Audio Engineering Society

Convention Paper 6678

Presented at the 120th Convention
2006 May 20–23 Paris, France

Application of Fisher Linear Discriminant Analysis to Speech/Music Classification

Enrique Alexandre[1], Manuel Rosa[1], Lucas Cuadra[1], and Roberto Gil-Pita[1]

[1] *Departamento de Teoría de la Señal y Comunicaciones. Universidad de Alcalá. 28805 - Alcalá de Henares, Madrid, Spain*

Correspondence should be addressed to Enrique Alexandre (enrique.alexandre@uah.es)

ABSTRACT

This paper proposes the application of Fisher linear discriminants to the problem of speech/music classification. Fisher linear discriminants can classify between two different classes, and are based on the calculation of some kind of centroid for the training data corresponding with each of these classes. Based on that information, a linear boundary is established, which will be used for the classification process. Some results will be given demonstrating the superior behavior of this classification algorithm compared with the well-known K-nearest neighbor algorithm. It will also be demonstrated that it is possible to obtain very good results in terms of probability of error by using only one feature extracted from the audio signal, being thus possible to reduce the complexity of this kind of systems in order to implement them in real-time.

1. INTRODUCTION

This work presents some preliminary results of the use of a particular kind of classifier, a Fisher linear discriminant [1], for the problem of speech/music classification. This kind of classifiers is widely used in some other applications such as face recognition systems [2][3]. The results will be compared with those obtained using a K-nearest neighbor algorithm (K-NN), since this particular classifier has been widely used in many applications [4][5], and presents a very good behavior compared to other algorithms like those based on gaussian mixture models [6].

Our work will be focused on the particular problem of speech/audio classification, which is usually a first step in many musical genre classification systems. This problem has attracted a large amount of research efforts from the early work of Saunders [7], where a simple thresholding of the average zero-crossing rate and energy features was used. A different approach was proposed by Scheirer and Slaney

[†]This work has been partially financed by the Universidad de Alcalá (UAH PI2005/081) and Comunidad de Madrid (CAM-UAH2005/036).

in [8], with the use of multiple features and statistical pattern recognition classifiers. Another approach was proposed in [9], where a speech/music discriminator with a false alarm probability virtually zero is described, although it was not tested with noisy signals. In [10] the use of the Line Spectral Frequencies (LSFs) is proposed, providing a low-delay algorithm with very good results using a quadratic gaussian classifier and a nearest neighbor classifier.

The paper will be structured as follows: first, a brief introduction of the features and classifiers employed in this work will be given. It has been decided to use a very simple set of features at this early stage of the work, since the objective is to compare the performance of the classification algorithm. Then, the results obtained for speech/music discrimination task will be shown and discussed.

2. FEATURE EXTRACTION

The objective of the feature extraction process is to obtain a compact numerical representation that can be used to characterize a segment of audio. A large number of features has been proposed in the literature for the speech/music classification problem, some of them inherited from the speech recognition area [11]. These features can be usually classified into three different classes: timbre-related, rhythm-related and pitch-related. In this work, since the objective is to compare the classification algorithms, and to simplify the problem, only timbre-related features will be used. For the feature extraction, a 512-samples window is used, with no overlap between adjacent frames. The time-frequency decomposition is performed using either a Modified Discrete Cosine Transform (MDCT), or a Discrete Fourier Transform (DFT), as it will be commented below. For each of these frames, all the features are calculated and their mean and standard deviation are computed every 43 frames (1.85 seconds at our sampling rate). Thus a 2-dimensional vector, containing the mean and standard deviation computed every 43 frames, is obtained.

All the features considered in this work will be now briefly described, although more detailed descriptions can be found on [8], [12] or [13].

2.1. Spectral centroid

The spectral centroid can be associated with the measure of brightness of a sound, and is obtained by evaluating the center of gravity of the spectrum:

$$Centroid_t = \frac{\sum_{k=1}^{N} |X_t[k]| \cdot k}{\sum_{k=1}^{N} |X_t[k]|} \quad (1)$$

where $X_t[k]$ represents the k-th frequency bin of the spectrum at frame t, and N is the number of samples.

2.2. Spectral roll-off

The spectral roll-off is usually defined as the frequency, $RollOff_t$, below which a PR% of the magnitude distribution is concentrated:

$$\sum_{k=1}^{RollOff_t} |X_t[k]| = PR \cdot \sum_{k=1}^{N]} |X_t[k]| \quad (2)$$

A typical value for PR is PR=85%. The spectral roll-off can give an idea of the shape of the spectrum.

