Engineering Applications of Artificial Intelligence 23 (2010) 151–159

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

# Two-stage cascaded classification approach based on genetic fuzzy learning for speech/music discrimination

N. Ruiz Reyes *, P. Vera Candeas, S. García Galán, J.E. Muñoz

*Telecommunication Engineering Department, University of Jaén Polytechnic School, C/ Alfonso X el Sabio 28, 23700 Linares, Jaén, Spain*

ARTICLE INFO

*Article history:*
Received 2 October 2008
Received in revised form
20 February 2009
Accepted 4 June 2009
Available online 30 September 2009

*Keywords:*
Speech/music discrimination
Signal features
Statistical pattern recognition
Classifier
Fuzzy rules-based system
Genetic learning
Evolutionary computation
Multimedia

ABSTRACT

Automatic discrimination of speech and music is an important tool in many multimedia applications. The paper presents a robust and effective approach for speech/music discrimination, which relies on a two-stage cascaded classification scheme. The cascaded classification scheme is composed of a statistical pattern recognition classifier followed by a genetic fuzzy system. For the first stage of the classification scheme, other widely used classifiers, such as neural networks and support vector machines, have also been considered in order to assess the robustness of the proposed classification scheme. Comparison with well-proven signal features is also performed. In this work, the most commonly used genetic learning algorithms (Michigan and Pittsburgh) have been evaluated in the proposed two-stage classification scheme. The genetic fuzzy system gives rise to an improvement of about 4% in the classification accuracy rate. Experimental results show the good performance of the proposed approach with a classification accuracy rate of about 97% for the best trial.



## 1. Introduction

Automatic speech/music discrimination (SMD) 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; Richard et al., 2007; Scheirer and Slaney, 1997; El-Maleh et al., 2000; Harb and Chen, 2003; Keum and Lee, 2006; Wang et al., 2008). Each of them 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. Richard et al. (2007) have recently developed a combined supervised and unsupervised approach, which includes feature selection, for automatic segmentation of radiophonic audio streams.

In automatic speech recognition (ASR) of broadcast news, it is desirable to disable the input to the speech recognizer during the non-speech portion of the audio stream. Scheirer and Slaney (1997) developed a SMD system for ASR of audio sound tracks. Thirteen features and three classification schemes (MAP Gaussian, GMM and *k*- NN) were exploited, resulting in an accuracy of over 90%. Matsunaga et al. (2004) proposed in an audio source segmentation method for automatic indexing of broadcast news, which relies on spectral correlation features.

Low bit-rate audio coding is another application that can benefit from distinguishing between speech and music. Designing an universal coder to reproduce well both speech and music is the best approach. However, it is not a trivial issue. An alternative approach consists of designing a multi-mode coder that can accommodate different signals. The appropriate module is selected by using the output of a speech–music classifier (Tancerel et al., 2000; Exposito et al., 2007).

Automatic SMD is an important tool in many multimedia applications. El-Maleh et al. (2000) propose in a low delay (20 ms) speech/music classification approach, 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; Lu et al., 2002). Minami et al. (1998) proposed an audio-based approach to video indexing, where SMD is used to browse a video database. In Gong and Xiong-wei (2006), content-based indexing and retrieval of cognitive multimedia is performed using SMD. The gray correlation analysis method is applied, which makes the algorithm feasible for real-time multimedia applications.

Comparative analysis of different types of features for SMD is provided in Carey et al. (1999). Experimental results showed that 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. In Harb and

* Corresponding author. Tel.: +34 953648554; fax: +34 953648508.
*E-mail address:* nicolas@ujaen.es (N. Ruiz Reyes).

0952-1976/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2009.06.006

Chen (2003), MFSC's first order statistics are combined with neural networks, providing quite good results. MFCC have been widely used in ASR, and proposed in last few years for musical genre classification (Tzanetakis and Cook, 2002; Burred and Lerch, 2004; Ezzaidi and Rouat, 2007). Comparison of statistical and information theory measures for automatic musical genre classification is reported in Ezzaidi and Rouat (2007).

In the problem of audio classification, the requirements of low-complexity, high-accuracy and short delay are crucial for some practical scenarios. In Wang et al. (2008), the authors propose a real-time SMD method based on a hierarchical decision tree, which comes with promising short delay of 10 ms, high accuracy of 98% and low complexity. Only two signal characteristics (amplitude and mean frequency) are used in Panagiotakis and Tziritas (2005), giving also rise to a short delay (20 ms), high accuracy (95%) and low complexity SMD approach.

The main contribution of the paper is the two-stage cascaded classification scheme based on genetic fuzzy learning for SMD. It achieves a meaningful improvement in the classification accuracy rate by incorporating a fuzzy rules-based system (FRBS), which evolves using genetic learning algorithms. The proposed classification scheme consists of a classical statistical pattern recognition (SPR) classifier followed by the FRBS (two-stage cascaded classification scheme). It results in a robust and effective approach for SMD.

