SPEECH/MUSIC DISCRIMINATION USING A SINGLE WARPED LPC-BASED FEATURE

J.E. Muñoz-Expósito, S. Garcia-Galán, N. Ruiz-Reyes, P. Vera-Candeas and F. Rivas-Peña
Electronics and Telecommunication Engineering Department, University of Jaén
Polytechnic School, C/ Alfonso X el Sabio, 28
23700 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 for speech/music discrimination, which exploits only one simple feature, called Warped LPC-based Spectral Centroid (WLPC-SC). A three-component Gaussian Mixture Model (GMM) classifier is used because it showed a slightly better performance than other Statistical Pattern Recognition (SPR) classifiers. Comparison between WLPC-SC and the timbral features proposed in Tzanetakis and Cook (2002) is performed, aiming to assess the good discriminatory power of the proposed feature. Experimental results reveal that our speech/music discriminator is robust and fast, making it suitable for real-time multimedia applications.

Keywords: speech/music discrimination, LPC, spectral centroid, GMM.

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 (Saunders, 1996; Scheirer and Slaney, 1997; El-Maleh et al., 2000; Harb and Chen, 2003; Wang et al., 2003). Each of these uses different features and pattern classification techniques and describes results on different material.

Saunders (1996) 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 (1997) 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 as in ISO-IEC (1999) or Tancerel et al. (2000).

Automatic discrimination of speech and music is an important tool in many multimedia applications. El-Maleh et al. (2000) 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 (Zhang and Kuo, 2001). Minami et al. (1998) proposed an audio-based approach to video indexing, where a speech/music detector is used to help users to browse a video database.

Comparative view of the value of different types of features in speech music discrimination is provided in Carey et al. (1999), 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 Harb and Chen (2003), 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 (Tzanetakis and Cook, 2002; Burred and Lerch, 2004).

Unlike the previous works, speech/music discrimination approaches based on only one type of features are presented in Karneback (2001) and Wang et al. (2003), which result in fast and robust classification systems. The approach in Karneback (2001) 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 Wang et al. (2003) 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). This new feature is also psychoacoustic-based. Its simplicity and robustness make its application scope very wide, especially for the applications where low system cost is strongly demanded.

2 SPEECH/MUSIC DISCRIMINATION

2.1 New Warped LPC-based feature

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 85 values, which are updated each 1 s-length texture window. This large dimensional feature vector is difficult to be handled for classification tasks, giving rise to two main drawbacks: 1) too much computational cost, 2) possible too high misclassification rate. Therefore, it is required reducing the feature space to a few statistical values each 1 s-length texture window. In this work, the mean and variance of each feature vector are only computed.

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 (Härmä et al., 2000) 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 (Härmä et al., 2000; III and Abel, 1999). Generally, the Bark scale is performed via the allpass 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 (III and Abel, 1999). Applying (1), the Bark scale values can be approximated from frequency positions as follows (Härmä et al., 2000):

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

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.

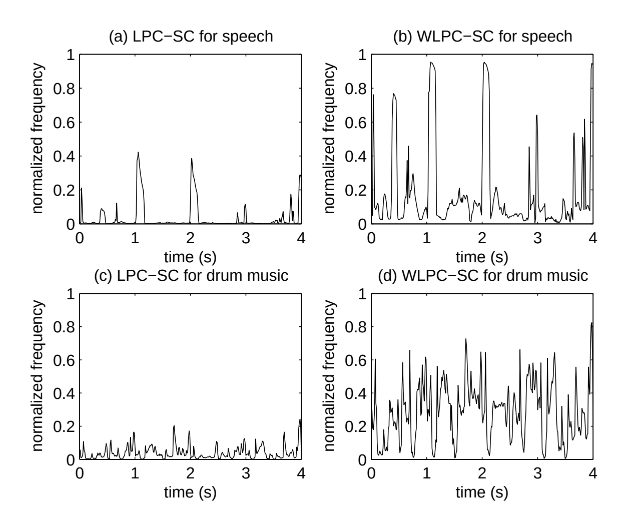

Figure 1: Example illustrating the values that LPC-SC and WLPC-SC takes for both speech and music signals.

