ha habido un descenso de más de 1.5 semitonos.

Además de estos movimientos, el módulo de transcripción fonética también etiqueta los casos en los que el rango aumenta considerablemente. Así, se han considerado como movimiento tonal superalto (¡H) los casos en que la diferencia entre dos puntos supera los 6 semitonos. Sin embargo, las cuestiones de rango son fonológicas en español y catalán en contadas ocasiones, por lo que en la capa fonética ancha el símbolo que indica nivel superalto desaparece la mayoría de veces. Se mantiene en el caso de L+¡H*, ya que está demostrado que contrasta fonológicamente con L+H* (Borràs-Comes, Vanrell, y Prieto, 2014).

Por otro lado, la etiqueta ¡ y, más concretamente, ¡H* también se utiliza en la transcripción para marcar que un movimiento tonal es un ascenso desde un punto ya alto, es decir, para transcribir ascensos desde una meseta alta a un punto aún más alto, como sería el caso de las interrogativas informativas en canarias (Vizcaíno Ortega, Cabrera Abreu, Dorta, & Hernández Díaz, 2007).

Los símbolos ¡ y ! no se usan en este transcriptor para marcar casos de escalonamiento tonal (*upstep, downstep*), por lo que el transcriptor se aparta en este aspecto de las convenciones aplicadas habitualmente en ToBI para el uso de estos símbolos.

[1]La etiqueta L+H*+L aparece en el español de Argentina.

Los acentos monotonales son los mismos, tanto en la transcripción estrecha como en la ancha.

2.3.4. Clasificación de acentos nucleares

Las reglas que se han explicado hasta ahora se podían aplicar sin tener en cuenta el tipo de acento de la palabra. Sin embargo, para los acentos nucleares esto no es posible, puesto que los puntos entre los que el script busca diferencias significativas tienen que ser distintos en el caso de las agudas yen el de las llanas y esdrújulas.

En el caso de las palabras llanas y esdrújulas el script busca diferencias entre 3 puntos de la pretónica (inicio, centro y final), 5 de la tónica (inicio, centro, final, mínimo y máximo) y 5 de la postónica (inicio, centro, final, mínimo y máximo).También toma el máximo valor en la tónica y el mínimo en las postónicas.

Las reglas que se aplican funcionan como en el caso de los acentos nucleares, con la única diferencia de que el inventario tonal es algo diferente. Por ejemplo, en posición tonemática no se pueden encontrar los tonos L+<H* o L*+H, por lo que cuando estos aparecen en la transcripción fonética estrecha se reescriben en la transcripción ancha teniendo en cuenta el tono de frontera posterior.

Sin embargo, en el caso de las agudas, al no tener postónicas, el script solo puede tomar valores de la pretónica y de la tónica. Se considera que los primeros valores de la tónica se corresponderían con los movimientos tonales del acento nuclear tónico y los últimos, con los del tono de frontera, por lo que su clasificación se verá en la sección siguiente. De esta manera se resuelven los casos de compresión tonal.

2.3.5. Clasificación de tonos de frontera

Después de transcribir el acento nuclear, el script transcribe el tono de frontera. Para ello, las sílabas postónicas se dividen en 6 intervalos de tiempo iguales de los que se extrae la f_0. También se identifican los valores mínimo y máximo de f_0 en las postónicas.

El script recupera para esta transcripción algunos de los datos de la última sílaba tónica (datos de la f_0 máxima y un dato nominal que indica si la sílaba tónica acabó en un nivel alto o bajo). El script calcula las diferencias en semitonos entre estos puntos y procede a realizar la transcripción. Por ejemplo, ante una diferencia mayor de 1.5 semitonos entre el primer punto de las postónicas y el último, el script considerará que ha habido un ascenso y transcribirá H%. Pero si no ha habido ningún movimiento que pase el umbral pero la tónica acabó en un tono alto (por ejemplo, H* o L+H*), el tono de frontera resultante también será H%.