This paper is structured as follows. Comprehensive review of the main existing approaches for SMD and the areas of application are discussed in Section 1. Section 2 is devoted to the proposed SMD approach. It consists of two parts: (1) brief description of classical features for SMD and (2) the proposed two-stage cascaded classification scheme. Experimental results are shown in Section 3, which allow to assess the performance of the proposed classification approach. Finally, Section 4 outlines some meaningful conclusions and future research lines.

## 2. The proposed speech/music discrimination approach

Section 2 is organized in two parts. First, Section 2.1 reviews the most commonly used features in audio classification tasks, which represent timbral texture, and are based on the short time Fourier transform (STFT). Then, the two-stage cascaded decision-taking scheme is described in Section 2.2, which also contains the motivation of using the proposed classification scheme.

### 2.1. Classical features for speech/music discrimination

Most of the works concerning audio classification rely on three types of features: timbral texture features, rhythmic content features and pitch content features (Tzanetakis and Cook, 2002; Burred and Lerch, 2004). However, timbral texture features are the most commonly used in audio classification when it is reduced to SMD. The features used to represent timbral texture are based on low-level signal features proposed for music–speech discrimination (Scheirer and Slaney, 1997) and speech recognition (Davis and Mermelstein, 1980). These features are based on the STFT and are calculated for every short-time frame of sound. The following low-level signal features, representing timbral texture, are used in this work for assessing the robustness of the proposed classification scheme:

(1) *Spectral centroid* (*SC*). The spectral centroid is defined as the center of gravity of the magnitude spectrum of the STFT:

$$C_t = \frac{\sum_{n=1}^{N} M_t[n] * n}{\sum_{n=1}^{N} M_t[n]} \quad (1)$$

where $M_t[n]$ is the magnitude of the Fourier transform at frame $t$ and frequency bin $n$. The centroid is a measure of spectral shape and higher centroid values correspond to "brighter" textures with more high frequencies.

(2) *Spectral rolloff* (*SR*). The spectral rolloff is defined as the frequency $R_t$ below which 85% of the magnitude distribution is concentrated:

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 * \sum_{n=1}^{N} M_t[n] \quad (2)$$

The rolloff is another measure of spectral shape.

(3) *Spectral flux* (*SF*). The spectral flux is defined as the squared difference between the normalized magnitudes of successive spectral distributions:

$$F_t = \sum_{n=1}^{N} (N_t[n] - N_{t-1}[n])^2 \quad (3)$$

where $N_t[n]$ and $N_{t-1}[n]$ are the normalized magnitude of the Fourier transform at the current frame $t$, and the previous frame $t-1$, respectively. The spectral flux is a measure of the amount of local spectral change.

(4) *Time domain zero crossings* (*ZC*). This timbral texture feature is defined as:

$$Z_t = \frac{1}{2} \sum_{n=1}^{N} |sign(x[n]) - sign(x[n-1])| \quad (4)$$

where the *sign* function is 1 for positive arguments and 0 for negative arguments and $x[n]$ is the time domain signal for frame $t$. Time domain zero crossings provide a measure of the noisiness of the signal.

(5) *Mel-frequency cepstral coefficients* (*MFCC*). Mel-frequency cepstral coefficients are perceptually motivated features that are also based on the STFT. After taking the log-amplitude of the magnitude spectrum, the FFT bins are grouped and smoothed according to the perceptually motivated Mel-frequency scaling. Finally, in order to decorrelate the resulting feature vector a discrete cosine transform (DCT) is performed. Although typically 13 coefficients are used for speech representation, it has been found that the first five coefficients provide the best performance for musical genre classification of audio signals (Tzanetakis and Cook, 2002). Therefore, we have also used the first five coefficients in our speech/music discrimination approach, without adding the derivatives. The use of MFCC to separate music and speech has been explored in Logan (2000).

In this work, an *analysis window* of 23.22 ms (1024 samples at 44,100 Hz sampling rate) is defined. It implies that parameter $N$ is equal to 1024. A *texture window* of approximately 1 s (43 analysis windows) is also defined. Overlapping with a hop size of 512 samples is performed, which results in a feature vector or matrix **F** for each texture window. When the low-level signal feature to be extracted is SC, SR, SF or ZC, a 86-length feature vector is obtained for each texture window. However, when the low-level signal features to be extracted are the first five MFCC, a feature matrix of size $5 \times 86$ is obtained. The texture window is shifted by 250 ms, which entails updating feature vector or matrix **F** every 250 ms. From feature vector or matrix **F**, statistical values (mean, standard deviation and skewness) are computed for each 1 second-length texture window.

Fig. 1 shows the windows scheme defined in this work to obtain the input values to the classification stage.

The values here considered for the length of the analysis and texture windows (23.22 ms and 1 s, respectively) are widely used in other related works dealing with SMD (El-Maleh et al., 2000;