REAL-TIME SPEECH/MUSIC CLASSIFICATION WITH A HIERARCHICAL OBLIQUE DECISION TREE

Jun Wang, Qiong Wu, Haojiang Deng, Qin Yan

Institute of Acoustics, Chinese Academy of Sciences, Beijing, China

{wangjun, wuq, denghj, yanq}@dsp.ac.cn

ABSTRACT

In the problem of classification of audio signals, the requirements of low-complexity, high-accuracy and short delay are crucial for some practical scenarios. This paper proposes a method of real-time speech/music classification with a hierarchical oblique decision tree. A set of discrimination features in frequency domain are selected together with a proposed simple harmonic structure stability feature, which is based on a rough estimation of the harmonic structure. A feature subset selection tool is used to select a subset of short and long term features to feed into a hierarchical oblique decision tree classifier. The method is evaluated and compared with the open loop selection mode in AMR-WB+. Experiments show the proposed approach gives a better performance (98.3%) compared to other prevailing approaches. In particular, it comes with promising short delay of 10 ms and low complexity of 1 wmops.

Index Terms—signal classification, FSS, harmonic structure, hierarchical oblique decision tree

1. INTRODUCTION

With rapid changes in the telecommunication network environment, the classification of audio signals is one key component in many multimedia systems. For instance, most codecs are designed to handle signals without discrimination and can not work properly in the existence of multimedia signals. This paper proposes a real-time speech/music classification algorithm which meets the conditions of low complexity, high accuracy and short delay.

Many efforts have been made in this area so far. A variety of classifier approaches have been researched and applied, such as simple manual thresholds approach [1, 2], classical modeling approaches such as Gaussian mixture model (GMM) and vector quantizer (VQ) model, multivariate-Gaussian classifier [3]. However, a hybrid of different methods will consistently outperform a single method. More research is focused on a hybrid analysis of different methods to obtain an ensemble scheme, such as [4, 5, 6]; particularly in [7], Scheirer constructs a refined speech/music discriminator by combining a Gaussian maximum a posteriori (MAP) estimator, a GMM classifier, a spatial partitioning scheme and a nearest-neighbor classifier to gain an improved accuracy.

Table 1: accuracy (%) of reference classifiers

Segment-level	[1]	[5]	[8]	[3]	[4]
	91.9	95.4	97	98	99.43
Frame-level	[6]	[9]	[7]	[2]	
	80.9	87	94.2	95	

Feature selection is an important step for classification. In general, there are two feature categories: time domain features and frequency domain features. Saunders [3] employs strict time domain features for low complexity; similarly Wang [8] selects just a single time domain feature. Both of them report an accuracy of more than 97%. In contrast, more research is based on features extracted in frequency domain, e.g. [6] is LSF-based. Furthermore, to overcome the deficiencies of short-term features, modulation-scale analysis for long-term features [1, 7], and run-time features such as averages and variances [2, 7, 9] are generally employed. Other typical works such as [4] obtain an optimal feature set out of a large feature space through feature subset selection (FSS) [10].

The audio classification problem can also be regarded from another perspective: segment-level based and frame-level based. Most classifiers are segment-level based. The accuracy rates reported by segment-level and frame-level based references are shown in Table 1. As they are not tested under identical conditions, these results are not listed for comparison, but just for a glimpse of the achievements of existing approaches.

This paper is organized as follows. Section 2 briefly describes a set of features. In section 3 the feature set is selected by FSS; the output optimal feature sets are then fed into a hierarchical oblique decision tree (DT) classifier to construct a frame-level classification framework. Finally, in section 4, experiment results are provided, and the proposed method is evaluated and compared with other

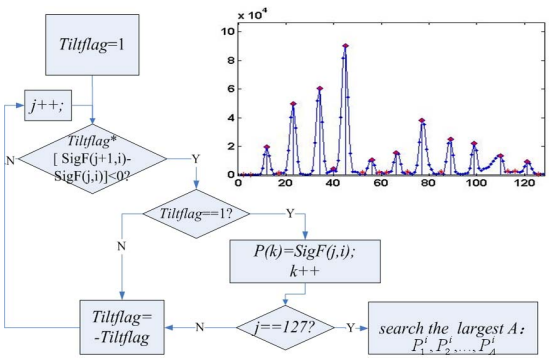

Figure 1: flowchart of harmonic structure estimation and output illustration.

classification algorithms.

2. FEATURES

The input signal is divided into 10ms length frames with non-overlapped windowing and transformed to *SigF* with FFT size of 256. Features are listed below.

