SPEECH/MUSIC DISCRIMINATION BASED ON A NEW WARPED LPC-BASED FEATURE AND LINEAR DISCRIMINANT ANALYSIS

J.E. Muñoz-Expósito, S. Garcia-Galán, N. Ruiz-Reyes, P. Vera-Candeas, F. Rivas-Peña

Electronics and Telecommunication Engineering Department. University of Jaén
Polytechnic School, Linares, Jaén, SPAIN
{jemunoz,sgalan,nicolas,pvera,rivas}@ujaen.es

ABSTRACT

Automatic discrimination of speech and music is an important tool in many multimedia applications. The paper presents a low complexity but effective approach, which exploits only one simple feature, called Warped LPC-based Spectral Centroid (WLPC-SC). Comparison between WLPC-SC and the classical features proposed in [9] is performed, aiming to assess the good discriminatory power of the proposed feature. The length of the vector for describing the proposed psychoacoustic based feature is reduced to a few statistical values (mean, variance and skewness), which are then transformed to a new feature space by applying LDA with the aim of increasing the classification accuracy percentage. The classification task is performed by applying SVM to the features in the transformed space. The classification results for different types of music and speech show the good discriminating power of the proposed approach.

1. INTRODUCTION

Automatic discrimination between speech and music has become a research topic of interest in the last few years. Several approaches have been described in the recent literature for different applications [1][2] [3][4][5]. Each of these uses different features and pattern classification techniques and describes results on different material.

Saunders [1] proposed a real-time speech/music discriminator, which was used to automatically monitor the audio content of FM audio channels. Four statistical features on the zero-crossing rate and one energy-related feature were extracted, a multivariate-Gaussian classifier was applied, which resulted in an accuracy of 98%.

In automatic speech recognition (ASR) of broadcast news, it's desirable to disable the input to the speech recognizer during the non-speech portion of the audio stream. Scheirer and Slaney [2] developed a speech/music discrimination system for ASR of audio sound tracks. Thirteen features to characterize distinct properties of speech and music, and three classification schemes (MAP Gaussian, GMM and k-NN classifiers) were exploited, resulting in an accuracy of over 90%.

Another application that can benefit from distinguishing speech from music is low bit-rate audio coding. Designing an universal coder to reproduce well both speech and music is the best approach. However, it is not a trivial problem. An alternative approach is to design a multi-mode coder that can accommodate different signals. The appropriate module is selected using the output of a speech-music classifier [6].

Automatic discrimination of speech and music is an important tool in many multimedia applications. Khaled El-Maleh et al. [3] combined the line spectral frequencies and zero-crossings-based features for frame-level narrowband speech/music discrimination. The classification system operates using only a frame delay of 20 ms, making it suitable for real-time multimedia applications. An emerging multimedia application is content-based indexing and retrieval of audiovisual data. Audio content analysis is an important task for such application [7].

Comparative view of the value of different types of features in speech music discrimination is provided in [8], where four types of features (amplitudes, cepstra, pitch and zero-crossings) are compared for discriminating speech and music signals. Experimental results showed cepstra and delta cepstra bring the best performance. Mel Frequencies Spectral or Cepstral Coefficients (MFSC or MFCC) are very often used features for audio classification tasks, providing quite good results. In [4], MFSC's first order statistics are combined with neural networks to form a speech music classifier that is able to generalize from a little amount of learning data. MFCC are a compact representation of the spectrum of an audio signal taking into account the nonlinear human perception of pitch, as described by the mel scale. They are one of the most used features in speech recognition and have recently proposed in musical genre classification of audio signals [9][10].

Unlike the previous works, speech/music discrimination approaches based on only one type of features are presented in [11] and [5], which result in fast and robust classification systems. The approach in [11] takes psychoacoustic knowledge into account in that it uses the low frequency modulation amplitudes over 20 critical bands to form a good discriminator for the task, while the approach in [5] exploits a new energy-related feature, called modified low energy ratio, that improves the results obtained with the classical low energy ratio.

In this paper, we present our contribution to the design of a robust speech/music discrimination system. The paper presents a low complexity but effective approach, which also exploits only one simple feature, called Warped LPC-based Spectral Centroid (WLPC-SC). The length of the vector for describing our psychoacoustic based feature is reduced to a few statistical values (mean, variance and skewness), which are then transformed to a new feature space by applying Linear Discriminant Analysis (LDA). LDA intends to provide a significant improvement in the classification accuracy percentage. The classification task is performed by applying Support Vector Machines (SVM) to the features in the transformed space.