2.3. Zero Crossing Rate

The Zero Crossing Rate (ZCR) is computed from the temporal signal $x[n]$ using the expression:

$$ZCR_t = \frac{1}{2} \sum_{n=1}^{N} |sign(x[n]) - sign(x[n-1])| \quad (3)$$

Where $sign(\cdot)$ represents the sign function, which returns 1 for positive arguments and -1 for negative ones. This parameter gives an idea of how noisy a signal is.

2.4. High Zero Crossing Rate Ratio

This feature, proposed in [14], is computed from the previously-defined ZCR, and is defined as the number of frames whose ZCR is 1.5 times above the mean ZCR on a window containing M frames. Mathematically:

$$HZCRR = \frac{1}{2M} \sum_{t=0}^{M-1} [sign(ZCR_t - 1.5 \cdot avZCR) + 1] \quad (4)$$

where $avZCR$ is the mean ZCR within the M-frames window. It can be demonstrated [14] that the HZCRR takes higher values for speech than for music since speech is usually composed by alternating voiced and unvoiced fragments, while music does not follow this structure.

2.5. Short Time Energy

The Short-Time Energy (STE) is defined as the mean energy of the signal within each analysis frame (N samples):

$$STE_t = \frac{1}{N}\sum_{k=0}^{N-1}|X_t[k]|^2 \tag{5}$$

2.6. Low Short-Time Energy Ratio

Similarly to the HZCRR, the LSTER is obtained from the STE, and defined as the ratio of frames whose STE is 0.5 times below the mean STE on a window that contains M frames. Mathematically,

$$LSTER = \frac{1}{2M}\sum_{t=0}^{M-1}[sign(0.5avSTE - STE_t) + 1] \tag{6}$$

2.7. Mel-frequency cepstral coefficients

Mel-frequency Cepstral Coefficients (MFCC) are a set of perceptual parameters calculated from the STFT [11] that have been widely used in speech recognition. They provide a compact representation of the spectral envelope, such that most of the signal energy is concentrated in the first coefficients. The application of these parameters for music modeling was discussed by Logan in [15]. To obtain the MFCCs, first the log-magnitude of the spectrum is computed. Then, the FFT bins are grouped and smoothed according the perceptually motivated Mel-frequency scaling, and de-correlated by means of a discrete cosine transform. To represent speech, 13 coefficients are commonly used, although it has been demonstrated that for classification tasks, it is enough to take into account only the first five coefficients [16].

2.8. Voice2White

This parameter, proposed in [17], is a measure of the energy inside the typical speech band (300-4000 Hz) respect to the whole energy of the signal. Mathematically,

$$v2w_t = 10log\frac{\sum_{300Hz}^{4kHz}|X_t[k]^2|}{\sum_{\forall k}|X_t[k]^2} \tag{7}$$

2.9. Activity level

The activity level of the audio signal is calculated according to the method for the objective measurement of active speech published by the ITU-T in its recommendation P.56 [18].

3. CLASSIFICATION ALGORITHMS

A number of different classification algorithms have been proposed in the literature. This paper will focus its attention on two of them: the K-nearest neighbor algorithm and the Fisher linear discriminant [1]. Both will be now briefly described.

3.1. K-nearest neighbor

The K-nearest neighbor (K-NN) is a very simple, yet powerful classification algorithm. Let us assume that we have a training set with L vectors grouped into C different classes. To obtain the class corresponding to a new observed vector $\mathbf{X}$, the algorithm has simply to look for the K nearest neighbors to the test vector $\mathbf{X}$, and weigh their class numbers they belong to, usually using a majority rule. Although it is possible to use different distance measures, most implementations employ a euclidean measure.

To express this idea in a more formal way, let us consider a set of training vectors $\{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_L}\}$ with $\mathbf{x_i} \in \mathbb{R}^n$ organized into C different classes y_i. Let $\mathbb{R}^n(\mathbf{x}) = \mathbf{x}' : ||\mathbf{x} - \mathbf{x}'|| \leq r^2$ be a ball centered in the vector $\mathbf{x}$ in which lie K prototype vectors $\mathbf{x_i}$. The K-nearest neighbor classification rule is defined as:

$$q(\mathbf{x}) = arg_{max}v(\mathbf{x}, y) \tag{8}$$

Where $v(\mathbf{x}, y)$ is the number of prototype vectors $\mathbf{x_i}$ with hidden state $y_i = y$, which lie in the ball $\mathbf{x_i} \in \mathbb{R}^n(\mathbf{x})$.