- *Normalized spectral Flux between frames(Flux)*

$$Flux = \frac{\sum_{j=FLUX\_F1}^{FLUX\_F2} |SigF(i,j) - SigF(i-1,j)|}{norm} \quad (1)$$

- *Normalized spectral Flux between sub-bands (SFlux)*

$$SFlux = \frac{\sum_{j=SFLUX\_F1}^{SFLUX\_F2} |SigF(i,j) - SigF(i,j-1)|}{norm} \quad (2)$$

Where *norm* (·) is the normalized function, and *FLUX_F1*, *FLUX_F2*, *SFLUX_F1 and SFLUX_F2* are frequency boundaries. The variances of *Flux* and *SFlux* within *run_on_len* frames (typically *run_on_len* may be 20), named *varFlux*, *varSFlux*, as well as the moving average of those variances, named *varmovFlux* and *varmovSFlux*, are also exploited. All of these features are typically higher for speech than for most music.

- *stda_short, stda_long*

These are the standard deviations of energy levels [11]. They describe the variations of the frequency-band energy within short and long windows. Music tends to have less variation than speech does.

- *Energy ratio (hpl)*

This feature describes the relationship between higher frequency bands and lower frequency bands. The energy of higher frequency bands (such as 2500-4000Hz) *LevH*, is devided by the energy of lower frequency bands (such as 1000-2000Hz) *LevL* to create *hpl*. The moving average *mov_hpl* and the variance *var_hpl* within *run_on_len* frames are also exploited.

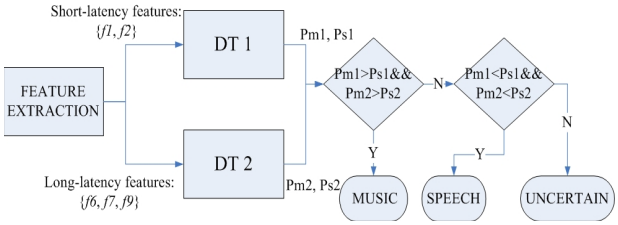

Figure 2: hierarchical DT

- *Harmonic structure stability* (*hss*)

A discriminating attribute of music and speech is the stability of their harmonic structure. Zhang [12] proposes the Average Harmonic Structure (*AHS*) and Harmonic Structure Stability (*HSS*) to measure the stability of harmonic structures. Considering the complexity of harmonic computing, we propose a rough yet efficient algorithm to estimate the harmonic structure. First, we consider the FFT spectral amplitude as a discrete multimodal function, and adopt the multimodal function optimization approach [13] to find the first *A* amplitude peaks of each spectral frame. Fig. 1 shows how we determine the monotone increasing and monotone decreasing interval of the function by a single flag *Tiltflag*. The transform point between monotone increasing and monotone decreasing interval is determined to be a peak. The 128 frequency points need to be searched only once, and all the local and global optimal peaks can be found, as shown in Fig. 1.

Second, we search the highest *A* amplitude peaks: $P_1^i, P_2^i, ..., P_A^i$, which describe the amplitudes of harmonic. Here, *i* describes the frame count. Then the peaks are translated into normalized log-scale peaks: $LP_1^i, LP_2^i, ..., LP_A^i$, as defined in (3)

$$LP_j^i = \log(P_j^i) - \log(\sum_j P_j^i) \qquad j = 1,...,A \quad (3)$$

Thirdly, the variances of *LP* within *run_on_len* frames, named *hss*, are calculated [12].

3. FEATURE SELECTION AND CLASSIFIERS

Usually, short-term features, such as {*f1, f2, f3, f4, f5*}: {*varFlux*, *varSFlux*, *hpl*, *stda_short*, *zcr*} have much more discrimination impurity but less delay; whereas long-term features such as {*f6, f7, f8, f9, f10*}: {*varmovFlux*, *varmovSFlux*, *stda_long*, *mov_hpl*, *var_hpl*} have more delay for frame-level classification, but less impurity. If both short-term and long-term features are put into a single set for training, the short-term features tend to be NOT selected by FSS due to their impurity and their high

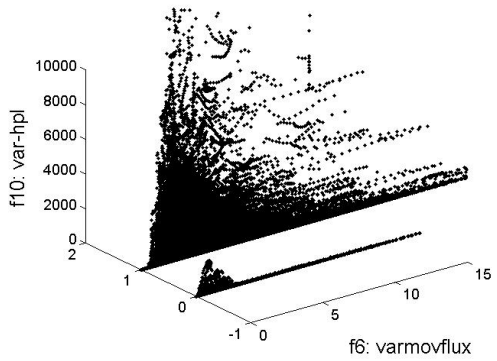

Figure 3: scatter plot of {*f6, f10*}.
Front: music; back: speech.

correlation with the long-term ones. That will be further confirmed by the FSS result at the end of this section.

Here, the hierarchical DT structure is proposed. It is organized with double layers of DTs trained with short and long term-based feature sets. As shown in Fig. 2, *Pm1*, *Ps1* are respectively the accuracy of music and speech classification in the training of *DT1*; similarly *Pm2*, *Ps2* are the accuracy of music and speech classification in the training of *DT2*.