2. SPEECH/MUSIC DISCRIMINATION

2.1. New Warped LPC-based feature

We propose the use of the centroid frequency each analysis window to discriminate between speech and music excerpts. Usually, speech signals has a low centroid frequency, which varies sharply at a voiced-unvoiced boundary. Instead, music signals show a quite changing behavior. There is no a specific pattern for such signals. We compute the centroid frequency by a one-pole lpc-filter. Geometrically, the lpc-filter minimizes the area between the frequency response of the filter and the energy espectrum of the signal. The one-pole frequency tells us where the lpc-filter is frequency-centered. Therefore, someway, the one-pole frequency informs us where most of the signal energy is frequency-localized.

However, the human auditory system is nonuniform in relation to the frequency. According to this statement, the Mel, the Bark and the ERB (Equivalent Rectangular-Bandwidth) scales [12] are defined for audio processing. For speech/music discrimination, it would be desirable to use a feature that works directly on some of these auditory scales, resulting in frequency-warped audio processing. The transformation from frequency to Bark scale is a well studied problem [12] [13]. Generally, the Bark scale is performed via the all-pass transformation defined by the substitution in the z domain

$$z = A_\rho(\zeta) \equiv \frac{\zeta + \rho}{1 + \zeta\rho} \qquad (1)$$

which takes the unit circle in the z plane to the unit circle in the ζ plane, in such a way that, for $0 < \rho < 1$, low frequencies are stretched and high frequencies are compressed. Parameter ρ depends on the sampling frequency of the original signal [13]. Applying (1), the Bark scale values can be approximated from frequency positions as follows [12]:

$$b = 13arctan(0.76f(kHz)) + 3.5arctan(\frac{f(kHz)}{7.5})^2 \qquad (2)$$

We propose the use of a one-pole warped-lpc filter based on this bilinear transformation to compute the WLPC-SC feature each analysis window.

As can be seen in Fig. 1, the WLPC-SC feature shows clear differences between voiced and unvoiced phonemes due to the frequency-warped processing. Besides, these differences are bigger than in a drum-based music signal. The results in Fig. 1 suggest us that WLPC-SC could be a profitable low complexity feature to design a robust music/speech discriminator. It will be assessed in section 3.

In our system, an *analysis window* of 23 ms (1024 samples at 44100 Hz sampling rate) and a *texture window* of 1 s (43 analysis windows) are defined. Overlapping with a hop size of 512 samples is performed. Hence, the vector for describing the proposed feature consists of 86 values, which are updated each 1 s-length texture window. This large dimensional feature vector is difficult to be handled for classification tasks. Therefore, it is required reducing the feature space to a few statistical values each 1 s-length texture window. Mean, variance and skewness of the feature vector are here computed. However, better classification results can be achieved if the statistical feature data are transformed or projected into a new feature space in which the classes are easier to be distinguished and a more robust decision rule can be found.

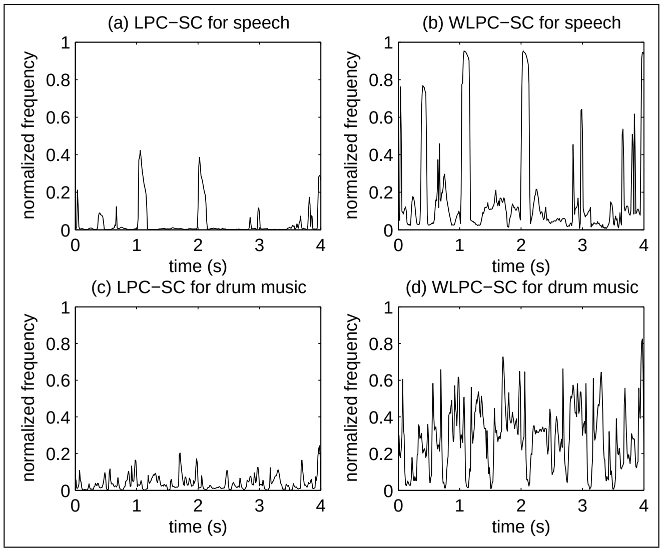

Figure 1: Example illustrating the values that LPC-SC and WLPC-SC takes for both speech and music signals.

