Design of an Efficient Music-Speech Discriminator

Lorenzo J. Tardón,[a)] Simone Sammartino, and Isabel Barbancho

Dept. Ingeniería de Comunicaciones

E.T.S. Ingeniería de Telecomunicación

Universidad de Málaga

Campus Universitario de Teatinos s/n

E29071 Málaga Spain

Phone: +34952131188

Fax: +34952132027

(Dated: September 23, 2009)

Abstract

In this paper, the problem of the design of a simple and efficient music-speech discriminator for large audio data sets in which advanced music playing techniques are taught and voice and music are intrinsically interleaved is addressed. In the process, a number of features used in speech-music discrimination are defined and evaluated over the available data set. Specifically, the data set contains pieces of classical music played with different and unspecified instruments (or even lyrics) and the voice of a teacher (a top music performer) or even the overlapped voice of the translator and other persons. After an initial test of the performance of the features implemented, a selection process is started, which takes into account the type of classifier selected beforehand, to achieve good discrimination performance and computational efficiency, as shown in the experiments. The discrimination application has been defined and tested on a large data set supplied by Fundación Albéniz, containing a large variety of classical music pieces played with different instrument, which include comments and speeches of famous performers.

PACS numbers: 43.60.Dh,43.60.Lq, 43.75.Xz

Keywords: Music and Speech Discrimination, Fisher Classifier, Linear Discrimination Analysis

I. INTRODUCTION

For many years, the problem of discrimination between music and speech has been an important task for audio/video processing research groups, due to the growing need of supplying classified audio archives over large data sets of radio broadcasts, musical recordings etc. Nowadays, there is a strong source of demanding applications, some of them due to the increasing web access to audio-video contents, based on tagged audio-video retrieval; this is the specific context for which we develop our system. The search and definition of efficient descriptors for the discrimination of audio obtained by different sources is handled since '60s[1] and since then, many authors dealt with the problem from different points of view. In Aarts and Dekkers[2], a simple electronic setup for audio signal pre-processing and features extraction is shown, which employs a fuzzy combiner based on two specific features for real time discrimination of speech and music (SMD); in the paper by Saunders[3], a method for SMD in radio broadcast audio is shown, it is based on the definition of a threshold on a measure of the skew of the distribution of the zero crossing rate (ZCR). In the paper by Carey et al.[4], a more general view of features for SMD is presented, the performance of some of them is evaluated using a simple classifier based on the extracted Gaussian mixture model parameters (GMM)[5], although the work is not intended to a specific application. In Jarina et al.[6], another point of view is considered: how to deal with encoded audio; in this case, MPEG-1 audio is considered and the volume of segments, directly extracted from coded data, is used in the discrimination. In Harb and Chen[7], a multi layer perceptron is used to perform discrimination, using, as features, the mean and variance of the samples of the Mel-filtered spectrum. Goodwin and Laroche[8] argue on the importance of the features selected for MSD and propose to incorporate a wide variety of features to the discrimination system and, also, linear discriminant analysis is said to be applied to reduce dimensionality of features vectors and the paper is focused on the design of the cost functions for a dynamic programming algorithm to find data clusters. In Panagiotakis and Tziritas[9], the combination of features

[a)]Electronic address: lorenzo@ic.uma.es

extracted, related to root mean square measures (RMS) and zero-crossing rate (ZCR), are the core of the study, where the behavior of the features is, in this case, analyzed in detail. In this paper, means and variances of several features are considered and a classification based on some specific tests is described.

In this paper, we consider some of the ideas of these authors, and others, so we will pay attention to a number of features which can be useful for discrimination and we make an analysis of their behavior before using them blindly in our system. We also consider the choice of a simple and efficient classification strategy and we search for an optimal combination of the features for the classification, according to a certain criterion.

This paper is organized as follows: in section II, the overall set of features, mostly extracted from the selected bibliography, is shown and an analysis of the behavior of the most relevant ones is done. In section III, we present the type of classifier selected for the target application, and in section IV, the discrimination performances of the implemented features and the selection of features to optimize the performance of the system with the classification strategy chosen are shown. Finally, in section V, some conclusions are drawn.

II. FEATURES FOR SPEECH-MUSIC DISCRIMINATION

As mentioned before, a main stage of a SMD process consists on the selection and extraction of numerical features from the different audio sources. A wide range of features have been used and proposed for the purpose of SMD, and some of them are considered as extracted in the time domain and others in the frequency domain and their usage reaches other areas of audio processing like speech recognition[10,11] and music classification[12].

A common stage previous to feature extraction is the division of the audio data into small pieces in order to classify each of them as speech or music. As this granularity has certain influence on the performance of the features selected, we present the relevant parameters before the description of the features and their analysis. In our implementation, the audio files are sampled at 44,100 Hz and they are divided into segments of one second (T_s) of

duration and each descriptor is calculated on a number of portions of these segments; we will call them frames and their duration will be 20 msec (T_f). Using these global parameters, a number of features are extracted to create a feature vector with audio data divided into segments and frames, as described, and then, sample means and variances are obtained for time and frequency domain features. Although high order sampled statistics are more sensitive to noise, the analysis of the sampled variance is important due to the different behavior of voiced and unvoiced parts of speech with respect to different features[13].

In the following, all the implemented features are discussed, while the ones that were finally selected for the optimal subset, are exposed with more detail. Although more features could have been considered, even non-linear combinations of some of them, as in Panagiotakis and Tziritas[9], we have limited our work to the ones shown in the following sections, which lead to adequate discrimination performance. In the following, the input data will be denoted as $\overline{x}$ (we drop the segment index to simplify the notation), the 20 millisecond length vector will be denoted as $\overline{x}_j = [x_j(1), x_j(2), \ldots, x_j(n)]^t$, where j indexes the frame under analysis, with $j = 1, 2, \ldots, 50$, where 50 is the number of frames per segment in our implementation, and $\frac{T_s}{T_f}$, in a general case. Hence, for each segment, the mean and standard deviation of each descriptor (generally denoted as d) will be computed and stored in the vector $\overline{d} = [d_1, d_2, \ldots, d_{50}]^t$. Note that, for simplicity, we have omitted the segment index.

A. *Selected Features for the optimal subset*

In this section, we expose a brief description and illustrate the specific behavior of the selected features on the training data set available for the project. Descriptions are explicitly provided to ensure the reproducibility of the experiments and results obtained in this work, thus, specific parameters that may be involved in the definition of the feature extraction process will be defined. The frames and the segments defined to obtain the features will be disjoint and multiplied by a rectangular window function, unless otherwise indicated for selected features.

1. Root mean squares

The root mean squares (RMS) is related to the signal power[14] and it is estimated as:

$$RMS_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_j^2(i)} \tag{1}$$

where n is the number of samples in the audio frame under analysis.

Generally the feature reflects the higher occurrence of lower power frames in speech and the more homogeneous distribution of sound power in music. In Fig. 1, the scatter plot of means and standard deviations of the RMS of a set of hundreds of 1 second length audio segments is shown. Illustrative ellipses have been drawn to facilitate the observation of the different behavior of the features extracted from music and speech samples. The ellipses have their center at the mean of each feature, the semi-axes are 1.5 times the standard deviations of the features (σ_{mean} and σ_{std}) and the rotation angle is obtained using $atan\left(\frac{\sigma_{std}}{\sigma_{mean}}\right)$, it has been observed that this choice is not heavily affect by the outliers.

Note that the presence of silence or quasi-silence instants, more frequently in speech than in music, leads to a higher ratio between the spreads of means and standard deviations in speech than in music.

2. Zero crossing rate

The zero crossing rate (ZCR) measures the rate of zero amplitude crossings of a sound signal with respect to its length. It is a good indication of the dominant frequency[15], and it is highly correlated with the spectral centroid[16]. This feature is computed counting the number of zero-crossing of the audio signal per second of sample (for a given sampling rate). In our implementation, we use:

$$ZCR_j = \frac{1}{2n}\sum_{i=1}^{n} |sign(x_j(i)) - sign(x_j(i-1))| \tag{2}$$

where $sign(x_j(i))$ is 1 for positive amplitudes and -1 for negative ones and n is the number of samples in the audio frame.