Para poder transcribir los tonos medios el script usa el rango de la IP. El rango del hablante se divide en tres niveles (L, !H y H) y si el final del movimiento tonal acaba dentro del nivel medio se transcribe como tono medio.

Como se avanzaba en la sección anterior, este análisis es diferente en el caso de las palabras agudas. Cuando el script detecta que la última sílaba acentuada es también la última sílaba de IP, la transcripción del acento nuclear y el tono de frontera se realiza de manera conjunta.

La sílaba se divide en 12 partes iguales y se toma un valor de la sílaba pretónica y 8 de la tónica. De los 8 valores que se toman de la tónica, los 3 primeros corresponderían al acento nuclear y los restantes al tono de frontera. A partir de estos datos, las fórmulas asignan directamente una configuración nuclear completa.

Como se ha dicho, esta estrategia resuelve los casos de compresión y de diferencias en el alineamiento tonal pero no resuelve los casos de truncamiento, ya que en estos los tonos subyacentes no se pueden recuperar a partir de los datos acústicos. Estos casos requieren un conocimiento fonológico de la lengua que solo puede tener un investigador humano o un transcriptor prosódico que incluya información semántica del contexto para realizar desambiguaciones.

2.3.6. Estandarización de configuraciones nucleares

El último paso del módulo de anotación tonal es la estandarización de las configuraciones nucleares. El script usa 23 reglas para realizar esta tarea.

Los tonos nucleares y de frontera incluyen los tonos medio (*mid)* y superalto en la capa fonética estrecha siempre que se cumplan las condiciones acústicas necesarias, como por ejemplo que el rango supere los 6 semitonos.

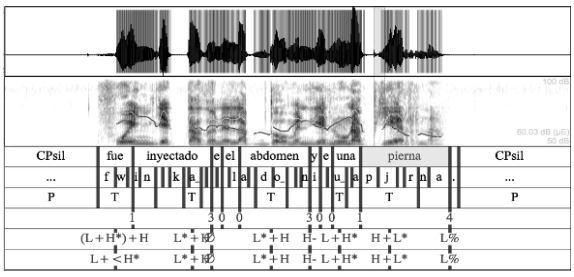
Sin embargo, estos tonos de frontera solo son susceptibles de ser fonológicos cuando aparecen al lado de algunas configuraciones nucleares. Por ejemplo, en posición nuclear, L+¡H* y L+H* solo pueden ser susceptibles de ser fonológicos si van seguidos de un tono L%. Igualmente, el tono de frontera !HH% solo es posible cuando va precedido de L+H*.

2.4. Salida

El script tiene dos salidas. La primera, por defecto, es el TextGrid con la transcripción ToBI y la segmentación prosódica. La segunda, opcional, es una figura con el resultado.

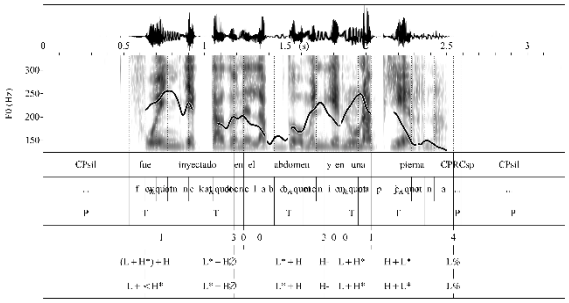
La anotación tonal resultante del proceso anterior se guarda en el TextGrid de entrada (Figura 6) en el que aparecen las transcripciones fonéticas estrecha y ancha en capas separadas, además de la segmentación prosódica y la segmentación en sílabas generadas por el primer módulo.

Figura 6: Onda sonora, espectrograma, curva de f_0 y anotación generado por FonetiToBI correspondiente al enunciado del español *'fue inyectado en el abdomen y en una pierna'*, pronunciado por un hablante femenino.



Elvira-García, W. & Garrido, J. M. (2019). FonetiToBI, una herramienta para la anotación prosódica automática de corpus. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 133–142). Málaga: Universidad de Málaga.

