Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (Istanbul), pp. 2445–2448, June 2000

SPEECH/MUSIC DISCRIMINATION FOR MULTIMEDIA APPLICATIONS

Khaled El-Maleh *Mark Klein* *Grace Petrucci* *Peter Kabal*

Dept. Electrical & Computer Engineering
McGill University, Montreal, Quebec H3A 2A7, Canada

ABSTRACT

Automatic discrimination of speech and music is an important tool in many multimedia applications. Previous work has focused on using long-term features such as differential parameters, variances, and time-averages of spectral parameters. These classifiers use features estimated over windows of 0.5–5 seconds, and are relatively complex. In this paper, we present our results of combining the line spectral frequencies (LSFs) and zero-crossing-based features for frame-level narrowband speech/music discrimination. Our classification results for different types of music and speech show the good discriminating power of these features. Our classification algorithms operate using only a frame delay of 20 ms, making them suitable for real-time multimedia applications.

1 INTRODUCTION

A human listener can discriminate easily between speech and music signals by listening to a short segment (i.e., few seconds) of an audio signal. In recent years, different systems have been proposed for the automatic discrimination of speech signals and music signals. Saunders [1] proposed a real-time speech/music discriminator to be used in radio receivers for the automatic monitoring of the audio content of FM radio channels. In automatic speech recognition (ASR) of broadcast news, it is important to disable the speech recognizer during the non-speech portion of the audio stream. Recently, Scheirer and Slaney [2] and Williams and Ellis [3] developed and evaluated different speech/music discrimination systems for ASR of audio sound tracks.

Another application that can benefit from distinguishing speech from music is low bit-rate audio coding. Traditionally, separate codec designs are used to digitally encode speech and music signals. Generally, speech coders do better on speech, and audio coders do better on music [4]. In many emerging multimedia applications such as the Internet, the sound stream carries both speech and music. Designing a universal coder to reproduce well both speech and music is the best approach—however, this 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. This approach has been already employed in the parametric coder of the MPEG-4 standard [5] and recently in the multi-mode audio coder proposed by Ramprashad [4], and in [6] for mixed wideband speech and music coding.

Web page: `http://WWW.TSP.ECE.McGill.CA`.

An emerging multimedia application is a content-based audio and video retrieval. Audio classification is an important part of such systems. Automatic classification would remove the subjectivity inherent in the classification process and ultimately speed up the retrieval process. Zhang and Kuo [7] developed a content-based audio retrieval system that classifies audio signals as speech or music or noise. Minami *et al.* [8] proposed an audio-based approach to video indexing. A speech/music detector is used to help users to browse a video database.

Existing speech/music classification systems use long-term features such as variances and time-averages of spectral parameters [1], [2]. Tonality and pitch have also been combined into several designs [9]. Typically, these features are estimated over audio segments of 0.5–5 seconds. While these classifiers show high accuracy in distinguishing speech and music, they are not suitable for delay-sensitive applications such as interactive communications.

In this paper, we present our contribution to the design of a robust narrowband speech/music discrimination system. We propose new sets of classification features and we assess their discrimination properties. We demonstrate that by combining the line spectral frequencies (LSFs) and zero-crossing-based features we get good classification results using short audio segments. Unlike other designs, our classification system operates using only a frame delay of 20 ms, making it suitable for real-time applications with low computational complexity constraints.

2 FRAME-LEVEL CLASSIFICATION

2.1 Classification Features

A major step in the design of a signal classification system is the selection of a "good" set of features that are capable of separating the signals in the feature space. The choice of classification features is usually based on *a priori* knowledge of the nature of the signals to be classified. Features that capture the temporal and spectral structure of the input signal are often used.

Different features have been suggested in previous speech/music classification work. A common set of features includes zero-crossings information [1], energy, pitch, and spectral parameters such as cepstral coefficients [2] and [9]. In this paper we have selected the Line Spectral Frequencies (LSFs) as the core feature set for speech/music classification. This was motivated by our recent work in using the LSFs for classifying different types of background acoustic noises and speech [10].

The LSFs parameters are another transformation of lin-

ear prediction (LP) coefficients that have been successfully used in speech coding applications. Recently, Tourneret [11] has studied the statistical properties of the LSFs and discussed their merits for pattern recognition problems.

The LP coefficients were calculated using a 10th order narrowband (8 kHz sampling frequency) linear prediction analysis performed on a frame-by-frame basis (frame length is 20 ms) using the auto-correlation method. A Hamming window of 240 samples was used. The LP coefficients were calculated using the Levinson-Durbin algorithm and then bandwidth expanded using a factor of ($\gamma = 0.994$). The LP coefficients were then transformed to the LSF domain.