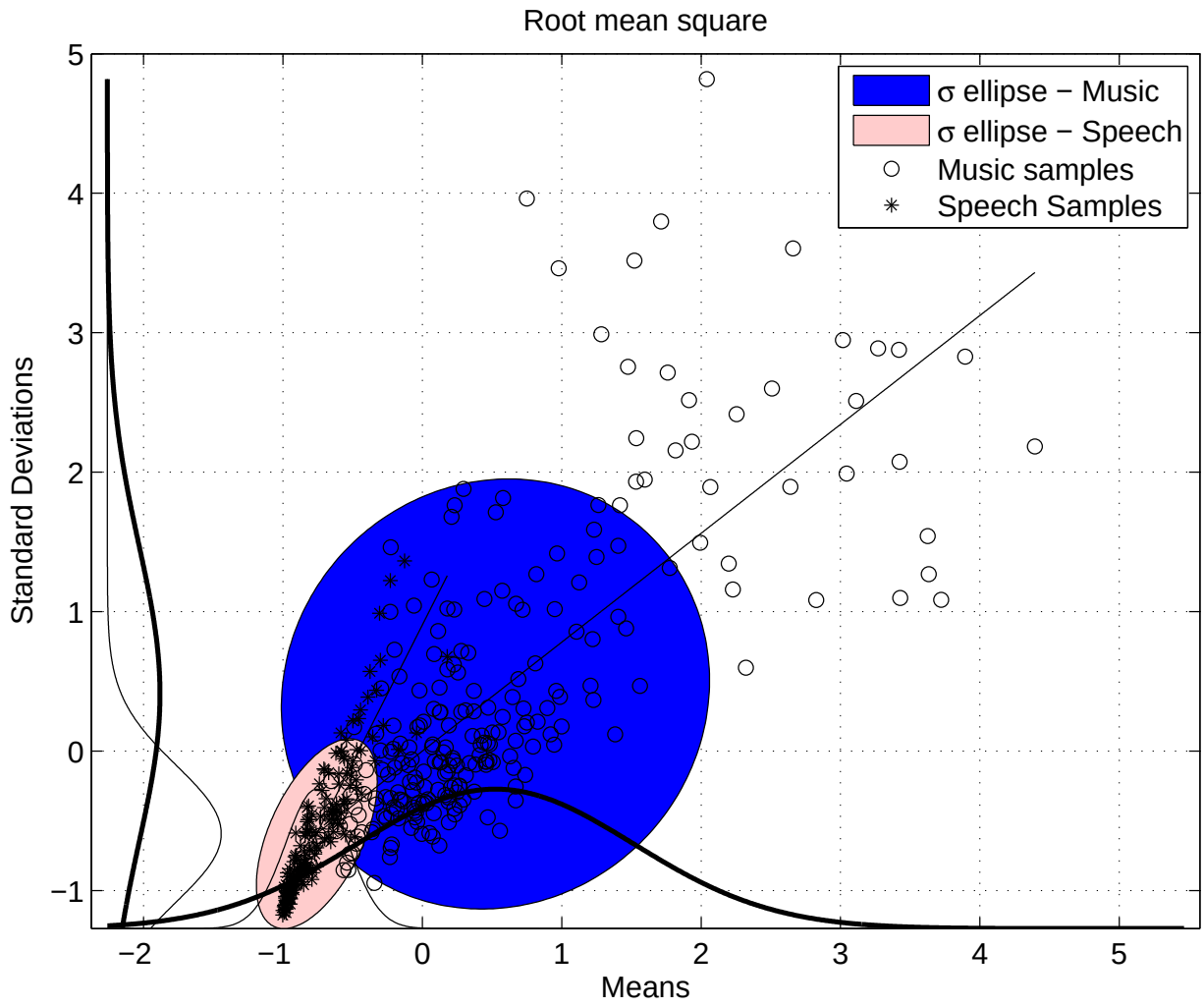

FIG. 1. (color online) Scatter plot of speech and music root mean squares means and standard deviations of a large 1 second length samples set. Linear fits, 1.5σ ellipses and normal pdfs for means and standard deviations are shown too.

Generally, in speech samples, voiced and unvoiced segments exhibit characteristic low and high zero-crossing rates[13].

Thus, a relevant difference in variance between the standard deviations of speech and music is easily observed in the scatter plot shown in Fig. 2.

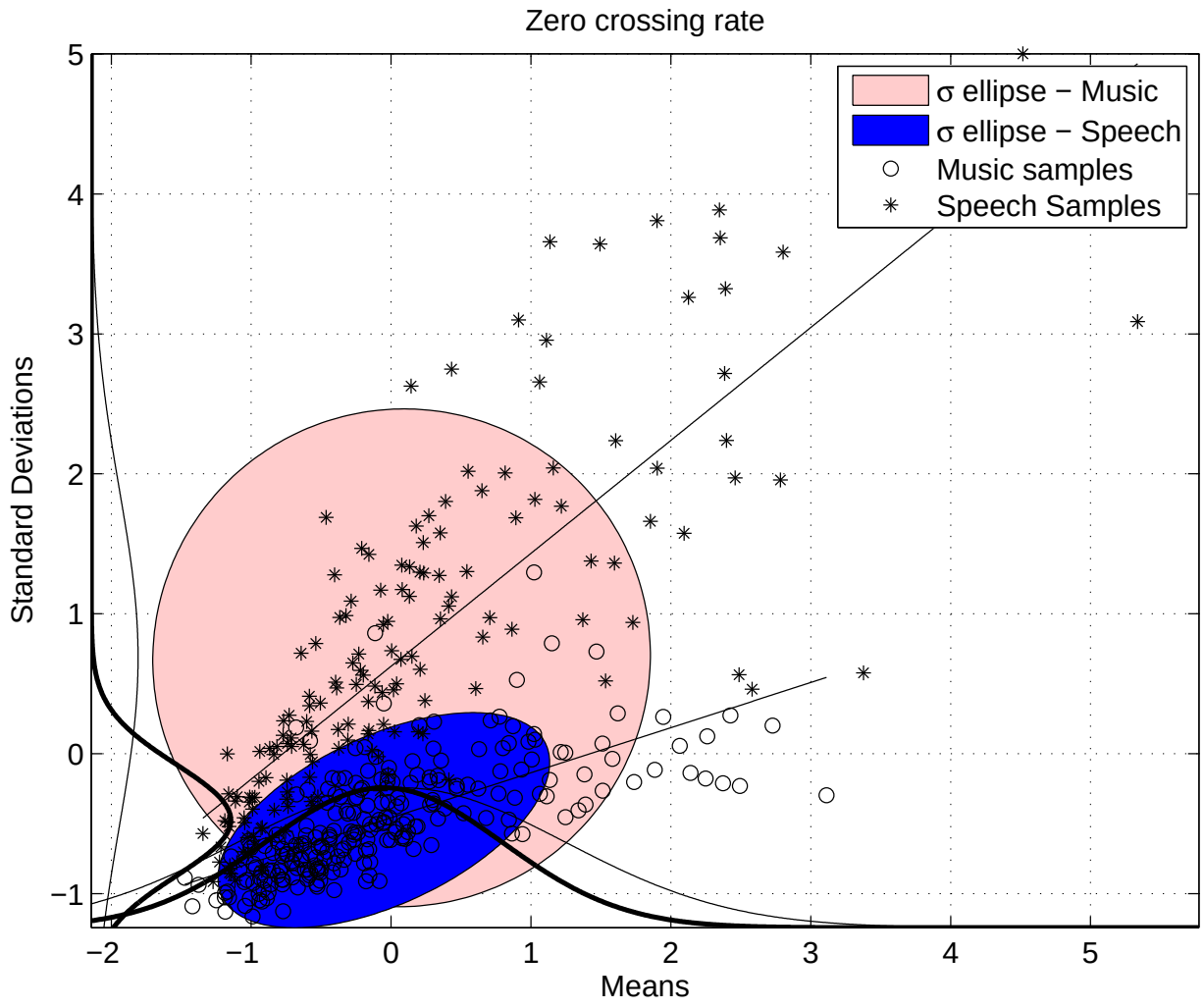

FIG. 2. (color online) Scatter plot of speech and music zero crossing rate means and standard deviations of a large 1 second length samples set. Linear fits, 1.5σ ellipses and normal pdfs for means and standard deviations are shown too.

3. Cepstrum residuals

The term cepstrum is a form of assessing the spectrum shape of an audio signal[16], in the so called cepstrum domain[1]. The cepstrum residuals give a measure of the rate of change of the bands of the spectrum.

In this work, the cepstrum residuals (CER) are computed as the Euclidean distance between the cepstrum of signal and its smoothed form[17], specifically, we use the following

expression:

$$CER_j = \sqrt{\sum_{k=1}^{m} \left(C_j(k) - SC_j(k)\right)^2} \tag{3}$$

where $C_j(k)$ is the k-th element of the raw cepstrum obtained with a DFT of length 1024 of the frame j and $SC_j(k)$ is the k-th element of the j-th smoothed cepstrum using a moving average method with a 5 samples window. The cepstrum $\overline{C}_j = [C_j(1), C_j(2), \ldots]^T$ ares obtained as follows:

$$\overline{C}_j = DFT^{-1}\left\{ln(|DFT(\overline{x}_j)|)\right\} \tag{4}$$

where $DFT\,()$ represents the function to obtain the Discrete Fourier Transform and $\overline{x}_j$ is is padded with zeros before the calculation of the DFT.

The behavior of CER is shown in Fig. 3. Cepstrum residuals show higher mean and variance for speech samples with respect to music ones, where good class-separability characteristics are exhibited, since speech contains higher variations of the frequency contents than music and their smoothing process causes a rising of the residuals.

4. Spectral flux

The spectral flux (SPF) is a measure of the variation rate between short fractions of audio vectors. From a mathematical point of view, such feature has been computed as the Euclidean norm of the difference between the magnitude vectors of two adjacent audio fragments[17]. In this work, such residuals are extracted for each pair of frames in each segment, using a DFT of length 1024.

Basically, the spectral flux measures how fast the signal spectrum changes. In the literature, different interpretations of the behavior of the feature are found[16,17], but concerning our work, we consider that the type of audio source is a main aspect that influences the outcomes of the feature and its behavior.