Figura 7: Ejemplo de figura obtenida con FonetiToBI: onda sonora, espectrograma, curva de f_0, transcripción de unidades entonativas y transcripción prosódica correspondiente al enunciado del español *'fue inyectado en el abdomen y en una pierna'*, pronunciado por un hablante femenino.



Además, la herramienta ofrece la posibilidad de crear una figura para cada uno de los ficheros de sonido analizados. La figura se guarda en formato png de 300 dpi (Figura 7). También puede visualizarse al final del proceso si se selecciona la opción correspondiente.

3. EVALUACIÓN

La fiabilidad del script se ha evaluado mediante la transcripción de un corpus de 100 grupos entonativos para el catalán y 1 186 para el español, y la posterior comparación de la transcripción obtenida con la proporcionada por transcriptores humanos. Los resultados de fiabilidad que se describen en esta sección se analizan con más detalle en Elvira-García *et al.* (2016).

3.1. Corpus

3.1.1. Corpus del español

El corpus que se ha usado para el español está constituido por 1 186 grupos entonativos producidos por 4 hablantes de diferentes variedades entonativas de español, a saber, español norteño (Cantabria), español castellano (Madrid), español de Cataluña (Barcelona) y un punto de español meridional (Sevilla). Las hablantes, de sexo femenino, tenían una edad comprendida entre 21 y 28 años y estudios superiores. La Tabla 1 muestra la distribución de los grupos entonativos evaluados en función del punto de encuesta.

El corpus se elicitó mediante una variación de la tarea de compleción del discurso (DCT; Blum-Kulka, 1982), y se grabó con una grabadora Marantz PMD620 conectada a un micrófono Shure SM58. Por lo tanto, se trata de un corpus con calidad de habla de laboratorio.

Tabla 1: Número de grupos entonativos por punto de encuesta en el corpus de evaluación en español.

Punto de encuesta	Número de IP
Barcelona	300
Cantabria	286
Madrid	300
Sevilla	300

3.1.2. Corpus del catalán

En el caso del catalán, el script se ha probado sólo con la variedad de catalán central. El corpus usado contiene 100 frases producidas por 20 informantes de entre 20 y 30 años que tenían el catalán como L1. Como en el caso anterior, se trata de un corpus elicitado mediante DCT y grabado con una grabadora Marantz PMD620 conectada a un micrófono Shure SM58.

3.2. Procedimiento

Tanto en el caso del catalán como en el del español se realizaron pruebas para comprobar que la transcripción de la herramienta coincidiera con la de transcriptores humanos. Sin embargo, los experimentos realizados en cada una de las lenguas fueron ligeramente diferentes.

3.2.1. Procedimiento para medir la fiabilidad en español

El test de fiabilidad del español consistió en la comparación de las transcripciones propuestas por FonetiToBI con las hechas por un transcriptor humano. Como las fórmulas para cada uno de los tipos de eventos tonales difieren, la comparación de los acentos prenucleares, los nucleares y los tonos de frontera se realizó por separado. Para cada uno de los tipos de evento tonal se calculó el número de veces que la transcripción coincidía y se determinó el nivel de acuerdo entre los dos mediante el coeficiente kappa de Cohen (Cohen, 1960).

3.2.2. Procedimiento para medir la fiabilidad en catalán

Los datos del catalán, sin embargo, no se compararon con un único transcriptor sino con cuatro. Los transcriptores, que desconocían que su transcripción se compararía con un transcriptor automático, recibieron un entrenamiento de 30 minutos en el que se les instruyó sobre el tipo de etiquetaje que debían realizar (un etiquetaje de carácter fonético ancho basado en ToBI, en el que se excluyeron los casos de truncamiento).

La comparación se realizó por pares, comparando el resultado de cada uno de los transcriptores con el transcriptor automático.

3.3. Resultados

3.3.1. Resultados para el español

Los resultados para el español muestran un nivel de coincidencia bueno para los tonos de frontera y muy bueno para los dos tipos de acentos tonales (Tabla 2). Más en concreto los acentos prenucleares (APN) tienen un nivel de coincidencia del 94.94 % y un kappa de 0.9.

