Chapter 6

FEATURES FOR AUDIO CLASSIFICATION

Jeroen Breebaart and Martin F. McKinney

Abstract Four audio feature sets are evaluated in their ability to differentiate five audio classes: popular music, classical music, speech, background noise and crowd noise. The feature sets include low-level signal properties, mel-frequency spectral coefficients, and two new sets based on perceptual models of hearing. The temporal behavior of the features is analyzed and parameterized and these parameters are included as additional features. Using a standard Gaussian framework for classification, results show that the temporal behavior of features is important for automatic audio classification. In addition, classification is better, on average, if based on features from models of auditory perception rather than on standard features.

Keywords Audio classification, automatic content analysis

6.1 Introduction

Developments in Internet and broadcast technology enable users to enjoy large amounts of multimedia content. With this rapidly increasing amount of data, users require automatic methods to filter, process and store incoming data. Examples of applications in this field are automatic setting of audio equalization (e.g., bass and treble) in a playback system, automatic setting of lighting to correspond with the mood of the music (or vice versa), automatic cutting, segmenting, labeling, and storage of audio, and automatic playlist generation based on music similarity or some other user specified criteria. Some of these functions will be aided by attached *metadata*, which provides information about the content. However, due to the fact that metadata is not always provided, and because local processing power has increased tremendously, interest in *local* automatic multimedia analysis has increased. A major challenge in this field is the automatic classification of audio. During the last decade, several authors have proposed algorithms to classify incoming audio data based on different algorithms [Davis & Mermelstein, 1980; Wold et al., 1996; Spina &

W. Verhaegh et al. (eds.), Algorithms in Ambient Intelligence, 113-129

Zue, 1996; Scheirer & Slaney, 1997; Spina & Zue, 1997; Scheirer, 1998; Zhang et al., 1998; Wang et al., 2000a; Wang et al., 2000b; Zhang & Kuo, 2001; Li et al., 2001]. Most of these proposed systems combine two processing stages. The first stage analyzes the incoming waveform and extracts certain parameters (features) from it. The feature extraction process usually involves a large information reduction. The second stage performs a classification based on the extracted features.

A variety of signal features have been proposed for general audio classification. A large portion of these features consists of low-level signal features, which include parameters such as the zero-crossing rate, the signal bandwidth, the spectral centroid, and signal energy [Davis & Mermelstein, 1980; Wold et al., 1996; Scheirer & Slaney, 1997; Scheirer, 1998; Wang et al., 2000a; Wang et al., 2000b]. Usually, both the averages and the variances of these signal properties are included in the feature set. A second important feature set which is inherited from automatic speech recognizers consists of mel-frequency cepstral coefficients (MFCC). This parametric description of the spectral envelope has the advantage of being level-independent and of yielding low mutual correlations between different features for both speech [Hermansky & Malayath, 1998] and music [Logan, 2000]. Classification based on a set of features that are uncorrelated is typically easier than that based on features with correlations.

Both low-level signal properties and MFCC have been used for general audio classification schemes of varying complexity. The simplest audio classification tasks involve the discrimination between music and speech. Typical classification results of up to 95% correct have been reported [Toonen Dekkers & Aarts, 1995; Scheirer & Slaney, 1997; Lu & Hankinson, 1998]. The performance of classification schemes usually decreases if more audio classes are present [Zhang et al., 1998; Zhang & Kuo, 2001]. Hence, the use of features with high discriminative power becomes an issue. In this respect, the MFCC feature set seems to be a powerful signal parameterization that outperforms low-level signal properties. Typical audio classes that have been used include clean speech, speech with music, noisy speech, telephone speech, music, silence and noise. The performance is roughly between 80 and 94% correct [Foote, 1997; Naphade & Huang, 2000a; Nahphade & Huang, 2000b; Li et al., 2001].

For the second stage, a number of classification schemes of varying complexity have been proposed. These schemes include Multivariate Gaussian models, Gaussian mixture models, self-organizing maps, neural networks, k-nearest neighbor schemes and hidden Markov models. Some authors have found that the classification scheme does not influence the classification accuracy [Scheirer & Slaney, 1997; Golub, 2000], suggesting that the topology of the feature space is relatively simple. An important implication of these results