Specifically, in the samples supplied by Fundación Albéniz for the development of the discrimination application, the music played or the lyrics sung are short fragments interrupted by teacher's or student's speech with mean and variance of the spectral flux larger

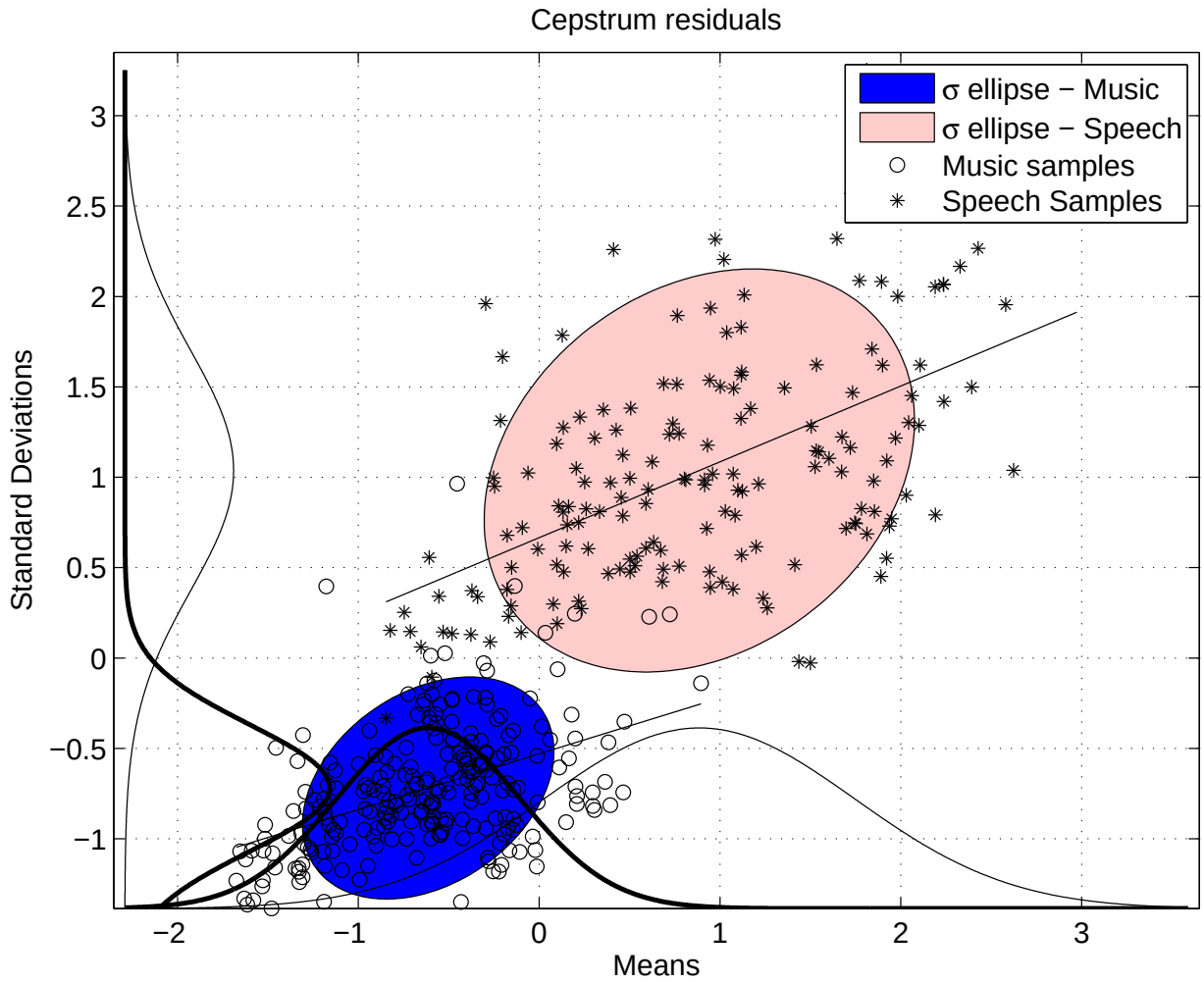

FIG. 3. (color online) Scatter plot of speech and music cepstrum residuals means and standard deviations of a large 1 second length samples set. Linear fits, 1.5σ ellipses and normal pdfs for means and standard deviations are shown too.

in music segments than in speech ones, as shown in Fig. 4.

5. Magnitude

Magnitude (MAG) refers to the behavior of the magnitude of the DFT calculated for each segment (not sectioning it into frames). The mean and standard deviation of the magnitude samples of the DFT are calculated. A detailed description of the algorithm used

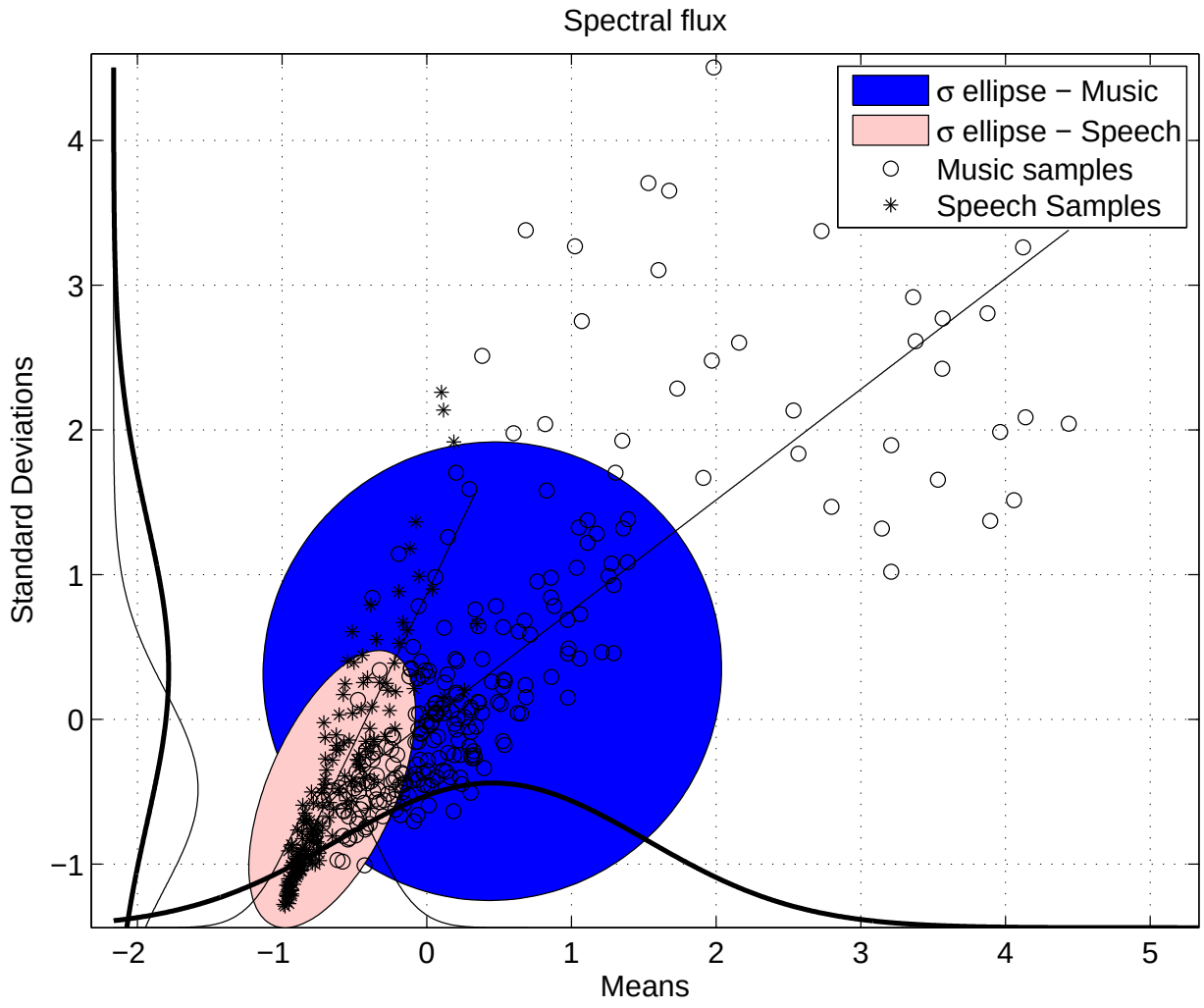

FIG. 4. (color online) Scatter plot of speech and music spectral flux means and standard deviations of a large 1 second length samples set. Linear fits, 1.5σ ellipses and normal pdfs for means and standard deviations are shown too.

follows. Let $\overline{X} = DFT(\overline{x})$, where the DFT is of length m, with k the smallest integer such that $m = 2^k \geq N$, with N the length of the segment, then the mean values and the standard deviation of the bins if $\left|\overline{X}\right|$ are obtained. Fig. 5 shows, visually, that discrimination based on both mean and variance can be performed, we have found that this feature is one of the simplest features with usable discrimination capabilities between speech and music.