Los acentos nucleares (AN) muestran un nivel de coincidencia del 88.11 % y un kappa de 0.8. El nivel más bajo de coincidencia se obtiene en los tonos de frontera (81 % y 0.7 de kappa). Este efecto probablemente se debe al hecho de que los tonos de frontera final coinciden con el final absoluto de la

Tabla 2: Coincidencia en porcentaje, valor kappa y valoración delacoincidencia entre la transcripción de FonetiToBI y la del transcriptor humano del corpus del español.

Evento tonal	n	%	Kappa	Evaluación
APN	1660	94.94 %	0.907	muy buena
AN	1186	88.11 %	0.831	muy buena
TF	1186	81.28 %	0.756	buena

Tabla 3: Coincidencia en porcentaje, valor kappa y valoración delacoincidencia entre la transcripción de FonetiToBI y la de los transcriptores humanos del corpus del catalán.

Evento tonal	T	%	Kappa	Evaluación
AN	1	85.71 %	0.772	buena
	2	85.71 %	0.770	buena
	3	82.65 %	0.722	buena
	4	78.79 %	0.657	buena
TF	1	92.86 %	0.884	muy buena
	2	92.86 %	0.885	muy buena
	3	93.88 %	0.900	muy buena
	4	90.82 %	0.851	buena

emisión, ya que el corpus estaba formado por frases aisladas. En final de emisión, las vocales pueden aparecer ensordecidas y, en general, hay una pérdida sustancial de cualidad de voz con la aparición de fenómenos como la voz rota (*creaky voice*) que pueden hacer que el algoritmo de detección de f_0 de Praat no funcione correctamente y aparezcan fallos.

3.3.2. Resultados para el catalán

Los resultados para el catalán (Tabla 3) varían dependiendo del transcriptor, pero aun así todos los valores obtenidos se sitúan en niveles buenos o muy buenos. Para los acentos nucleares los valores en porcentaje oscilan entre el 78 % y el 85 %, mientras que los kappa tienen valores de entre 0.65 y 0.77. Los tonos de frontera, sin embargo, muestran en catalán mayor porcentaje de acierto (entre el 90 % y el 93 %) y sus valores kappa se mueven entre el 0.85 y el 0.88.

3.3.3. Análisis cualitativo de los errores

En cuanto al tipo de errores que comete el script, se ha podido observar que para algunas configuraciones nucleares hay más discrepancias entre el script y los etiquetadores humanos que para otras.

En la matriz de confusiones de la Tabla 4, obtenida a partir de los resultados de la evaluación para el catalán, se puede comprobar que, en la mayoría de casos, las configuraciones nucleares han sido transcritas igual por el transcriptor automático y los humanos (diagonal de la tabla)

En la Tabla 4 se observa también que las configuraciones nucleares que suponen más problemas corresponden a, por un lado, confusiones entre L+H* L% y H* L% y, por otro, confusiones entre L+H* L% y L+¡H* L%. La primera confusión tiene que ver con la implementación fonética de los tonos fonológicos. Cuando en el acento L+H* hay un escalonamiento ascendente del tono L debido a un tono alto anterior, dado que el tono fonético resultante es alto, el script no puede detectar la diana tonal baja por lo que transcribe solo los tonos altos que sí se han realizado.La segunda confusión tiene que ver con el rango de los eventos tonales. El requisito para etiquetar un tono como extra-alto se marcó en el script como una diferencia de más de 6 semitonos con la sílaba anterior. Este número se obtuvo a partir de la comparación de algunos estudios de percepción de la literatura. Sin embargo, viendo la cantidad de discrepancias entre el script y los etiquetadores parece claro que hay que revisar ese dato.