Differential Line Spectral Frequencies (DLSFs) are defined by taking successive differences of the LSFs. Small differentials indicate that the energy peaks are tightly packed while larger values can be interpreted as a broader distribution. We have included this LSF-based features as a means to capture the fine spectral variations between speech and music.

The zero-crossing count (ZCC) of a waveform can be a useful measure of the spectral centroid of a signal. Zero-crossings features alone can not effectively discriminate between speech and music [9]. Kedem in his book [12] has extended the theory of zero-crossings of a signal to include zero-crossing counts of filtered versions of the input signal. The ZCCs of these filtered signals are known as higher order crossings (HOC). He gives different examples where the HOC measurements can give more spectral information about the signal dynamics than the signal ZCC. In our experiments, we used high-pass differentiator filters and measured the ZCC of the filter-output signals. Filtering was limited to six stages as the discriminatory powers of HOCs decrease with successive differencing.

In this paper, we propose a new feature, the linear prediction zero-crossing ratio (LP-ZCR). It is defined as the ratio of the ZCC of the input and the ZCC of the output of the LP analysis filter. In LP analysis, the output signal (the LP residual) is a decorrelated version of the input signal and thus will have a higher ZCC. The LP-ZCR takes values between zero and one. It quantifies the correlation structure of the input sound. For example, a highly correlated sound such as voiced speech will have a low LP-ZCR, while unvoiced speech will have a value above 0.5. For a white noise the LP-ZCR is ideally one.

2.2 Classification Algorithms

For this study we have selected two different classification algorithms: a quadratic Gaussian classifier (QGC) and a nearest neighbor (NN) classifier. This selection of classification algorithms will enable us to compare our results with existing speech/music classifiers and will highlight the effect of the classification algorithm on the classification results.

A Gaussian classifier is based on the assumption that feature vectors of each class obey a multivariate Gaussian distribution. Estimates of the parameters of the Gaussian PDF of each class (mean and covariance) using the labeled training data are computed. In the classification stage, an input vector is mapped to the class with the largest likelihood [13]. In nearest neighbor classifiers, for each input feature vector, a search is done to find the label of the vector in the dictionary of stored training vectors with the minimum distance. Euclidean distance is commonly used as the metric to measure neighborhood. In k-NN decision rule, the input feature vector is assigned the label most frequently represented among the k nearest patterns in the training dictionary. One major disadvantage of NN classifiers is the need to store large number of training vectors resulting in a large amount of computations [13].

3 EVALUATION EXPERIMENTS

The training data consisted of both music and speech audio recordings with 8 kHz sampling frequency. The speech data originated from ten speakers, five males and five females. Music was selected from various categories including *classical, instrumental, opera, rock, dance, rap and pop.* The training vectors correspond to 28 000 frames (i.e., 9.3 minutes) for speech and 32 000 frames (i.e., 10.7 minutes) for music. Additional music and speech samples were set aside for independent testing. The test music vectors were chosen from the same categories as the training set and the test speech vectors were taken from an *InGroup* which were speakers used in the training set and from an *OutGroup*, speakers who were not.

The Bayes error rate and empirical error estimation were used to gauge the performance of the features and classifiers prior to independent testing. The Bayes error rate provides a measure of the discriminating power of the various groups of features. A lower bound on the Bayes error rate P_{Bayes} is a function of the asymptotic error rate of the nearest neighbor decision rule P_{NN} [10], given as

$$P_{Bayes} \geq \frac{M-1}{M}\left(1 - \sqrt{1 - \frac{M}{M-1}P_{NN}}\right) \qquad (1)$$

where M is the number of classes (i.e., $M = 2$ in our case).

Empirical error estimation gives an indication of the performance of a classifier with a given feature set. This error rate was estimated by using the Hold-out method [13]. This technique involves first dividing the data set into two parts, a training set and a testing set. Vectors for each set are chosen at random. The classifier is then induced using the training set and testing is carried out with the remaining samples. After several iterations, the mean error rate is calculated to give estimates of the average error rate.

4 CLASSIFICATION RESULTS

Four feature sets were used for experimentation. They were line spectral frequencies (LSF), differential line spectral frequencies (DLSF), line spectral frequencies with higher order crossings (LSF-HOC) and line spectral frequencies with LP-ZCR (LSF-ZCR). Table 1 contains the Bayes error rate for the aforementioned feature sets and error estimations using the k-NN and quadratic Gaussian classifiers.