The hierarchical DT provides output in three classes: *MUSIC*/*SPEECH*/*UNCERTAIN*. Afterward, the classification is refined with a counter *music_Cnt*, which indicates how many continuous previous frames up to that point are classified as music. If the output is *UNCERTAIN* and *music_Cnt* is larger than *thr_mu_cnt*, then the current frame is set as *MUSIC*. Otherwise the current frame is set as *SPEECH*.

Fig. 3 shows the scatter plot of {*f6, f10*}. It should be intuitively clear that the underlying concept of speech/music classification is defined by a polygonal space partitioning. Accordingly, a DT with a linear combination of attributes should perform a better discrimination than an axis-parallel DT. Here, we choose the oblique DT OC1 [14], which offers a good oblique split in the form of a hyperplane at each node of the DT. The hyperplane takes the form: $\sum_{i=1}^{d} a_i x_i + a_{d+1} > 0$, where d is the number of attributes, $a_1,..., a_{d+1}$ are real-valued coefficients and x_i are real-valued attributes.

For comparison, we select the Brieman's classical classification and regression trees [15]. This is an axis-parallel method which is a powerful yet simple method for speech/music classification. This tree branches out to several ending nodes, each of which outputs a decision with a music probability and a speech probability,

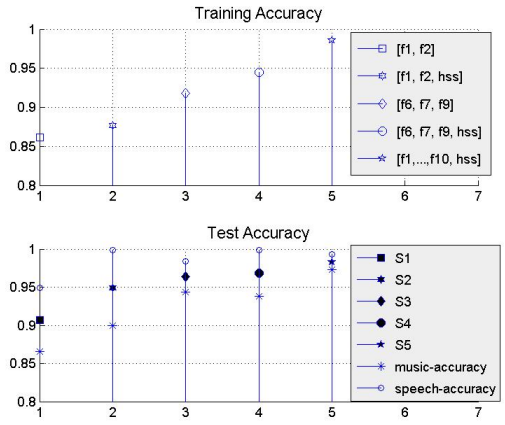

Figure 4: training accuracy and test accuracy

respectively.

For speech/music classification, there are several advantages in using the oblique methods compared to the axis-parallel methods:

· More compact trees. Above all, concerning this particular issue, the OC1 method produces a single hyperplane, while the axis-parallel method produces a tree with roughly 10 mid-nodes and 10 ending-nodes after pruning.

· Simpler DSP implementation. A hyperplane is superior to an axis-parallel tree, as the latter has quite a few logic branches that will hamper DSP implementation.

· More accuracy. Both the previous analysis of Fig. 3 and the experiment results in section 4 demonstrate that OC1 is a better method.

· More flexible system with more efficient code. A hyperplane makes adapting the system to different use scenarios much easier, e.g. adapting from digital media storage applications to portable voice/audio recorders which may be operating in extremely noisy environments. We simulated these scenarios by training features of noisy speech/music at different SNR levels, such as 6db and 15db; we found that the fitted axis-parallel trees at the different SNR levels are different from each other, both in shapes and in each node's logics, whereas a hyperplane simply needs to adapt its coefficients $a_1,..., a_{d+1}$.

The oblique DT induction method can benefit substantially by using a feature selection method which selects a subset from the original attribute set. As some classical FSS methods will work fine within OC1, we chose FSS-naive-bayes as the feature selection method. Through FSS-naive-bayes, subset {*f1*, *f2*} is selected from short-term feature set {*f1*, *f2*, *f3*, *f4*, *f5*}; subset {*f6*, *f7*, *f9*} is selected from long-term feature set {*f6*, *f7*, *f8*, *f9*, *f10*}. The results of FSS are well compatible with the correlation coefficients results; for instance, *f7* and *f8* have a correlation coefficient as high as 0.31, so that one of them is discarded.

4. EXPERIMENTS AND RESULTS

The training data consists of 228,512 music frames and 237,671 speech frames, with 10ms length per frame and 16 kHz sampling frequency. The speech data covers about 20 English and 20 Chinese multiple speakers; for each language, half of the speakers are male and half are female. Music was selected from various genres including jazz, rock, symphony, Chinese folk music, and so on. Separate music and speech test sequences were chosen for independent testing. There are also 6 speech test sequences, including respectively multiple male and female speakers of Chinese, English and French, each sequence is about 5 minutes long. There are 6 music test sequences of jazz, piano, saxophone, folk, symphony and concerto; the sequences vary in length from 5 minutes to 11 minutes.