4. CONCLUSIONES Y POSIBLES MEJORAS

A la vista de los resultados obtenidos en las pruebas de evaluación, puede afirmarse que FonetiToBI es capaz de anotar prosódicamente enunciados simples en español y catalán con un nivel de precisión aceptable, similar al de los etiquetadores humanos, por lo que puede aplicarse con garantías al etiquetado automático de grandes corpus, y con un considerable ahorro en el tiempo de etiquetado. Sin embargo, hay que tener en cuenta que los test se llevaron a cabo con enunciados grabados en laboratorio y en los que la mayoría de sonidos eran sonoros, por lo que la fiabilidad del script podría disminuir si se usan corpus con calidad inferior o con abundantes sonidos sordos.

Otro aspecto que es importante resaltar es el hecho de que ofrezca, con un mismo sistema de transcripción, dos niveles diferentes de anotación tonal: uno, el de transcripción fonética estrecha, más basado en la forma acústica de la curva de f_0, con un inventario de marcas tonales más rico (es decir, más fonético); y otro, el denominado aquí como transcripción fonética ancha, con un inventario de símbolos más reducido y más cercano a las convenciones establecidas en Cat_ToBI y Sp_ToBI, que pretenden recoger solo los eventos tonales contrastivos (más fonológico, por tanto). Esta posibilidad de transcribir fonética y fonológicamente con un mismo inventario de símbolos los eventos tonales de una lengua, existente ya desde hace tiempo para la transcripción de los elementos segmentales con el Alfabeto Fonético Internacional, aumenta de forma importante las aplicaciones potenciales de esta herramienta y del sistema mismo de anotación. Por ello, una de las tareas previstas para un futuro próximo es mejorar el inventario de etiquetas usado en el nivel fonético para permitir una transcripción lo más fiel posible de los eventos tonales observados en los contornos.

El hecho de que FonetiToBI admita también como entrada transcripciones fonéticas realizadas en SAMPA abre la posibilidad, como se ha mencionado al principio, de utilizar herramientas automáticas, como SPPAS o el propio Praat, para obtener el fichero TextGrid que requiere como entrada. Se podría llegar a obtener así, de forma completamente automática, una transcripción fonética completa (segmental y suprasegmental) a partir de un fichero wav con la onda sonora de un enunciado y de un fichero de texto con su transcripción ortográfica. Otra de las líneas en las que se está trabajando

Tabla 4: Matriz de confusiones de las diferencias entre el etiquetaje de FonetiToBI y los transcriptores humanos en la evaluación del corpus en catalán.

	H* L%	H+L* L%	L* H%	L+H* L%	L+¡H* L%	L* HL%	L+H* L!H%	L+H* LH%	L* L%
H* L%	52	6		4	1			1	4
H+L* L%	6	12							1
L* H%			179	4		2			
L+H* L%	11		4	7	20				
L+¡H* L%	1			20	11				
L* HL%			2			18			3
L+H* L!H%							0	3	
L+H* LH%	1						3	24	
L* L%	4	1				3			8

actualmente es en el desarrollo de una herramienta que permita realizar esta tarea.

Finalmente, hay que destacar el hecho de que FonetiToBI es una herramienta de libre distribución, disponible bajo licencia GNU de manera gratuita en:

- http://stel.ub.edu/labfon/en/praat-scripts;
- https://sites.google.com/site/juanmariagarrido/research/resources/tools/fonetitobi

5. REFERENCIAS

Bigi, B. (2012). SPPAS: a tool for the phonetic segmentations of speech *The Eight International Conference on Language Resources and Evaluation, Istanbul (Turkey)* (pp. 1748–1755). Recuperado de http://www.lrec-conf.org/proceedings/lrec2012/pdf/1116_Paper.pdf.

Blum-Kulka, S. (1982). Learning to say what you mean in a second language: A study of the speech act performance of learners of Hebrew as a second language. *Applied Linguistics, 3*(1), 29–59.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *IFA Proceedings 17*, 97–110.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*, 341–345.

Borràs-Comes, J., Vanrell, M. del M., & Prieto, P. (2014). The role of pitch range in establishing intonational contrasts. *Journal of the International Phonetic Association, 44*, 1–20.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

De Looze, C. (2010). *Analyse et interprétation de l'empan temporel des variations prosodiques en français et en anglais* (Tesis Doctoral). Aix-en-Provence: Université Aix-en-Provence.