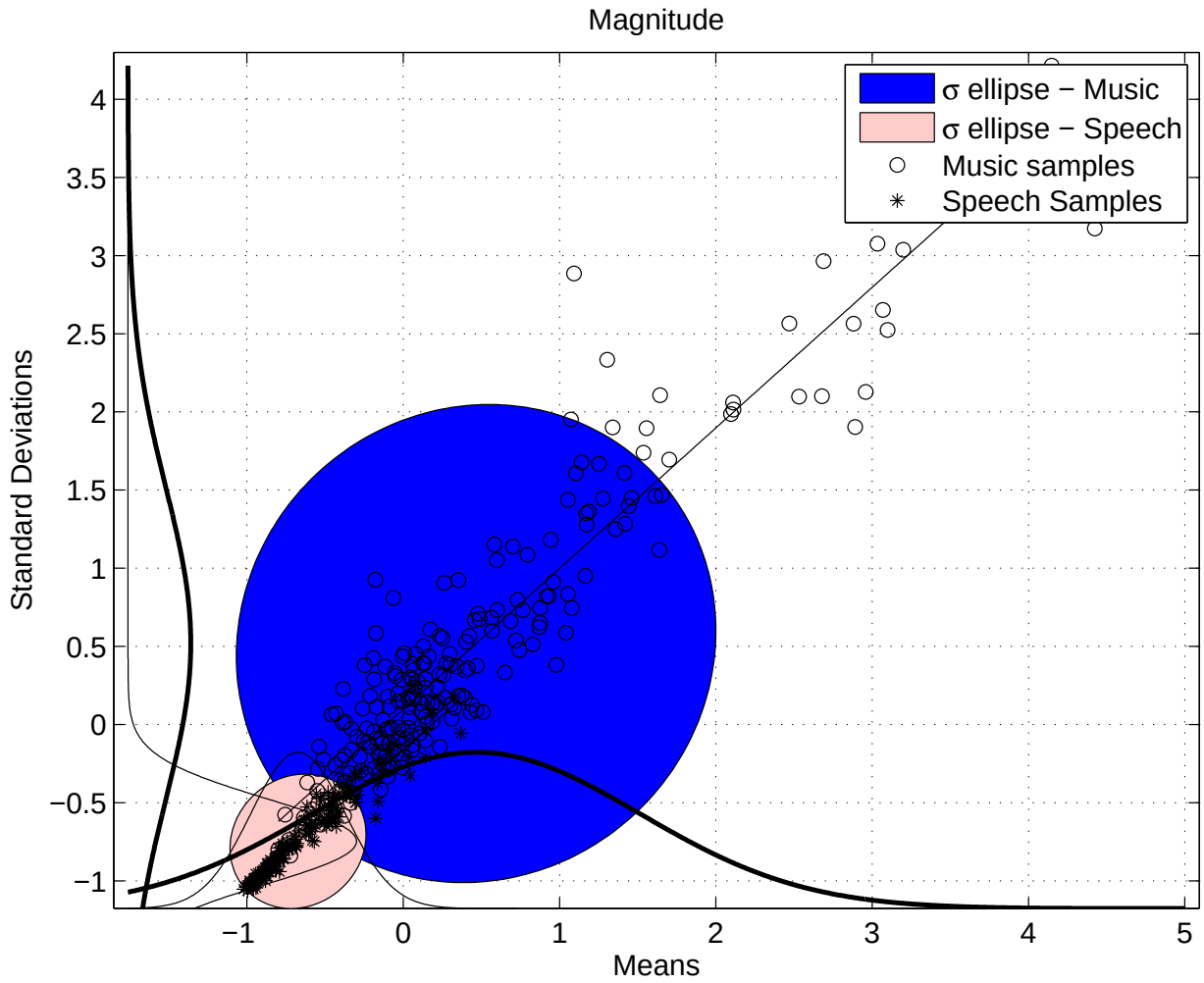

FIG. 5. (color online) Scatter plot of speech and music magnitude means and standard deviations of a large 1 second length samples set. Linear fits, 1.5σ ellipses and normal pdfs for means and standard deviations are shown too.

6. Mel frequency cepstral coefficients

The analysis of the human perception of frequency revealed that logarithmic arrangement of frequency for audio samples is perceived as linearly stepped by a group of listeners in standard conditions. This study gave rise to the so called Mel-scale[18]. One of the main usages of the Mel scale in audio signal processing, is the derivation of the Mel frequency cepstral coefficients (MFCC's), regarded as one of the most efficient tools for speech compression

and retrieval[19].

The computation process of the MFCC's[11,20] includes a pre-emphasis filter applied before the 512 point DFT. After that, a Mel filter bank, with 40 filters, is applied to the amplitude of the DFT. Pre-emphasis of the signal, $\overline{x}$, consists on the application of a FIR filter defined by the following expression $x'(i) = x(i) - ax(i-1)$, with $a = 0.97$[20]. A Hamming window is applied before the DFT to reduce the discontinuity between frames in a segment and the magnitude of the side lobes in the DFT due to the windowing[20]. A 50% of overlap is employed when analyzing the signal. After the Mel filtering, the discrete cosine transform DCT of the logarithm of the output is calculated. Note that there may be slight differences in the calculation of the MFCCs due to the differences in the pre-filtering or even in the specific Mel filter-bank used[21], in our case, the integral of each filter, in the bank of Mel filters, is normalized to one. In this work, the first 5 coefficients have been selected[12]. Each coefficient is considered a single feature, whose mean and standard deviation are extracted.

In Fig. 6, a scatter plot of the second coefficient is shown, where its discriminating capability is evident.

7. Volume dynamic ratio

The spread of RMS over the frames in a segment, normalized with respect to its maximum value, depends on the type of sound source[22] and it can also discriminate efficiently music from speech. This feature is referred to as volume dynamic ratio (VDR) and in our implementation we use:

$$VDR = 1 - \frac{min_j(RMS_j)}{max_j(RMS_j)} \tag{5}$$

In each segment, speech presents higher mean values and lower variance than music (Fig. 7).

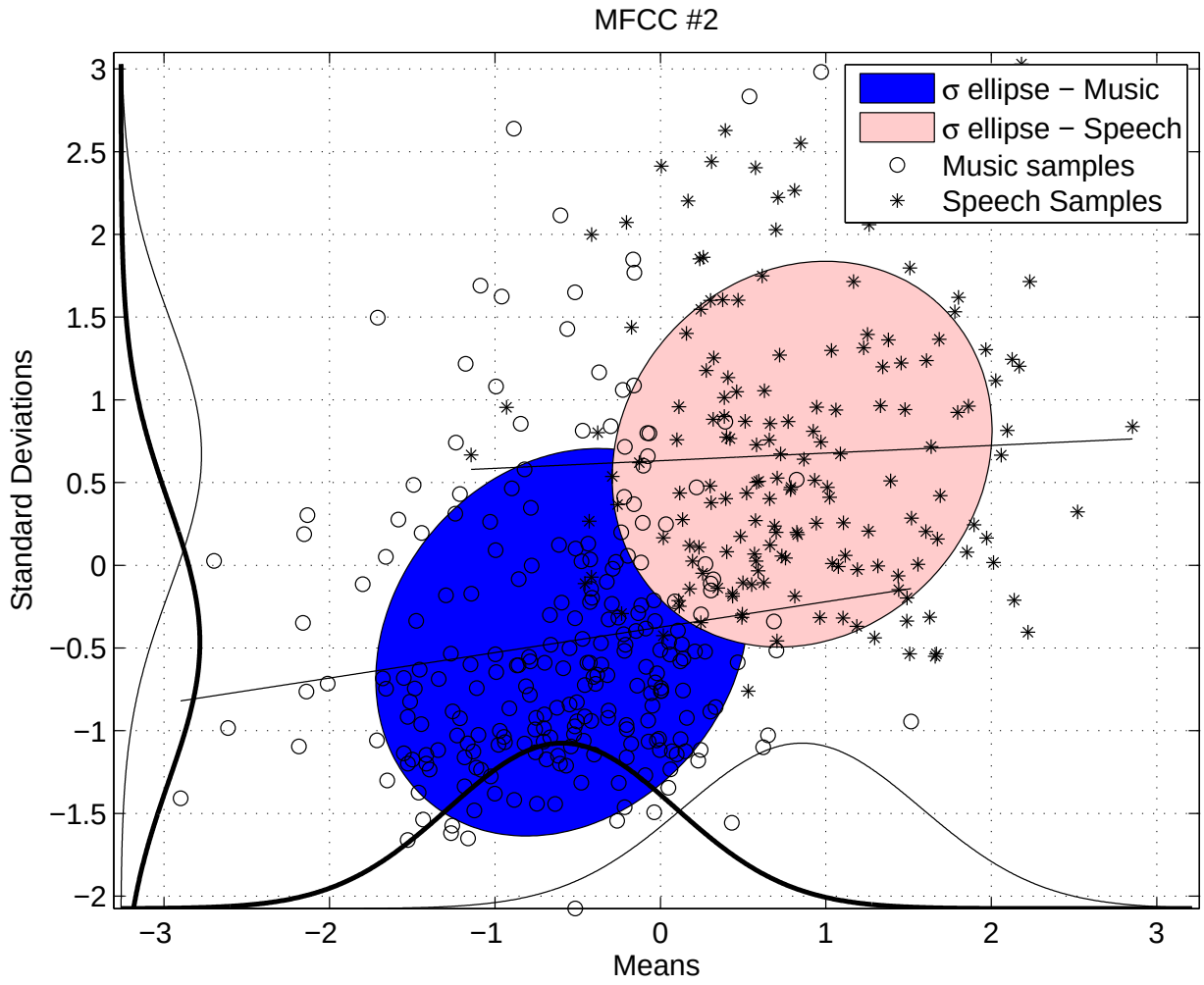

FIG. 6. (color online) Scatter plot of speech and music Mel frequency cepstral coefficient #2 means and standard deviations of a large 1 second length samples set. Linear fits, 1.5σ ellipses and normal pdfs for means and standard deviations are shown too.