Several observations can be made from the results in Table 1. First, the LSFs when used alone, have a low Bayes

Table 1 Error estimation for the classification features

Features	Bayes Rate (%)	Error (%)		
		1-NN	3-NN	QGC
LSF	4.6	8.8	9.6	21.1
LSF-ZCR	5.9	11.2	9.9	18.0
LSF-HOC	6.7	12.6	11.4	18.7
DLSF	7.3	13.5	13.9	23.3

error rate showing that they have the potential to effectively discriminate music from speech. Second, the error rates from using a Gaussian classifier and NN classifiers demonstrate the effect of classification algorithms on the results. The QGC classifier has error rate that is around 16% more than the Bayes rate. This can be explained by our observation that the LSFs features of speech and music deviate from the Gaussianity assumption. Combining zero-crossings features with the LSFs slightly improves the error rate for the QGC by reducing the overlapping of the feature spaces of music and speech.

The nearest neighbor classifiers have a superior holdout rate, approaching the Bayes error rate. This would indicate that they are a better choice than the Gaussian classifier. However, hold-out estimations for k-NN classifiers tend to be biased due to the large frame-to-frame correlation within the training set. Only independent testing gives a true measure of a classifier's performance. Moreover, the computational complexity and memory requirements of NN classifiers make them not practical for implementation. As a remedy to this problem, only prototype vectors from the training data can be computed and stored (i.e., prototype nearest-neighbor classification) [13].

Table 2 Accuracy (%) testing results for different music types (QGC)

music	LSF	DLSF	LSF-ZCR	LSF-HOC
Classical	93.6	93.2	96.2	89.5
Instrumental	79.3	80.3	92.8	90.1
Opera	77.3	76.3	73.7	53.9
Rock	72.1	69.4	87.6	84.4
Dance	68.2	59.5	86.6	83.9
Rap	60.7	54.6	80.9	77.1
Pop	57.8	57.3	84.3	82.4
Average	72.7	70.1	86.0	80.2

4.1 Music Test Results

Table 2 shows the QGC results of independent testing on different categories of music. We can observe that the accuracy rate depends on the music type. For example, using the LSFs features alone, *Classical* music is 93.6% detected as music while *Rap* music is detected as music 60.7% of the time. This could be attributed to the speech-music content of each music type. *Classical* music tends to be devoid of any speech content while *Rap* is dominated by rhythmic speech. Clearly, a large speech content will result in music being classified as speech. Combining additional features with the LSFs improved the decision accuracy with the troublesome categories. For instance, a gain of 20% in accuracy has been scored for the *Rap* music by combining the LP-ZCR feature with the LSFs. More than a 13% average increase in accuracy has been obtained for all types of music by using the LSF-ZCR feature set.

Table 3 Accuracy (%) results for music using the LSF features

music	QGC	3-NN
Classical	93.6	92.3
Instrumental	79.3	76.6
Opera	77.3	94.6
Rock	72.1	78.3
Dance	68.2	87.2
Rap	60.7	52.7
Pop	57.8	72.9
Average	72.7	79.2

In Table 3 we compare the music testing results from the Gaussian and the 3-NN classifiers. In general, the 3-NN has a better discrimination of music than the QGC. For example, *Opera* music was 94.6% accurately identified as music using the 3-NN, compared to 77.3% using the QGC. This shows that the estimated parameters of the Gaussian classifier are not capable of completely covering the large variations in music feature space.

Table 4 Accuracy (%) testing results for speech (QGC)

speech	LSF	DLSF	LSF-ZCR	LSF-HOC
InGroup	75.6	71.8	74.0	78.3
OutGroup	72.9	68.8	70.6	73.9
Average	74.3	70.3	72.3	76.1

4.2 Speech Test Results

The results of independent testing on speech using the quadratic Gaussian classifier are shown in Table 4. *InGroup* speech was classified with slightly better accuracy than *OutGroup* speech. This could be attributed to the LSFs tendency to model the vocal tract of the speaker. Generally, *InGroup* speech will always have a higher probability of being classified correctly. Table 5 shows that the 3-NN also outperforms the QGC by about 8% in distinguishing speech using the LSFs features.

Table 5 Accuracy (%) results for speech using the LSF features

speech	QGC	3-NN
InGroup	75.6	84.3
OutGroup	72.9	80.6
Average	74.3	82.5

4.3 Segment-level Classification

To make a fair comparison with previous speech/music classifiers, the quadratic Gaussian classifier was modified to make decisions over 50 frames (1 second). This was done by first making decisions for the individual frames. A global decision was then made for the entire block by choosing the class that appeared most frequently. By incorporating 50 frames of information into one decision, rather than one

frame per decision, the accuracy of the classifier rises noticeably.