3.2. Fisher linear discriminant

The basic idea behind Fisher linear discriminants is that the data are projected onto a line, and the classification is performed in this one-dimensional space.

The projection maximizes the distance between the means of the two classes while minimizing the variance within each class.

Let $\{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_L}\}$ where $\mathbf{x_i} \in \mathbb{R}^n$ be a set of binary labeled training vectors, with n_1 samples in class 1 denoted C_1 and n_2 samples in class 2 denoted C_2. As it can be observed, contrary to the K-NN algorithm, which can classify among an arbitrary number of different classes, the Fisher linear discriminant only performs the classification between two classes. This limitation however can be overcome by using, for example, one against all techniques.

The class separability function in a direction $\mathbf{w} \in \mathbb{R}^n$ is defined as:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}} \tag{9}$$

Where S_B and S_W are the between-class and within-class scatter matrixes respectively:

$$S_B = (\mathbf{m_1} - \mathbf{m_2})(\mathbf{m_1} - \mathbf{m_2})^T \tag{10}$$

$$S_W = \sum_{i=1,2} \sum_{\mathbf{x} \in \mathbf{C_i}} (\mathbf{x} - \mathbf{m_i})(\mathbf{x} - \mathbf{m_i})^T \tag{11}$$

The sample mean of the respective classes, $\mathbf{m_i}$ is defined as:

$$\mathbf{m_i} = \frac{1}{n_i} \sum_{\mathbf{x} \in \mathbf{C_i}} \mathbf{x} \tag{12}$$

The Fisher linear discriminant is given by the vector $\mathbf{w}$ that maximizes the class separability function $J(\mathbf{w})$. It can be observed that expression (9) is a particular case of the generalized Rayleigh quotient, and thus, and assuming that S_W is a non-singular matrix, it is possible to find an analytic expression for $\mathbf{w}$ which maximizes $J(\mathbf{w})$:

$$\mathbf{w} = S_W^{-1}(\mathbf{m_1} - \mathbf{m_2}) \tag{13}$$

This expression allows for calculating the optimal projection direction $\mathbf{w}$ that ensures that the samples belonging to each one of the two classes will be as much separated as possible. It is possible to demonstrate, [19], that assuming normal distributions and equal covariance for the two different classes, the resulting linear discriminant function is in the same direction as the Bayes optimal classifier.

4. RESULTS

4.1. Database used

All the experiments here described have been done using the sound database for speech/music classification provided by Dan Ellis, which was used in several publications [8][20]. This database was recorded directly from the radio, using a sampling frequency of 22050 Hz, 16 bits per sample and only one channel (mono). It contains a set of files for the training process with a total length of 45 minutes (180 files with a length of 15 seconds each), belonging to the classes: speech alone (60 files), speech in presence of music or background noise (60 files) and music (60 files). For the test a total of 15.25 minutes of audio material are available, divided into 366 files with a length of 2.5 seconds each. There are 120 speech files (with and without background music), 126 of music with no vocals, and 120 of music with vocals.

4.2. Results obtained and discussion

Table 1 shows the probabilities of error obtained for the speech/music classification task using all the features individually. The results shown have been obtained using the Fisher's linear discriminant analysis and the K-Nearest Neighbors algorithms, with K=1 and K=3. As it can be observed, the best results are obtained with the Fisher's linear discriminant, except for the STE feature, which works best with the 1-NN classifier. The best results are obtained using the MFCC and the Voice2White features, with a probability of error of 4.09% and 4.91% respectively. As it can be observed, for the spectral centroid the results are better if the Modified Discrete Transform (MDCT) is used, while for the roll-off the best results are achieved using the Discrete Fourier Transform. With the classification algorithms considered, the combination of two or more of these features does not seem to improve the results. Combining the MFCC and the Voice2White features with a Fisher linear discriminant classifier, leads to a probability of error equal to 4.09%, the same than for the MFCC alone.

Feature	Probability of error		
	Fisher	1-NN	3-NN
Centroid (MDCT)	8.74%	17.48%	21.85%
Centroid (DFT)	16.66%	29.23%	30.60%
Roll-off (MDCT)	14.48%	25.40%	21.85%
Roll-off (DFT)	8.19%	13.11%	13.11%
ZCR	9.83%	19.67%	18.03%
HZCRR	25.13%	39.89%	36.33%
STE	48.63%	22.40%	22.67%
LSTER	11.74%	33.87%	23.77%
MFCC	4.09%	22.13%	26.50%
Voice2White	4.91%	6.28%	6.01%
Activity level	12.84%	18.03%	18.85%

Table 1: Probabilities of error obtained using each one of the features individually.