B. *Other features evaluated*

In this section, the audio features evaluated that were not selected for the final discrimination subset are presented. The specific implementation used in our study is shown but no figures illustrate the behavior, in this case.

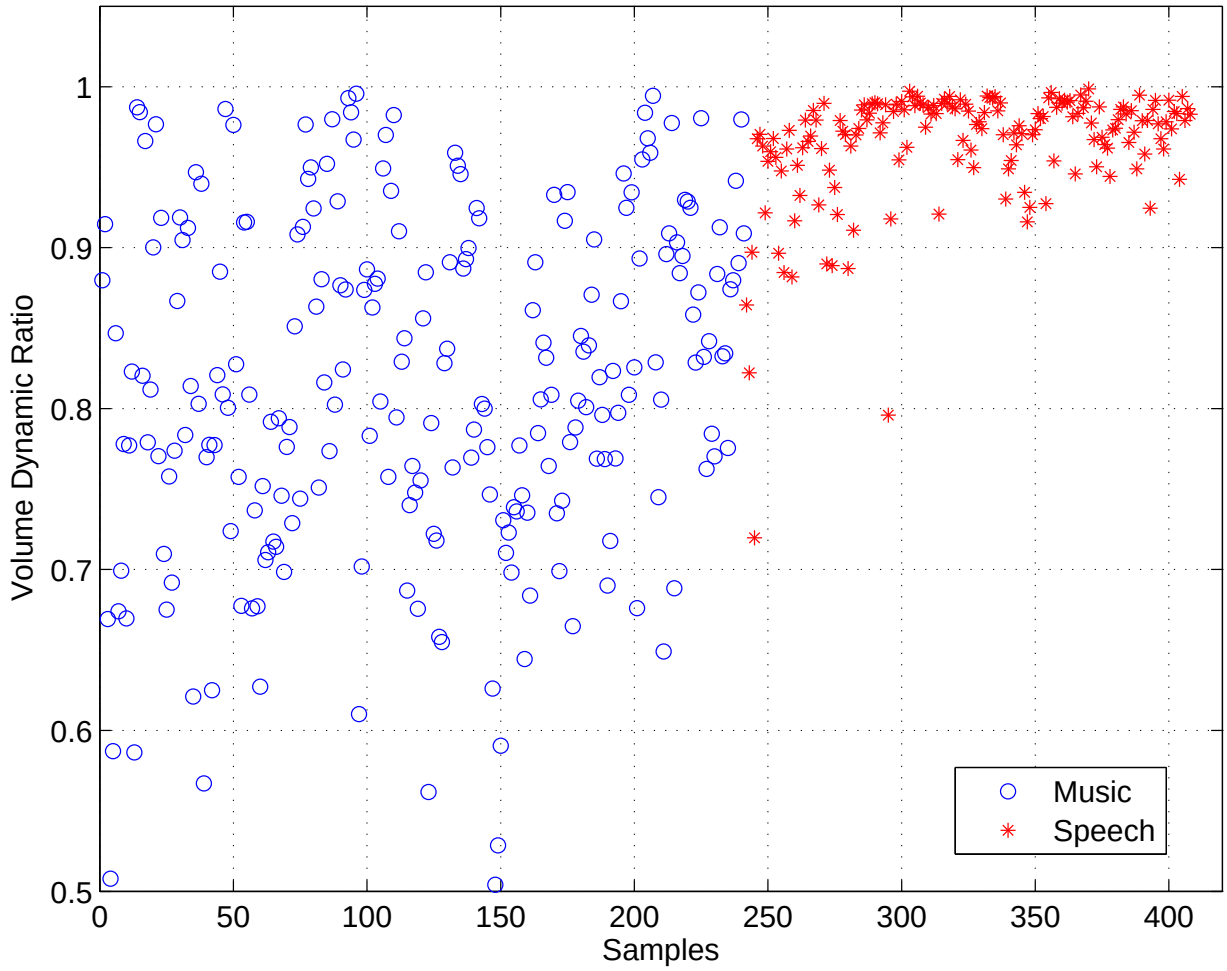

FIG. 7. (color online) Scatter plot of speech and music volume dynamic ratio outcomes.

1. Silence ratio

One segment is considered 'silence' if its amplitude is lower than a fixed threshold[14]. In this work, the silence ratio is calculated tagging the frames with RMS_j lower than 10% of the RMS of the entire segment:

$$SIR = \frac{card\{j : RMS_j \leq 0.1 \cdot RMS\}}{50} \tag{6}$$

where the function $card\{\}$ returns the number of elements that accomplish the specified condition.

2. Spectral centroid

This is the center of gravity of the spectrum itself[12] and it is estimated in each frame as follows[23]:

$$SPC_j = \frac{\sum_{k=1}^{m} F_j(k) \cdot M_j(k)}{\sum_{k=1}^{m} M_j(k)} \tag{7}$$

where $F_j(k)$ and $M_j(k)$ represent the k-th frequency and DFT magnitude samples of the j-th frame, respectively. The DFT was calculated with length 1024.

3. Spectral rolloff point

It is the frequency value up to which 95% of signal energy resides[16]. We estimate this value for each frame using:

$$SRP_j = F_j(t) : \left[\sum_{k=1}^{t<m} M_j(k) = 0.95 \cdot M_j \right] \tag{8}$$

with $M_j = \sum_{k=1}^{m} M_j(k)$, where $M_j(k)$ is the magnitude of k-th sample of the DFT of the j-th frame. $F_j(t)$, with $t < m$, is the frequency that corresponds to the 95-th percentile of the energy of the DFT.

4. Bandwidth (spectral centroid range)

The bandwidth (BDW) represents, here, a frequency range around the spectral centroid, where the signal energy is concentrated[14]; it is estimated as follows:

$$BDW_j = \frac{1}{m} \sum_{k=1}^{m} M_j(k) \cdot (F_j(k) - SPC_j) \tag{9}$$

where SPC_j is the spectral centroid of the j-th frame.

5. Frame energy and segment energy

The frame energy represents the energy of each frame and, correspondingly, the segment energy. We use the following measures obtained in the frequency domain:

Frame energy:

$$FRE_j = \sum_{k=2}^{m} (M_j(k))^2 \tag{10}$$

where $(M_j(k))^2$ represents the squared magnitude of the DFT (with length m) of each frame.

Segment energy:

$$SGE = \sum_{k=2}^{m} (M(k))^2 \tag{11}$$

where $(M(k))^2$ represents the squared magnitude of the DFT (with length m) of each segment. In both equations, m is the smallest integer such that $m = 2^k \geq N$, with N the number of samples of frame and segment, respectively. Note that, now, the magnitude of the DFT sample at frequency 0 (offset) is not included in the measures. Segment energy represents a scalar magnitude and it is integrated in the vector of descriptors as it is, while the mean and variance are computed for the frame energy.

6. Fundamental frequency

The fundamental frequency, in our context, is related to pitch, although these can be considered different views of the frequency perception by humans and computers[24]. A way to robustly estimate the fundamental frequency is based on the autocorrelation function ACF[25], in this work, we proceed as follows. In each frame:

- Compute the ACF $y_j(i) = \sum_{k=0}^{n-1} x_j(k) \cdot x_j(k+i)$.
- Find the largest local maxima of $y_j(i)$, with $i > 0$.
- Obtain the frequency corresponding to the index selected:

$$F_j = F_j(k) \leftrightarrow \underset{i,i>0}{max}(y_j(i)) \tag{12}$$

Note that although the method is not completely robust against errors in the identification of the fundamental frequency (other harmonics can be taken as fundamental frequency using this technique), the method is valid for us since we do not really need the fundamental frequency, but a dominant frequency.

7. Salience of pitch

This feature quantifies the prominence of the fundamental frequency. We use the following expression:

$$SOP_j = \frac{\underset{i\geq 1}{local\ max}\left(\sum_{k=0}^{n-1} x_j(k)\cdot x_j(k+i)\right)}{\sum_{i=1}^{n}|x_j(i)|^2} \tag{13}$$

where the function $local\ max_{i\geq 1}(y(i))$ returns the largest $y(j)$, with $j \geq 1$, such that $y(j) \geq y(j-1)$ and $y(j) \geq y(j+1)$.

III. CLASSIFICATION OF THE VECTORS OF FEATURES

Although the discrimination between music and speech sometimes is done using heuristic tests with a few features at hand[9] or by direct thresholding[6], a more common practice is the usage of several features as inputs to a more elaborate classifier. Carey[4] uses a Gaussian mixture model GMM to model the classes of audio signal and the difference of the log-likelihood is used for classification. Saunders uses a multivariate Gaussian model[3] (the tests performed are not specified in this paper). A set of fuzzy rules is used in[2], a neural network is used in[7] using as inputs the first order statistics of the spectra of the audio segments employed; the k-Nearest Neighbor classifier is also considered. Goodwin and Laroche[8] turn their attention to the identification of cost functions for a dynamic programming procedure that uses linear discriminant analysis for dimensionality reduction. In our case, we initially deal with a large number of features, as shown, and we expect to use them in an efficient way in a multi-feature classifier. We choose the Fisher linear discriminator to deal with multiple features as a technique to reduce the problem of dimensionality[26] and a Gaussian model for

the projections of the vectors of features[27], then, the complementary error function is used to measure the probability that a certain sample is at least that distance away from the mode of the distribution, and the comparison of the probabilities for the different classes is used for the classification.

