Semi-Supervised Feature Selection for Audio Classification

Xu-Kui Yang
Zhengzhou Information Science and Technology Institute
gzyangxk@163.com

Wei-Qiang Zhang
Department of Electronic Engineering, Tsinghua University
wqzhang@tsinghua.edu.cn

ABSTRACT

Audio classification, the segmentation of an audio signal into broad categories such as speech, non-speech, and silence, is an important front-end problem in speech signal processing. Dozens of features have been proposed for audio classification. Unfortunately, these features are not directly complementary and combining them does not improve classification performance. Feature selection provides an effective mechanism for choosing the most relevant and least redundant features for classification.

1. INTRODUCTION

Initial segmentation of audio signals into broad categories such as speech, non-speech, and silence, provides useful information for audio content understanding and analysis [1], and it has been used in a variety of commercial, forensic and military applications [2]. Most audio classification systems involve two processing stages: feature extraction and classification. There is a considerable amount of literature on audio classification regarding different features [3] or classification methods [4]. Many features [5] have been developed to improve classification accuracy. Nevertheless, using all of these features in a classification system may not enhance but instead degrade the performance. The underlying reason is that there can be irrelevant, redundant, and even contradictory information among these features. Choosing the most relevant features to improve the classification accuracy is a challenging problem [6].

2. SEMI-SUPERVISED FEATURE SELECTION FOR AUDIO CLASSIFICATION

Assuming that an audio signal has been divided into a sequence of audio segments using a segmentation algorithm or timestamp, audio classification focus on the classi-fication of categorizing these audio segments into a set of predefined audio classes.

Fig. 1 illustrates the process of audio classification. In an audio classification sys-tem, every audio signal is first divided into mid-length segments which range in duration from 0.5 to 10 seconds. After this, the selected features are extracted for each segment using short-term overlapping frames. The sequence of short-term features in each segment is used to compute feature statistics, which are used as inputs to the classifier. In the final classification stage, the classifier determines a segment-by-segment decision.

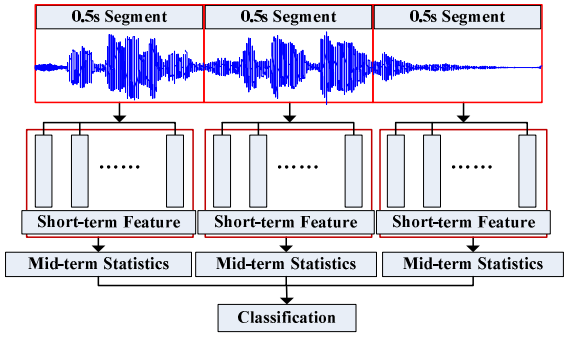

Figure 1. The audio classification framework.

In audio analysis and classification there are dozens of features which can be used. Widely-used time-domain features [5] include short-term energy [7], zero-crossing rate [8], and entropy of energy [9]. Common frequency-domain features include spectral centroid, spectral spread, spectral entropy [10], spectral flux, spectral roll-off, MFCCs, and chroma vector [11].

There is a lot of complementary information among these features which can im-prove classification accuracy when used together; however, there is also a lot of re-dundant and even contradictory information which can decrease performance. It is hard to judge which combination of features is most likely to have a positive effect on classification. Furthermore, it is computationally infeasible to select the optimal feature subset by exhaustive search. Thus, it's important to implement an effective feature selection method for this task.

Most supervised feature selection methods are dependent on labeled data. Unfortunately, it is difficult to obtain sufficient labeled data for audio classification, while unlabeled data is readily available. Semi-supervised feature selection methods can take good use of both labeled and unlabeled data, thus this approach is more practical.

3. REFERENCES

[1] J. Foote, "Content-based retrieval of music and audio," in *Proc. SPIE*, vol. 3229, pp. 138–147, 1997.

[2] Lie Lu, Hong-Jiang Zhang, and Hao Jiang, "Content Analysis for Audio Classification and

Segmentation," in *IEEE Trans. Speech Audio Processing*, vol. 10, no. 7, 2007, pp. 504-515.

[3] M. F. McKinney and J. Breebaart, "Features for Audio and Music Classification," in *Proc. ISMIR 2003, 4th International Conference on Music Information Retrieval*, Baltimore, Maryland, USA, October 27-30, 2003.

[4] L. Honglak, L. Yan, P. Peter, and Y. N. Andrew, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Proc. Neural Information Processing Systems (NIPS) 22*, 2009, pp. 1096-1104.

[5] T. Giannakopoulos, and A.Pikrakis, *Introduction to Audio Analysis: A MATLAB Approach*, Elsevier Academic Press, 2014.

[6] Z. Zhao and H. Liu, *Spectral Feature Selection for Data Mining* (Data Mining and Knowledge Discovery Series). Boca Raton, FL, USA: Chapman and Hall-CRC, 2012.

[7] C. Panagiotakis and G. Tziritas, "A speech/music discriminator based on RMS and zero-crossings," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 155-166, 2005.

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

[9] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, "Gunshot detection in audio streams from movies by means of dynamic programming and Bayesian networks," in *Proc. ICASSP*, 2008.

[10] H. Misra, S. Ikbal, H. Bourlard, and H. HERMANSKY, "Spectral entropy based feature for robust ASR, in *Proc. ICASSP*, 2004.

[11] M. A. Bartsch and G. H. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," IEEE Trans. Multimedia, vol. 7, no. 1, pp. 96-104, 2005.