2.2 Classification

For classification purposes, a number of standard Statistical Pattern Recognition (SPR) classifiers (Duda et al., 2000) were evaluated. The basic idea behind SPR is to estimate the probability density function (pdf) for the feature vectors of each class. In supervised learning a labeled training set is used to estimate the pdf of each class. In the simple Gaussian (GS) classifier, each pdf is assumed to be a multidimensional Gaussian distribution whose parameters are estimated using the training set. In the Gaussian Mixture Model (GMM) classifier, each class pdf is assumed to consist of a mixture of a specific number K of multidimensional Gaussian distributions. Unlike the k-NN classifier, which needs to store all the training feature vectors in order to compute the distances to the input feature vector, the GMM classifier only needs to store the set of estimated parameters for each class. The iterative EM algorithm can be used to estimate the parameters of each Gaussian component and the mixture weights.

In this work a three-component GMM classifier with diagonal covariance matrices is used because it showed a slightly better performance than other SPR classifiers. The performance of the system does not improve when using a higher number of components in the GMM classifier. The GMM classifier is initialized using the K-means algorithm with multiple random starting points. Modern classification techniques, such as Neural Networks (NN), Support Vector Machines (SVM), fuzzy systems and dynamic programming, could also be used. We decided to use standard SPR classifiers because this work is mainly focussed on the feature selection task for speech/music discrimination, proposing a new psychoacoustic-based feature (WLPC-SC), rather on the classification task.

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 capability of the proposed feature. To achieve such goal, comparison with the timbral features proposed in Tzanetakis and Cook (2002) is performed. The WLPC-SC feature is separately compared to all timbral texture features proposed in Tzanetakis and Cook (2002). The vector for describing our psychoacoustic based feature consist of the mean and variance over each texture window.

The following specific features are used in Tzanetakis and Cook (2002) 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 (Tzanetakis and Cook, 2002). 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.

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

FEATURE	SPEECH (%)	MUSIC (%)	GLOBAL (%)
SC	92.26	95.54	93.90
SR	95.24	90.18	92.60
SF	90.09	71.81	80.56
ZC	93.26	88.54	90.80
MFCC	92.08	99.20	95.83
LE	88.75	78.00	86.19
WLPC-SC	94.87	91.63	93.20

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 Tzanetakis and Cook (2002) for speech/music discrimination. The Spectral Centroid (SC) performs as well as the Warped LPC-based Spectral Centroid (WLPC-SC), while the Mel-Frequency Cepstral Coefficients (MFCC) give slightly better accuracy percentages than the same feature. The good discrimination capability provided by the SC 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. Note that the WLPC-SC feature does not require a DFT computation, while both SC and MFCC features need this computation. As shown in table 1, the proposed feature achieves high accuracy percentages while maintaining the complexity at a reduced degree.

Now, we are interested in comparing MFCC with all timbral features and all timbral features plus WLPC-SC. We intend to know if the proposed feature improves the classification accuracy percentage when it is added to all timbral features for speech/music discrimination. Table 2 shows how the inclusion of the WLPC-SC feature within the feature set used for speech/music discrimination entails a discrimination capability improvement. The classification accuracy percentage goes up a value of 2% up to reach the value of 97%. However, it must be noted that no improvement is accomplished when all timbral texture features in Tzanetakis and Cook (2002) are used for speech/music discrimination regarding the case of using only the MFCC feature.

Table 2: Discrimination capability improvement when the WLPC-SC feature is included within the feature set.

FEATURE	SPEECH (%)	MUSIC (%)	GLOBAL (%)
MFCC	92.08	99.20	95.83
All timbral features	93.27	97.27	95.36
All timbral features + WLPC-SC	94.86	99.45	97.28

Finally, we are interested in knowing how much warping transformation influences in speech/music discrimination. Table 3 compares the classification accuracy results for both the proposed feature (WLPC-SC) and the same feature without warping transformation (LPC-SC).