Now, we turn our attention to Fisher linear discriminant function[26]. Consider a n-dimensional vector of observations (audio features in our context), the Fisher linear discriminant functions is aimed to maximize the clustering of the two classes[28], maximizing the separation between them, i.e. maximizing the inter-class variance and minimizing the intra-class variance[26]. From a geometrical point of view, this is accomplished by means of a projection in a favorable direction $\overline{v}$. To this end, the function:

$$J\left(\overline{v}\right) = \frac{\left(\mu_m - \mu_s\right)^2}{S_m^2 + S_s^2} \tag{14}$$

must be maximized[26]. Where, in our specific context:

- $\overline{v}$ is the projecting vector.
- $\mu_{m,s} = \frac{1}{n_{m,s}} \sum_{i=1}^{n_{m,s}} p_i$ are the means of projected points p_i, for music (m) and speech (s) samples.
- $\tilde{S}_{m,s}^2 = \sum_{i=1}^{n_{m,s}} \left(p_i - \mu_i\right)^2$ are the so called scatters of projected points p_i, for music and speech samples.

In order to solve Eq. (14), let:

- $p = \overline{v}^t \overline{d}$, denote the projected form p of the vector of descriptors $\overline{d}$;
- $S_W = \sum_{i=1}^{n_m} \left(p_i - \mu_i\right) + \sum_{i=1}^{n_s} \left(p_i - \mu_i\right) = S_m + S_s$, is the sum of the music and speech scatter matrices, also known as *intra-class scatter matrix*;
- $S_B = \left(\mu_m - \mu_s\right) \cdot \left(\mu_m - \mu_s\right)^t$, represents the *inter-class scatter matrix*;

then, eq. (14) can be rewritten as:

$$J\left(\overline{v}\right) = \frac{v_t S_B v}{v_t S_W v}. \tag{15}$$

Then, the problem can be solved converting it into a well known generalized eigenvalue problem whose solution is given by[27]:

$$v = S_W^{-1} \left(\mu_m - \mu_s \right), \tag{16}$$

Ideally, two separated clouds of projected feature vectors should be found. However the features themselves behave as random-variables r.v's with unspecified probability density function (pdf) and, so, the projection of the vector of r.v's, as a sum of r.v's, can be taken to be a normal r.v., according to the central limit theorem (Stark and Woods[27] and Papoulis[29]). Then, the class membership of each projection is decided comparing the probability that such projection is at least that distance away from the mode of each class distribution. Thus, a sample p will be assigned to the class C_x with $x = m$ or s, for the music class and the speech class, respectively, using:

$$C_x = C_c : c \mid \underset{c=\{s,m\}}{max} \left(erfc \left\| \frac{p - \mu_c}{\sigma_c} \right\| \right) \tag{17}$$

Note that this is a monotonic function[30], so, from a computational point of view, we decide to use the argument for the classification. With this strategy, the computational load of the classifier is mainly due to the process of extraction of the selected features and, also, the number of parameters that the classifier needs to store is small.

Now, a number of features, which should attain the best possible classification behavior, with respect to the selected classifier, should be found. In the next section, we analyze the behavior of the features described previously, in relation with the selected classifier, to make the final decision about the features that will be finally used.

IV. PERFORMANCE ANALYSIS AND FINAL CLASSIFIER DESIGN

In order to check the performance of the features and to select the features that should be finally used in the classification, the samples given by Fundación Albéniz were manually cut into fragments of a few seconds of length, such that each of them contained a single type

of audio, either speech or music. Then, a large set of 1 second audio segments randomly selected was extracted for training and testing the classifier and the features. Initially, the classification performance of the isolated features was observed, then an optimal subset of features, for the training data set and classification strategy selected was found.

A. Performance test of single features. Simple grouping

The classification performance of each feature is evaluated. In Fig. 8 the music-to-speech and speech-to-music error rates are shown for an implementation of the classifier based on a single feature. Training and test audio samples are different and the complete set is about 1000 audio segments.

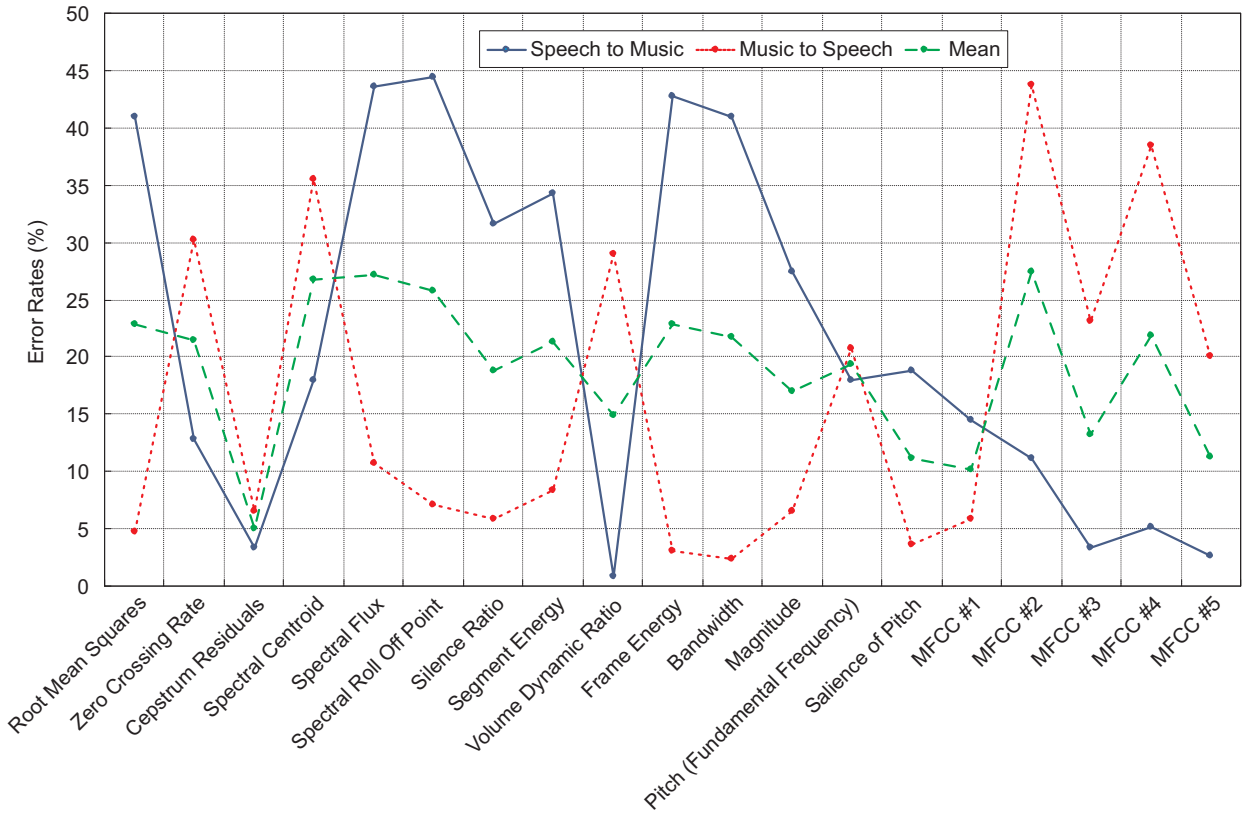

FIG. 8. (color online) Classification error rates. Speech-to-music (solid), music-to-speech (dotted) and mean values (dashed).

This figure shows that cepstrum residuals seems to be the best features for classification, achieving a mean error rate of 5% (assuming that speech and music are equally likely). Also, in the figure, it can be observed that the volume dynamic ratio shows the lowest speech-to-music error. With the data at hand, we turn to the main target of the work which is to find a subset of features that, combined in the selected classifier, attains the best classification performance with the classification strategy selected, based on the usage of the Fisher linear discriminant function. This search is motivated by the fact that a proper combination of favorable features should attain better classification performance than single features. For example, consider the 'Time domain' features (root mean squares, zero crossing rate, silence ratio and volume dynamic ratio), this subset attains 5.1% and 12.4% for speech-to-music and music-to-speech error rates, respectively, when combined using the ideas of the Fisher LDA, those features attain a global error rate of 8.8% versus a 14.5% of averaged global error rate of that subset of features. The remaining features achieve a better classification error rate, see Table I. So, we will search for a subset of features that attains the best possible performance with the classifier selected.

B. Selection of features and definition of the classifier

In order to select a subset of features that attains the best performance with the selected classification strategy, a recursive search has been used. As nineteen features are initially available, one could try to find the parameters of the classifier and evaluate its performance for all the possible combinations of features. Unfortunately, the number of subsets to evaluate[31] makes the full search computationally intractable. With the computational power at hand, we find that we can carefully evaluate subsets of up to 12 features, so, we test the performance of all the subsets of 12 features or less, with this, more than $1.8 \cdot 10^9$ combinations have been evaluated. Fig. 9 shows the evolution of the error rates of the best subset for a maximum number of features in the subset of $k = 1, 2, 3, \ldots, 12$.