Elvira-García, W., Roseano, P., & Fernández Planas, A. M. (2015). Una herramienta para la transcripción prosódica automática con etiquetas Sp_ToBI en Praat. En A. Cabedo Nebot (Ed.) *Perspectivas actuales en el análisis fónico del habla: tradición y avances en la fonética experimental* (pp. 455–464). València: Universitat de València.

Elvira-García, W., Roseano, P., & Fernández Planas, A. M. y Martínez Celdrán E. (2016). A tool for automatic transcription of intonation: Eti-ToBI a ToBI transcriber for Spanish and Catalan. *Language Resources and Evaluation, 50*(4), 767–792.

Garrido, J. M. (2010). A tool for automatic F0 stylization, annotation and modelling of large corpora. *SpeechProsody 2010, Chicago, May 2010*. Recuperado de http://speechprosody2010.illinois.edu/papers/100041.pdf.

Garrido, J. M. (2013). SegProso: A Praat-based tool for the automatic detection and annotation of prosodic boundaries. *Proceedings of TRASP 2013* (pp. 74–77). Recuperado de http://www.lpl-aix.fr/~trasp/Proceedings/19864-trasp2013.pdf.

Goldman, J. P. (2011). EasyAlign: an automatic phonetic alignment tool under Praat. *Proceedings of InterSpeech, September 2011, Firenze, Italy*.

Hirst, D. (2011). The analysis by synthesis of speech melody: from data to models. *Journal of Speech Sciences, 1*(1), 55–83.

Hirst, D., & Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix, 15*, 71–85.

Hualde, J. I., & Prieto, P. (2016). Towards an International Prosodic Alphabet (IPrA). *Laboratory Phonology: Journal of the Association for Laboratory Phonology, 7*(1).

International Phonetic Association (en línea). *The International Phonetic Alphabet and the IPA Chart*. Recuperado de https://www.internationalphoneticassociation.org/content/ipa-chart.

Martínez Celdrán, E., & Fernández Planas, A.M. (2003). Taxonomía de las estructuras entonativas de las modalidades declarativa e interrogativa del español estándar peninsular estándar según el modelo AM en habla de laboratorio. En E. Herrera & P. Martín (Eds.), *La tonía: dimensiones fonéticas y fonológicas* (pp. 267–295). México DF: El Colegio

de México.

Pamies, A., Fernández Planas, A. M., Martínez Celdrán, E., Ortega-Escandell, A., & Amorós Céspedes, M. C. (2002). Umbrales tonales en español peninsular. *Actas Del II Congreso de Fonética Experimental* (pp. 272–278).

Prieto, P. & Cabré, T. (Eds.) (2013). *L'entonació dels dialectes catalans*. Rubí: Publicacions de l'Abadia de Montserrat.

Prieto, P., & Roseano, P. (Eds.) (2010). *Transcription of intonation of the Spanish language*. München: Lincom Europa.

Rietveld, A., & Gussenhoven, C. (1985). On the relation between pitch excursion size and prominence. *Journal of Phonetics*, *13*, 299-308.

Roseano, P., & Fernández Planas, A. M. (2013). Transcripció fonètica i fonològica de l'entonació: una proposta d'etiquetatge automàtic. *Estudios de Fonética Experimental*, *22*, 275–332.

Vizcaíno Ortega, F., Cabrera Abreu, M., Dorta, J., & Hernández Díaz, B. (2007). La entonación de enunciados declarativos e interrogativos absolutos de Lanzarote. En J. Dorta (Ed.), *La prosodia en el ámbito lingüístico románico* (pp. 347–369). Santa Cruz de Tenerife: La Página Ediciones.

Elvira-García, W. & Garrido, J. M. (2019). FonetiToBI, una herramienta para la anotación prosódica automática de corpus. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 133–142). Málaga: Universidad de Málaga.