2.2. Linear Discriminant Analysis (LDA)

There are two major techniques for transforming data from a feature space to another where separation between classes can be easier and more accurate done: Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). We have evaluated both techniques and the results reveal that LDA outperform PCA for speech/music discrimination (see section 3). PCA often fails in classification problems since the principal axes do not necessary entail discriminatory features. Nevertheless, LDA achieves significant improvement in class separation since it separates the class means while attempting to sphere the data classes. Next, LDA is briefly described.

LDA is one technique for transforming raw data into a new feature space in which classification can be carried out more robustly. Let us assume that a set of N samples $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N\} \in \mathbb{R}^n$ is given. Each sample belongs to one of M classes $\{C_1, C_2, \ldots, C_M\}$. The within-class scatter matrix $\mathbf{S}_w$ is defined as:

$$\mathbf{S}_w = \sum_{j=1}^{M} \sum_{i=1}^{n_j} (\mathbf{f}_i^j - \mu_j)(\mathbf{f}_i^j - \mu_j)^T \quad \mu_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbf{f}_i^j \qquad (3)$$

where $\mathbf{f}_i^j$ is the i-th sample of the j-th class, n_j the number of samples for the j-th class, and μ_j the mean of the j-th class.

The between-class scatter matrix $\mathbf{S}_b$ is also defined:

$$\mathbf{S}_b = \sum_{j=1}^{M} (\mu_j - \mu)(\mu_j - \mu)^T \quad \mu = \frac{1}{N} \sum_{i=1}^{N} \mathbf{f}_i \qquad (4)$$

where μ represent the mean of the entire data set.

LDA maximizes the between-class scatter to within-class scatter ratio, which involves the separation between classes is maximized while the variance within a class is minimized. The solution to this optimization problem is obtained by a Singular Value Decomposition (SVD) of $\mathbf{S}_w^{-1}\mathbf{S}_b$. If the columns of a matrix $\mathbf{W}$ are the eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_b$, LDA maps the original data set into a new feature space as:

$$\mathbf{a}_i = \mathbf{W}^T\mathbf{f}_i \tag{5}$$

In general, the dimensionality of the transformed feature set is one less than the number of training classes, which can further reduced by incorporating in $\mathbf{W}$ only those eigenvectors corresponding to the largest singular values determined in the scatter SVD.

An example of PCA and LDA transformations, taken from [14], is shown in figure 2. The original data consists of two classes, and the task is to find a projection onto one dimension which separates the classes.

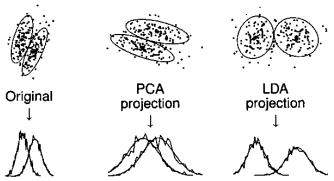

Figure 2: *A two-class example of PCA and LDA transformations.*

2.3. Classification by Support Vector Machines

SVM [15] have shown superb performance at binary classification tasks and handle large dimensional feature vectors better than other classification methods. Basically, a SVM aims at searching for an hyperplane that separates data points belonging to different classes with maximum margins.

Let's suppose a given training data set $\{(\mathbf{f}_1, \mathbf{y}_1), (\mathbf{f}_2, \mathbf{y}_2), \ldots, (\mathbf{f}_N, \mathbf{y}_N)\}, \mathbf{f}_i \in \mathbb{R}^n, \mathbf{y}_i \in (+1, -1)$, for $i = 1, \ldots, N$, with the aim of discriminating two classes (speech and music). If $\mathbf{y}_i = +1$, the data is music; otherwise, the data is speech. The SVM tries to determine an estimating function that allows to accurately classify a given data $\mathbf{f} \in \mathbb{R}^n$. The generic SVM estimating function is defined as:

$$\begin{array}{rcl} g(\mathbf{f}) & = & sign(p(\mathbf{f})) \\ p(\mathbf{f}) & = & \sum_{i=1}^{N} \mathbf{y}_i \alpha_i \Phi(\mathbf{f}_i, \mathbf{f}) + b \end{array} \tag{6}$$

where the weights α_i are obtained by maximizing the function

$$W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \mathbf{y}_i \mathbf{y}_j \Phi(\mathbf{f}_i, \mathbf{f}_j) \tag{7}$$

under the constrain

$$\sum_{i=1}^{N} \alpha_i \mathbf{y}_i = 0 \quad \forall i, \quad \alpha_i \geq 0 \tag{8}$$

The bias $b \in \mathbb{R}$ in (6) is obtained as:

$$b = -\frac{1}{2}(max(\mathbf{u}^T \mathbf{f}_i^+) + min(\mathbf{u}^T \mathbf{f}_i^-)) \tag{9}$$

where $\mathbf{f}_i^+$ and $\mathbf{f}_i^+$ are the training samples for which $\mathbf{y}_i = +1$ and $\mathbf{y}_i = -1$, respectively, and $\mathbf{u} = \sum_{i=1}^{N} \alpha_i \mathbf{y}_i \mathbf{f}_i$.