Observe that Fig. 9 indicates that the performance of the best subset for $k > 8$ is almost

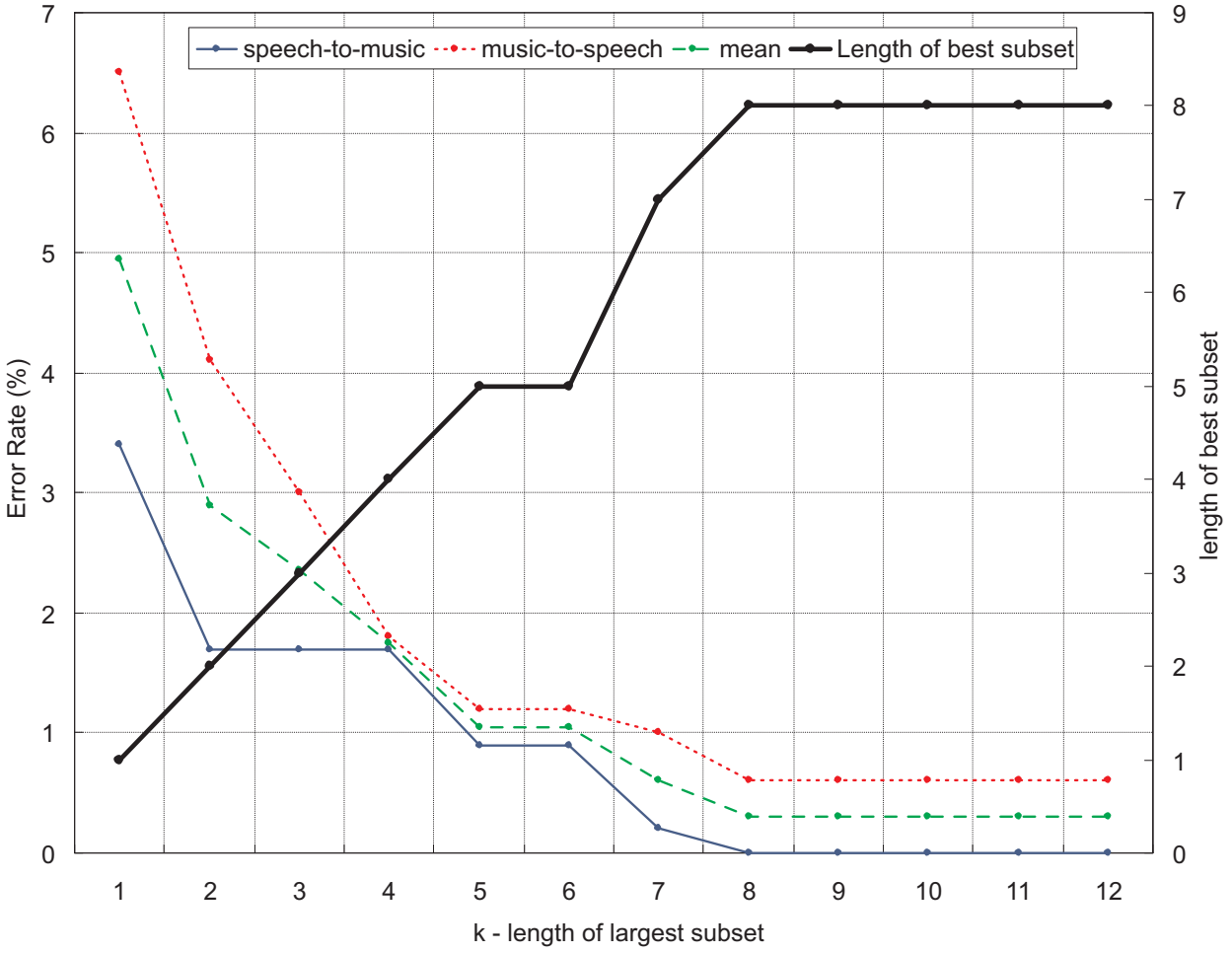

FIG. 9. (color online) Results of recursive classification using k varying length of subset. Speech-to-music, music-to-speech and mean error rates as shown with bold, dotted and dashed lines, respectively. The thick line shows the number of features of the best subset.

the same as for $k = 8$, discouraging the implementation of a costly recursive procedure for $k \gg 8$. The subset that requires the lowest computation time with the lowest number of features, that achieves the lowest error rate with the selected classification procedure, is composed by features shown in Table II. This subset of features, used in the Fisher classifier, with the Gaussian approximation of the projections of the classes, reaches a mean error rate of 0.3% (see Table II).

After finding this subset, we continue analyzing the performance of the classifier over

the subsets evaluated, paying attention to the subsets that attain the lowest error rates. Monitoring the best hundred subsets, we discover that the first 5 combinations attain the same mean error rate, about 0.3%.

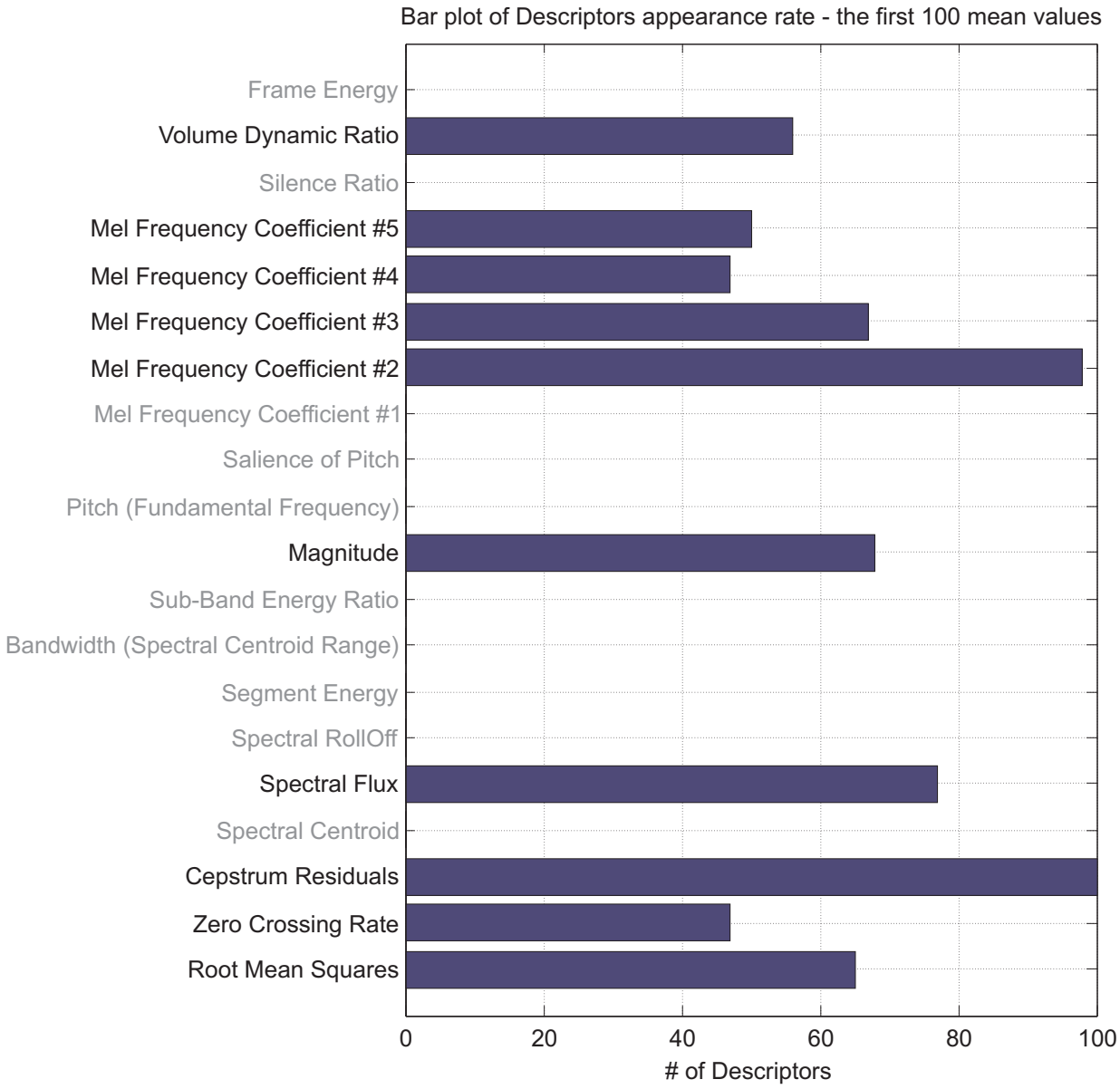

FIG. 10. (color online) Feature appearance on the best subsets for the Fisher linear classifier.

In Fig. 10, the appearance rate of all the features in this subset are shown. Note that no more than ten features are involved in these subsets, which constitute a computationally tractable subset, so, we select all these features for the final classification tool. This subset

attains a mean error rate of 0.3% over the test data set. Table III shows the best five subsets of eatures found by our iterative search, the features are represented by the acronyms defined in the section devoted to the description of each feature. In this table, subset 5 contains the features finally selected for the final discrimination application.