Table 3: Classification accuracy percentage. WLPC-SC vs. LPC-SC

FEATURE	SPEECH (%)	MUSIC (%)	GLOBAL (%)
WLPC-SC	94.87	91.63	93.20
LPC-SC	89.72	76.36	82.75

From the results in table 3, it can be said that warping transformation is a very important operation for the good performance of the feature proposed in this paper, because it entails psychoacoustic information is taken into account. Table 3 shows an improvement in the speech/music discrimination capability higher than 10% regarding the case of not using the warping transformation.

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). This new feature is is the main contribution of the paper. Its performance is assessed by different experimental tests. The classification stage is performed by a three-component GMM classifier with diagonal covariance matrices. We have proved that a higher number of components in the GMM classifier does not improve the performance of the system.

The experiment evaluation compares the proposed feature to other features commonly used in audio classification tasks. In particular, the timbral texture feature in Tzanetakis and Cook (2002) are used. The proposed feature performs better than most of the timbral features in Tzanetakis and Cook (2002) for speech/music discrimination with a reduced complexity. It is also assessed the improvement due to taking into account the nonlinear nature of the human earing system. The proposed feature (WLPC-SC) computed in the frequency domain (LPC-SC) seems to hide much potential. We achieve an improvement in the discrimination capability higher than 10% regarding the case of not using the warping transformation. The experiment results also demonstrate the robustness of the system. The classification accuracy percentage is higher than 93% with a wide range of styles. At the same time, its simplicity brings obvious advantages in constructing low cost systems.

References

J.J. Burred and A. Lerch. Hierarchical automatic audio signal classification. *Journal of the Audio Engineering Society*, 52:724–739, 2004.

M.J. Carey, E.S. Parris, and H. Lloyd-Thomas. A comparison of features for speech, music discrimination. *Proc. IEEE ICASSP'99*, pages 1432–1435, 1999.

R. Duda, P. Hart, and D. Stork. Pattern classification. *Wiley, New York*, 2000.

K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal. Speech/music discrimination for multimedia applications. *Proc. IEEE ICASSP'2000*, 6:2445–2448, 2000.

H. Harb and L. Chen. Robust speech music discrimination using spectrum's first order statistics and neural networks. *Proc. IEEE Int. Symp. on Signal Processing and Its Applications*, 2:125–128, 2003.

A. Härmä, M. Karjalainen, L. Savioja, V. Välimäki, U.K. Laine, and J. Huopaniemi. Frequency-warped signal processing for audio applications. *Journal of the Audio Engineering Society*, 48(11):1011–1031, 2000.

J.O Smith III and J.S. Abel. Bark and erb bilinear transforms. *IEEE Trans. Speech and Audio Processing*, 7: 697–708, 1999.

ISO-IEC. Mpeg-4 overview. *ISO/IEC JTC1/SC29/WG11 N2995 Document*, 1999.

S. Karneback. Discrimination between speech and music based on a low frequency modulation feature. *European Conf. on Speech Comm. and Technology*, pages 1891–1894, 2001.

K. Minami, A. Akutsu, H. Hamada, and Y. Tonomura. Video handling with music and speech detection. *IEEE Multimedia*, 5(3):17–25, 1998.

J. Saunders. Real-time discrimination of broacast speech/music. *Proc. IEEE ICASSP'96*, pages 993–996, 1996.

E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. *Proc. IEEE ICASSP'97*, pages 1331–1334, 1997.

L. Tancerel, S. Ragot, V.T. Ruoppila, and R. Lefebvre. Combined speech and audio coding by discrimination. *Proc. IEEE Workshop on Speech Coding*, pages 17–20, 2000.

G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Trans. on Speech and Audio Processing*, 10(5), 2002.

W.Q. Wang, W. Gao, and D.W. Ying. A fast and robust speech/music discrimination approach. *Proc. 4th Pacific Rim Conference on Multimedia*, 3:1325–1329, 2003.

T. Zhang and J. Kuo. Audio content analysis for online audiovisual data segmentation and classification. *IEEE Trans. on Speech and Audio Processing*, 9(4), 2001.