Finally, the function Φ that appears in (6) and (7) is known as kernel function. Among all the possibilities of kernel functions, we have chosen in this work Radial Basis Functions (RBF), which are defined as:

$$\Phi(\mathbf{f}_i, \mathbf{f}_j) = exp(-\frac{||\mathbf{f}_i - \mathbf{f}_j||}{2\sigma^2}) \tag{10}$$

where the parameter σ must be properly adjusted to ensure a good performance of the classifier.

Other modern classification techniques, such as Neural Networks (NN), fuzzy systems and dynamic programming, could also be used. We have decided to use SVM because it is one of the modern classification techniques that is achieving better results for classification purposes.

3. EXPERIMENT EVALUATION

First of all, the audio test database is carefully prepared. The speech data come from news programs of radio and TV stations, as well as dialogs in movies, and the languages involve English, Spanish, French and German with different levels of noise, especially in news programs. The speakers involve male and female with different ages. The length of the whole speech data is about an hour. The music consists of songs and instrumental music. The songs cover as more styles as posible, such as rock, pop, folk and funky, and they are sung by male and female in English and Spanish. The instrumental music we have chosen covers different instruments (piano, violin, cello, pipe, clarinet) and styles (symphonic music, chamber music, jazz, electronic music). Some music pieces in movies are also included, which are played by multiple different instruments. The length of the whole music data is also about an hour.

Next, we intend to assess the speech/music discrimination ability of the proposed feature. To achieve such goal, comparison with the timbral features proposed in [9] is performed. The WLPC-SC feature is separately compared to all timbral texture features proposed in [9]. The vector for describing our psychoacoustic based feature consist of the mean, the variance and the skewness over each texture window.

The following specific features are used in [9] to represent timbral texture: Spectral Centroid (SC), Spectral Rolloff (SR), Spectral Flux (SF), Time Domain Zero Crossings (ZC), Mel Frequency Cepstral Coefficients (MFCC) and Low Energy (LE) feature [9]. The last one (LE) is the only feature that is based on the texture window rather than the analysis window. Table 1 shows the classification accuracy percentage results when WLPC-SC is compared to the timbral features.

FEATURE	SPEECH (%)	MUSIC (%)	GLOBAL (%)
SC	91.67	97	94.33
SR	90.88	96.64	93.76
SF	81.77	96.44	89.10
ZC	91.28	90.82	91.05
MFCC	93.85	98.64	96.24
LE	90.29	88.45	89.37
WLPC-SC	95.07	94.36	94.71

Table 1: Classification accuracy percentage. WLPC-SC vs. timbral features

The results in table 1 are obtained by using a RFB-based SVM as classifier, which is properly trained and adjusted. At the sight of the results in table 1, we can say that the proposed feature performs better than most of the timbral features in [9] for speech/music discrimination. The Spectral Centroid (SC) and the Spectral Rolloff (SR) perform as well as the Warped LPC-based Spectral Centroid (WLPC-SC), while the Mel Frequency Cepstral Coefficients (MFCC) give slightly better accuracy percentages. The good discrimination ability provided by the SC, SR and MFCC features is achieved at the cost of a complexity increase regarding the WLPC-SC feature, which is much higher in the case of the MFCC feature.

Now, we are interested in knowing the improvement in the classification accuracy percentage due to PCA and LDA transformations. Table 2 shows the results obtained by the proposed approach (WLPC-SC + SVM) in three different cases: a) No transformation, b) PCA and c) LDA.

Case	Speech (%)	Music (%)	Global (%)
No transformation	95.07	94.36	94.71
PCA	95,77	95,21	95,49
LDA	96.56	95.95	96.25

Table 2: Classification accuracy percentage results when using: a) No transformation, b) PCA and c) LDA.

From the results in table 2, it can be said that feature space transformation is a very important operation for decreasing the misclassification rate. The best accuracy corresponds to LDA because it achieves a significant improvement in class separation.