The accuracy is calculated based on 10ms frame-level. The upper plot in Fig. 4 gives the training accuracy of different feature sets. The open-loop mode selection function in AMR-WB+ [11], which corresponds to a real-time 20ms frame-level speech/music classification, is chosen for the independent test. In the lower plot of Fig. 4, *S1* indicates the accuracy of AMR-WB+; *S2*: the hierarchical axis-parallel DT classifier with a short-term feature set {*f1*,…, *f5*} and a long-term feature set {*f6*,…, *f10*}, *S3*: the oblique DT classifier with a feature set {*f1*,…, *f10*}, *S4*: the hierarchical oblique DT classifier with the selected short-term feature set {*f1*, *f2*} and long-term feature set {*f6*, *f7*, *f9*}, and *S5*: the hierarchical oblique DT classifier with the selected feature set {*f1*, *f2,hss*} and {*f6*, *f7*, *f9,hss*}, provide orderly ascending test accuracies.

Experiments show that *S4* gives a frame-level accuracy of 96.8%. What's more, the proposed feature *hss* is weakly correlated with the other features, and brings a consistent improvement to the accuracy; for example, the hierarchical oblique DT classifier plus the feature *hss,* gives an 10ms frame-level accuracy of 98.3%, which is a promising result compared to other 16ms/20ms frame-level classifiers.

5. CONCLUSIONS

This paper presents a real-time speech/music classifier. Real-time features including harmonic structure stability (hss) based on a rough harmonic structure estimation are deployed. Oblique decision trees are trained and tested for classification. Our experiments show that the proposed system outperforms the open-loop selection mode in AMR-WB+ by a frame-level accuracy of 98.3%.

The system has a low complexity of approximately 1wmops in total and can be easily deployed in application scenarios where short delay and low complexity are essential. The proposed method gives output every 10ms and can be easily extended to segment-level based applications. Future work will focus on refining the method to the voiced/unvoiced speech classification and music-genre classifications.

REFERENCES

[1] J. Pinquier, J.L. Rouas, R. Andre-obrecht, "Robust Speech/ Music Classification in Audio Documents," Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP-02), Denver, USA, vol. 3, pp. 2005-2008, Sept. 2002.

[2] C. Panagiotakis, G. Tziritas, "A Speech/Music Discriminator Based on RMS and Zero-Crossings," IEEE Trans. Multimedia, vol. 7, pp. 155-166, Feb. 2005.

[3] J. Saunders, L.M. Co, N.H. Nashua, "Real-Time Discrimination of Broadcast Speech/Music," Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP-96), Atlanta, USA, pp. 993-996, May 1996.

[4] B. Schuller, B.J.B. Schmitt, D. Arsic, S. Reiter, M. Lang "Feature Selection and Stacking for Robust Discrimination of Speech, Monophonic Singing, and Polyphonic Music," Multimedia and Expo., 2005, ICME 2005, pp. 840-843, July 2005.

[5] S. Sukittanon, L.E. Atlas, J.W. Pitton, "Modulation-Scale Analysis for Content Identification," IEEE Trans. Signal Processing, vol. 52, pp. 3023-3035, Oct. 2004.

[6] K. El-Maleh, M. Klein, G. Petrucci, P. Kabal, "Speech/Music Discrimination for Multimedia Applications," Acoustics, Speech, and Signal Proceeding, 2000, ICASSP-00, vol. 4, pp. 2445-2448, 2000.

[7] E. Scheirer, M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator," Acoustics, Speech, and Signal Proceeding, 1997, ICASSP-97, vol. 2, pp. 1331-1334, April 1997.

[8] W.Q. Wang, W. Gao, D.W. Ying, "A Fast and Robust Speech/Music Discrimination Approach," Information, Communications and Signal Processing, 2003, vol. 3, pp. 1325-1329, Dec. 2003.

[9] N. Casagrande, D.Eck, B. Kegl, "Geometry in Sound: a Speech/Music Audio Classifier Inspired by an Image Classifier," ICMC 2005, 2005.

[10] R. Kohavi, G.H. John, "Wrappers for Feature Subset Selection," Aritificial Intelligence Journal, vol. 97, pp. 273-324, 1997.

[11] 3GPP, "Technical Specification Group Service and System Aspects; Extended Adaptive Multi-Rate-Wideband (AMR-WB+) codec; Transcoding functions," 3rd Generation Partnership Project, Jun. 2005.

[12] Y.G. Zhang, C. Zhang, "Separation of Voice and Music by Harmonic Structure Stability Analysis," Multimedia and Expo, 2005, ICME 2005, pp. 562- 565, July 2005.

[13] A.G. Zeng, X.R. Liang, M.W. He, "A Novel Approach to Multimodal Function Optimization," Computer Engineering and Applications, vol. 42, pp. 73-75, 2006.

[14] S. K. Murthy, S. Kasif, S. Salzberg, "A System for The Induction of Oblique Decision Trees," Journal of Artificial Intelligence Research, vol. 2, pp. 1-32, 1994.

[15] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, CRC Press, Boca Raton, 1993.