V. CONCLUSIONS AND DISCUSSION

In this paper, we have dealt with the problem of music-speech discrimination. We have evaluated the performance of a number of features, most of them taken from literature, using specific parameters driven from the requirements and the conditions of the data available, and, then, we have chosen a classification strategy based on Fisher linear discriminant analysis to define a classifier which is computationally tractable, in which, the computational load is due to the process of extraction of features. Also, the classifier selected does not require the storage of large sets of data or features to perform the classification, but only a small number of parameters.

With the selected classification strategy, an iterative search has been conducted to find a subset of features that attains the best possible classification performance with a tractable computational load. Ten out of nineteen features were selected on the basis of the analysis of the behavior of the features and their combinations in the selected classification strategy; thus, a fast classifier was developed that attained, on the training and evaluation data sets, mean error rates well below 1%.

Acknowledgments

This work has been funded by the Spanish *Ministerio de Industria, Turismo y Comercio* under project FIT-350201-2007-8 and by the Spanish *Ministerio de Educación y Ciencia* under project TSI-2007-61181.

The whole audio dataset used in this work was kindly provided by Fundación Albéniz.

References

[1] B.P. Bogert, M.J.R. Healy and J.W. Tukey, "The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo-autocovariance, Cross-Cepstrum, and Saphe Cracking", *John Wiley and Sons, New York*, 1963.

[2] R.M. Aarts and R.T. Dekkers, "A real-time speech-music discriminator", *J. Audio Eng. Soc.*, September 1999, 47, (9), pp. 720 - 725.

[3] J. Saunders, "Real-time discrimination of broadcast speech/music", *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '96*, May 1996, 2, pp. 993 - 996.

[4] M.J. Carey, E.S. Parris and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination", *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '99*, March 1999, 1, pp. 149 - 152

[5] E. Pampalk, "Computational Models of Music Similarity and their Application in Music Information Retrieval, PhD Thesis", *Vienna University of Technology, Vienna, Austria*, March 2006

[6] R. Jarina, N. Murphy, N. O'Connor and S. Marlow, "Speech-Music Discrimination from MPEG-1 Bitstream", *Kluev V.V and Mastorakis N.E. (Eds), Advances in Signal Processing, Robotics and Communications, WSES Press, 2001, SSIP'01 - WSES International Conference on Speech, Signal and image processing*, September 2001, pp. 174 - 178.

[7] H. Harb and L. Chen, "Robust speech music discrimination using spectrum's first order statistics and neural networks", *Signal Processing and Its Applications, 2003. proceedings. Seventh International Symposium on*, July 2003, 2, pp. 125 - 128.

[8] M.M. Goodwin and J. Laroche, "A dynamic programming approach to audio segmentation and speech / music discrimination", *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '04* May 2004, 4, pp. iv-309-iv-312

[9] C. Panagiotakis and G. Tziritas, "A speech/music Discriminator Based on RMS and

Zero-Crossings", *IEEE Transaction on Multimedia*, February 2005, 7, pp. 155 - 166.

[10] S.B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Trans Acoust. Speech and Signal Processing*, August 1980, 28, (4), pp 357 - 366.

[11] C. Lee, C. Chou, C. Han and R. Huang, "Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis", *Pattern Recognition Letters*, January 2006, 27, (2), pp. 93 - 101.

[12] G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signals", *IEEE Transactions on Speech and Audio Processing*, July 2002, 10, (5), pp. 293 - 302.

[13] C. Shahnaz, W.-P. Zhu and M.O. Ahmad, "A multifeature voiced/unvoiced decision algorithm for noisy speech", *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*, May 2006, pp. 2525 - 2528.

[14] M. Liu and C. Wan, "A Study on Content-Based Classification and Retrieval of Audio Database", *2001 International Database Engineering & Applications Symposium (IDEAS '01)*, 16-18 July 2001, pp. 339 - 345.

[15] B. Kedem, "Spectral analysis and discrimination by zero-crossings", *IEEE Proceedings*, 74, (11), November 1986, pp. 1477 - 1493.

[16] P. Arnaud, "Speech/Music Discriminator - Project Report", *Tampere University of Technology, Finland - http://www.cs.tut.fi/sgn/arg/arno* -, August 1999, last viewed 4/21/2009.

[17] E. Scheirer and M. Slaney, "Construction And Evaluation Of A Robust Multifeature Speech/music Discriminator", *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '97*, April 1997, 2, 1331.

[18] S.S. Stevens, J. Volkman, and E.B. Newman, "A scale for the measurement of the psychological magnitude pitch", *The Journal of the Acoustic Society of America*, January 1937, 8, pp. 185 - 190.

[19] H. Zhou, A. Sadka and R.M. Jiang, "Feature extraction for speech and music discrimination", *Content-Based Multimedia Indexing, 2008. CBMI 2008. International Workshop*

on, June 2008, pp. 170 - 173.

[20] W. Han, C. Chan, C. Choy and K. Pun, "An Efficient MFCC Extraction Method in Speech Recognition", *International Symposium on Circuits and Systems, Proceedings. 2006 IEEE*, May 2006, pp. 145 - 148.

[21] S. Sigurdsson, K.B. Petersen and T. Lehn-Schiøler, "Mel Frequency Cepstral Coefficients: An Evaluation of Robustness of MP3 Encoded music", *Proceedings of the Seventh International Conference on Music Information Retrieval (ISMIR)*, November 2006.

[22] M. Liu, C. Wan and L. Wang, "Content-based audio classification and retrieval using fuzzy logic system: towards multimedia search engines", *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, August 2002, 6 pp. 357 - 364.

[23] D. Hosseinzadeh and S. Krishnan, "Combining Vocal Source and MFCC Features for Enhanced Speaker Recognition Performance Using GMMs", *IEEE 9th Workshop on Multimedia Signal Processing, MMSP 2007*, October 2007, pp. 365 - 368.

[24] S. Suk, H. Chung and H. Kojima, "Voice/Non-Voice Classification Using Reliable Fundamental Frequency Estimator for Voice Activated Powered Wheelchair Control", *Lecture Notes in Computer Science. International Conference on Embedded Software and Systems*, July 2007, 4523, pp. 347 - 357.

[25] A. de Cheveigné and H. Kawahra, "YIN, a fundamental frequency estimator for speech and music", *Journal of the Acoustical Society of America*, April 2002, 111, (4), pp. 1917 - 1930.

[26] R.O. Duda, P.E. Hart and D.G. Stork, "Pattern Classification (2nd Edition)", *Wiley-Interscience*, New York, October 2000.

[27] H. Stark and J.W. Woods, "Probability and Random Processes with Applications to Signal Processing (3rd Edition)", *Prentice Hall*, New Jersey, August 2001.

[28] R.A. Fisher, "The statistical utilization of multiple measurements", *Annals of Eugenics*, 1938, 8, pp. 376 - 386.

[29] A. Papoulis, "Probability, Random Variables and Stochastic Processes (fourth ed.)", *McGraw-Hill*, New York, February 1991.

[30] M. Abramowitz and I.A. Stegun, "Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables", *Dover Publications*, New York,1972.

[31] L.J. Bain and M. Engelhardt, "Introduction to Probability and Mathematical Statistics (2nd Edition)", *Duxbury Press*, Pacific Grove (CA),March 2000.

TABLE I. Error rates for time and frequency domain subsets. Comparison among combined and averaged rates are shown.

Subset	Speech-to-Music	Music-to-speech	Average	
Time domain	5.1%	12.4%	8.8%	Combined
	21.6%	17.5%	19.5%	Single (mean)
Frequency domain	6.4%	3.0%	4.7%	Combined
	21.9%	15.7%	18.8%	Single (mean)

TABLE II. Error rates for each feature contributing to the optimal subset.

Feature	**Error Rate**		
	Speech as Music	**Music as Speech**	**Average**
Cepstrum Residuals	3.4%	6.5%	5.0%
Spectral Flux	43.6%	10.7%	27.2%
Magnitude	27.4%	6.5%	17.0%
Mel Frequency Coefficient #2	11.1%	43.8%	27.5%
Mel Frequency Coefficient #3	3.4%	23.1%	13.3%
Mel Frequency Coefficient #4	5.1%	38.5%	21.8%
Mel Frequency Coefficient #5	2.6%	20.1%	11.4%
Volume Dynamic Ratio	0.9%	29.0%	15.0%
Grouped features set	**0.0%**	**0.6%**	**0.3%**

TABLE III. Best subsets of features found for the classification.

Subset 1	**Subset 2**	**Subset 3**	**Subset 4**	**Subset 5**
CER	RMS	RMS	ZCR	RMS
SPF	ZCR	ZCR	CER	ZCR
MAG	CER	CER	SPF	CER
MFCC#2	SPF	SPF	MAG	SPF
MFCC#3	MAG	MFCC#2	MFCC#2	MAG
MFCC#4	MFCC#2	MFCC#3	MFCC#3	MFCC#2
MFCC#5	MFCC#3	MFCC#4	MFCC#4	MFCC#3
VDR	MFCC#4	MFCC#5	MFCC#5	MFCC#4
–	MFCC#5	VDR	VDR	MFCC#5
–	–	–	–	VDR

List of Figures