Finally, we would like to assess the performance of a SVM-based classifier for speech/music discrimination when compared to the classical Gaussian Mixture Model (GMM) classifier. Table 3 shows the results obtained when a RBF-based SVM and a three-component GMM with diagonal covariance matrices are used as classifiers for speech/music discrimination. In both cases LDA has been used.

Case	Speech (%)	Music (%)	Global (%)
3-GMM	95.47	92.96	94.21
RBF-SVM	96.56	95.95	96.25

Table 3: Comparing RBF-SVM and 3-GMM for speech/music discrimination when LDA is used.

As expected, better results are obtained when the SVM-based classifier is used for speech/music discrimination. Our experiments show that the differences between SVM and GMM are shortened when LDA is not performed. It seems that LDA only achieves a slight improvement in the performance of the GMM classifier.

4. CONCLUSIONS

This paper presents a simple but robust approach to discriminate speech and music. The method exploits only one feature, called Warped LPC-based Spectral Centroid (WLPC-SC). The experiment evaluation compares the proposed feature to other features commonly used in audio classification tasks. The approach proposed in this paper implements LDA transformation in order to achieve a significant reduction in the misclassification ratio. Comparison with PCA and no feature space transformation is reported. The system is completed with a SVM-based classifier, which was compared to the classical GMM classifier. The classification accuracy percentage of the proposed approach is above 96% for a wide range of audio styles. The experiment results demonstrate the robustness of the system, doing its application scope very wide.

5. REFERENCES

[1] Saunders, J. ”Real-time discrimination of broacast speech/music”, *Proc. IEEE ICASSP'96*, Atlanta, USA, pp. 993-996, 1996.

[2] Scheirer, E. and Slaney, M. ”Construction and evaluation of a robust multifeature speech/music discriminator”, *Proc. IEEE ICASSP'97*, Munich, Germany, pp. 1331-1334, 1997.

[3] El-Maleh, K., Klein, M., Petrucci, G. and Kabal, P. ”Speech/music discrimination for multimedia applications”, *Proc. IEEE ICASSP'2000*, vol. 6, pp. 2445-2448, 2000.

[4] Harb, H. and Chen, L. ”Robust speech music discrimination using spectrum's first order statistics and neural networks ”, *Proc. IEEE Int. Symp. on Signal Processing and Its Applications*, vol. 2, pp. 125-128, 2003.

[5] Wang, W.Q., Gao, W., Ying, D.W. ”A fast and robust speech/music discrimination approach”, *Proc. 4th Pacific Rim Conference on Multimedia.*, vol. 3, pp. 1325-1329, 2003.

[6] Tancerel, L., Ragot, S., Ruoppila, V.T. and Lefebvre, R. ”Combined speech and audio coding by discrimination”, *Proc. IEEE Workshop on Speech Coding*, pp. 17-20, 2000.

[7] Zhang, T. and Kuo, J. ”Audio content analysis for online audiovisual data segmentation and classification”, *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 4, 2001.

[8] Carey, M.J., Parris, E.S. and Lloyd-Thomas, H. ”A comparison of features for speech, music discrimination”, *Proc. IEEE ICASSP'99*, Phoenix, USA, pp. 1432-1435, 1999.

[9] Tzanetakis, G. and Cook, P. ”Musical genre classification of audio signals”, *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 5, 2002.

[10] Burred, J.J. and Lerch, A. ”Hierarchical automatic audio signal classification”, *Journal of the Audio Eng. Soc.*, vol. 52, pp. 724-739, 2004.

[11] Karneback, S. ”Discrimination between speech and music based on a low frequency modulation feature”, *European Conf. on Speech Comm. and Technology*, Alborg, Denmark, pp. 1891-1894, 2001.

[12] Härmä, A., Karjalainen, M., Savioja, L., Välimäki, V., Laine, U.K. and Huopaniemi, J. ”Frequency-warped signal Processing for audio applications”, *Journal of the Audio Engineering Society*, vol. 48, no. 11, pp. 1011-1031, November 2000.

[13] Smith III, J.O. and Abel, J.S. ”Bark and ERB bilinear transforms”, *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 697-708, November 1999.

[14] Goodwin, M.M. and Laroche, J. ”Audio segmentation by feature-space clustering using linear discriminant analysis and dynamic programming”, *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.*, New York, USA, pp. 131-134, 2003.

[15] Vapnik, V.N. ”Statistical learning theory”, Wiley, New York, 1998.