Table 6 Accuracy (%) results with decisions made over 50 frames (QGC)

input	LSF	DLSF	LSF-ZCR	LSF-HOC
Speech	93.8	96.8	95.2	100.0
Music	87.5	80.0	94.4	91.9
Average	90.7	88.4	94.8	95.9

As depicted in Tables 6 and 7, the performance over 50 frames compares favorably with the accuracy rates of Scheirer and Slaney's speech/music classifier [2] and the discriminator described in the MPEG-4 standard [5]. The accuracy rating for the Scheirer and Slaney classifier was obtained from their original testing using all of their 13 proposed features. The performance of the discriminator described in the MPEG-4 standard was measured through independent testing. The testing set for the LSF-based classifier was reused for the MPEG-4 testing. The reference software provided by the MPEG committee was used to classify the testing set.

Table 7 Accuracy results for two other speech/music discriminators

Classifier	Accuracy (%)
MPEG-4	97.6
Scheirer and Slaney	93.2

4.4 Post-Decision Processing

To improve the performance accuracy of the speech/music discriminator, post-decision processing was implemented. A secondary goal of post-decision processing was to minimize switching between the two classes. In the case of an audio coder, the overhead associated with switching modes would increase encoding time dramatically. An error correction scheme with delay and without delay were used.

- **Error correction with delay**

  At a given instant, the immediate preceding and proceeding decisions are noted. A majority-logic rule is then used to select either speech or music. To make the proceeding decisions available, a delay of one frame is required.

- **Error correction without delay**

  The current and the preceding two decision are used for decision processing of the present frame. To prevent error propagation, past decisions are reset after 15 frames.

These error correction schemes realized a 5%–10% improvement in performance when the accuracy was above 50%. Below this threshold, the error corrections schemes produced more errors than they corrected.

5 CONCLUSION

We have presented a narrowband frame-level speech/music discrimination system that requires only a frame delay of 20 ms. Line spectral frequencies have shown the potential to discriminate the spectral structure of speech and music signals. Additional features have been combined with the LSFs to increase the classification performance. This paper has examined two-way classification. To better accommodate mixed signals and spoken words with music (*Rap* music, for example), a three-way classifier (speech, music-only, and music with speech) may be used.

References

[1] J. Saunders, "Real-time discrimination of broadcast speech/music," *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing* (Atlanta, GA), pp. 993–996, May 1996.

[2] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing* (Munich, Germany), pp. 1331–1334, Apr. 1997.

[3] G. Williams and D. Ellis, "Speech/music discrimination based on posterior probability features," *Proc. European Conf. on Speech Commun. and Technology* (Budapest, Hungary), pp. 687–690, Sept. 1999.

[4] S. A. Ramprashad, "A multimode transform predictive coder (MTPC) for speech and audio," *Proc. IEEE Workshop on Speech Coding for Telecom.* (Porvoo, Finalnd), pp. 10–12, June 1999.

[5] ISO-IEC, *MPEG-4 Overview (ISO/IEC JTC1/-SC29/WG11 N2995 Document)*, Oct. 1999. `http://www.cselt.it/mpeg/standards/mpeg-4/`.

[6] R.-Y. Qiao, "Mixed wideband speech and music coding using a speech/music discriminator," *Proc. 1997 IEEE TENCON*, pp. 605–608, 1997.

[7] T. Zhang and C.-C. J. Kuo, "Hierarchical classification of audio data for archiving and retrieving," *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing* (Phoenix, AZ), pp. 3001–3004, Mar. 1999.

[8] K. Minami, A. Akutsu, H. Hamada, and Y. Tonomura, "Video handling with music and speech detection," *IEEE Multimedia*, vol. 5, no. 3, pp. 17–25, 1998.

[9] M. J. Carey, E. S. Parris, and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination," *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing* (Phoenix, AZ), Mar. 1999.

[10] K. El-Maleh, A. Samouelian, and P. Kabal, "Frame-level noise classification in mobile environments," *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing* (Phoenix, AZ), pp. 237–240, Mar. 1999.

[11] J.-Y. Tourneret, "Statistical properties of line spectrum pairs," *Signal Processing*, vol. 65, pp. 239–255, Mar. 1998.

[12] B. Kedem, *Time Series Analysis by Higher Order Crossings*. IEEE Press, 1994.

[13] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. Academic Press, 1999.