Classifier		Speech	Music
Fisher			
	Speech	104	16
	Music	2	244
1-NN			
	Speech	114	6
	Music	17	229
3-NN			
	Speech	116	4
	Music	18	228

Table 2: Confusion matrixes using the Voice2White feature with three different classification algorithms.

On the other hand, combining the different classification algorithms seems to be more positive. As an example consider Table 2, where the confusion matrixes obtained for the Voice2White feature using the three considered classifiers are shown. As it can be observed, the results are quite complimentary: the Fisher linear discriminant has a higher probability of error when the input is speech, while the nearest neighbor behaves worse when the input is music. If the results obtained from the three classifiers are combined using a majority rule, then the probability of error drops down to a 4.5%.

These results are very promising in the sense that the application of Fisher linear discriminant analysis to the task of speech/music classification seems to provide very good results.

5. CONCLUSION

This paper has shown some preliminary results on the application of Fisher linear discriminant analysis to the problem of speech/music discrimination. To evaluate the performance of the system, the results are compared with a nearest-neighbor classifier, which has been widely used in the literature. The results obtained show us that the Fisher linear discriminant analysis can provide very promising results using only one feature for the classification. Better results may be obtained combining the results obtained from two or more classifiers. Further work will imply using more complex classification algorithms, such as neural networks, to improve the performance of the system.

ACKNOWLEDGEMENT

The authors would like to thank Dan Ellis for providing them his speech/music database used in the experiments.

6. REFERENCES

[1] S. Mika, G. Rätsch, J. Weston, B. Schölkpf, and K. Müller, *Neural networks for signal processing.* IEEE, 1999, ch. Fisher discriminant analysis with kernel, pp. 41–48.

[2] C.-M. Liu and W.-C. Lee, "A unified fast algorithm for cosine modulated filter bamks in current audio coding standards," in *AES 104th Convention*, ser. Preprint Number 4729, April 1998.

[3] L. Qingshan, H. Rui, L. Hanqing, and M. Songde, "Face recognition using kernel-based fisher discriminant analysis," in *Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.

[4] P. Herrera, X. Amatriain, E. Batlle, and X. Serra, "Towards instrument segmentation for music content description: a critical review of instrument classification techniques," in *International Symposium on Music Information Retrieval*, 2000.

[5] P. Cano, M. Koppenberger, S. L. Groux, J. Ricard, N. Wack, and P. Herrera, "Nearest-neighbour generic sound classification with a

worldnet-based taxonomy," in *AES 116th Convention*, 2004.

[6] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in *ICASSP*, 2002.

[7] J. Saunders, "Real time discrimination of broadcast speech/music," in *ICASSSP*, 1996, pp. 993–996.

[8] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *ICASSP*, 1997.

[9] R. M. Aarts and R. T. Dekkers, "A real-time speech-music discriminator," *J. Audio Engineering Society*, vol. 47, no. 9, pp. 720–725, September 1999.

[10] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal, "Speech/music discrimination for multimedia applications," in *ICASSP*, 2000.

[11] S. Davis and P. Mermelstein, "Experiments in syllable-based recognition of continuous speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 357–366, August 1980.

[12] G. Tzanetakis and P. Cook, "Marsyas: A framework for audio analysis," *Organised Sound*, vol. 4, no. 3, 2000.

[13] J.-J. Aucouturier and F. Pachet, "Representing musical genre: a state of the art," *Journal of New Music Research*, vol. 32, no. 1, pp. 83–93, 2003.

[14] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transactions on speech and audio processing*, vol. 10, no. 7, pp. 504–516, October 2002.

[15] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Int. Symp. Music Information Retrieval (ISMIR)*, 2000.

[16] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, July 2002.

[17] E. Guaus and E. Batlle, "A non-linear rhythm-based style classification for broadcast speech-music discrimination," in *AES 116th Convention*, 2004.

[18] "Objective measurement of active speech level," ITU-T, Recommendation P.56, 1993.

[19] S. A. Billings and K. L. Lee, "Nonlinear fisher discriminant analysis using a minimum squared error cost function and the orthogonal least squares algorithm," *Neural networks*, vol. 15, pp. 263–270, 2002.

[20] G. Williams and D. P. Ellis, "Speech/music discrimination based on posterior probability features," in *Eurospeech*, 1999.