interspeech 2008 incorporating SST 08

A Combination of Data Mining Method with Decision Trees Building for Speech/Music Discrimination

Qiong Wu[1], *Qin Yan*[2], *Jun Wang*[1], *Jun Hong*[1]

[1] Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
[2] College of Computer and Information Technology, Hohai University, Nanjing, China
{wuq, wangjun, hongj}@dsp.ac.cn, yanqin@ieee.org

10.21437/Interspeech.2008-628

Abstract

Nowadays the applications in multimedia domain require that the Speech/Music classifier has many other merits in addition to the accuracy, such as short-time delay and low complexity. Here, we endeavor to form a Speech/Music classifier by using different data mining methods. The main work of this paper is to obtain such system by analyzing the inherent validity of diverse features extracted from the audio, combining them into two subsets, and building a hieratical structure of decision trees to maintain optimal performances. The classifier is evaluated by a set of 5-to-11-minutes 450 audio files of different types of speech and music, and outperforms AMR-WB+ by achieving 97.6% and 95.2% correct classification rate at the 10ms frame level in pure and high SNR (>=20dB) environment respectively. Besides, its complexity is lower than 1WMOPS which make it easily adapted to many scenarios.

Index Terms: real time discrimination, optimal feature subset, decision tree, hieratical.

1. Introduction

In many applications, for the purpose of constructing more and more "real-world" multimedia environment, there is a strong interest in classifying audio signals, whose characterization could be categorized as one kind of speech, music, or silence. Furthermore, the discrimination of active sound signals and inactive silence segments is the mainly purpose of Voice Active Detection (VAD), which has already had many dissertations to deal with. In this work, only the discrimination of speech and music is considered.

One of the basic issues in the design of a signal classifier is the selection of an appropriate feature set that captures the temporal and spectral structures of the signals. Many such features for Speech/Music discrimination have been investigated in the literatures. For instance, Saunders [1] proposed a real-time Speech/Music discriminator, which is used to automatically monitor the audio content. Scheirer and Slaney [2] exploited thirteen features to characterize distinct properties of music and speech signals, and a correct classification percentage of 94.2% is reported for 20 ms segments and 98.6% for 2.4s segments. There are also many other worthwhile works [3] to provide more than thirty features considering different characters of speech and music.

Another issue in the system design is the selection of a classification algorithm. A variety of algorithms for Speech/ Music classification have been proposed and implemented in the past for the needs of various applications. Different classifiers like the Bayesian Information Criterion (BIC) [4], Gaussian likelihood ratio (GLR) [1, 2], Gaussian mixture model (GMM) [2], quadratic Gaussian classifier (QGC) [3], nearest neighborhood classifier [2, 3] and hidden Markov model (HMM) [5] have been used for this purpose.

In this work, Feature subset selection (FSS) tool [6] is applied to induce optimal feature subsets, which can provides both cost-effective predictors and a better understanding of the underlying process that generated the data. Meanwhile, comparing with the methods mentioned above, a hieratical oblique decision tree is proposed to work efficiently in a shorter time-delay (10ms) environment so as to cope with requirements of nowadays multimedia applications. In the end, the advantages and details of the test results discuss comprehensively.

We conclude this introduction by describing the background of the problem and its basic characteristics as utilized in our work. Section 2 describes proposed classification algorithm in detail, expatiating the principle of extraction and selection of optimal acoustic features, and comparing two types of decision tree (DT) classifiers with several kinds of feature sets to construct a frame-level hieratical framework. In section 3, the applied database and experiments evaluation are described thoroughly followed by further discussion.

2. The proposed Speech/Music classification approach

2.1. Features extraction

As it is impossible to provide a detailed introduction of all relative attributes that can be extracted from the audio stream, we focus on the key ones as an initial feature set, listed in Table 1, which has the following characters:

- These features could characterize the underlying acoustic signal, such as pitch, energy, and spectral characteristics and the time signal itself.
- They do not depend on the spoken or musical content itself.

Accepted after peer review of full paper

Table 1. *Initial features used in the investigation*

Feature	Description
Var_Flux	The variance of spectral Flux in the nearest 20 frames
Var_mov_Flux	The moving average of the latest 25 values of the Var_Flux calculated above.
Var_subflux	The variance of low frequency sub-band Flux in a single frame
Var_mov_ subflux	The moving average of the latest 25 values of the Var_subFlux.
ZCR	Zero-Cross Rate in the frequency interval 0-16kHz.
Stda_long	12 sub-bands of standard energy deviation within 16 frames
Stda_short	12 sub-bands of standard energy deviation within 4 frames
Ratio	The ratio of low band energy to whole band energy.
Var_Ratio	The variance of the Ratio parameter within 20 frames
Mov_Ratio	The moving average of the Ratio parameter within 20 frames..
Hss	Harmonic structure stability of music and speech.

Var_Flux, *Var_mov_Flux*, *Var_subflux* and *Var_mov_subflux* are derived from the spectrum "Flux" [7], which has the character that speech alternates periods of transition (consonant-vowel boundaries) and periods of relative stasis (vowels), where music typically has a more constant rate of change. The parameters about energy *Ratio* and Harmonic structure stability (*Hss*) [2] are also considered to be effectively working in the process of discrimination.

The *ZCR* [1, 2] is correlated with the spectral centroid, and considered to be the measure of dominant frequency of the interval. Together with *Stda_short* and *Stda_long*, they are all omitted in the final version of the algorithm due to high correlation with the optimal feature subset in the specific algorithm, though the reference [8] indicates their advantages.

2.2. Feature selection for an optimal subset

Before applying to any specific classifier, it is necessary to find optimal sets of features as a whole rather than a combination of stand alone high performance attributes. Only through such means, can the proposed algorithm achieve the best possible accuracy with a particular classification algorithm, as well as saves the complexity in both space and time domain.

From viewpoints of both practicality and theory, the features with longer time-delay give a better performance but cannot keep pace with the switches between speech and music timely, and vice versa for the short time-delay features. If we simply put these features to be selected by FSS, the parameters which have short time-delay but relatively high correlation with some other longer time-delay features would be eliminated from the final subset. Thus, in the test results, there would be some undesirable errors at the transformations within audio files. In Section 3, the experiments' results validate this claim.

Table 2. *Training accuracy of different short-time delay feature subset*

Rank	Accuracy	Features Combination
1	7.36%±0.15%	{Var_Flux, Var_subflux}
2	9.00%±0.17%	{Var_Flux, Var_subflux }
3	9.12%±0.17%	{Var_Flux, Var_subflux, ZCR}
4	9.95%±0.17%	{Var_Flux, Var_subflux,Ratio, ZCR}
5	13.30%±0.2%	{Var_Flux, Var_subflux, Stda_short}
6	14.73%±0.2%	{Var_Flux, Var_subflux, Stda_short, ZCR}

Table 3. *Training accuracy of different long-time delay feature subset*

Rank	Accuracy	Features Combination
1	0.34%±0.03%	{Var_mov_flux, Var_mov_subflux, Mov_Ratio}
2	0.36%±0.03%	{Var_mov_flux, Var_mov_subflux, Stda_long, Mov_Ratio}
3	0.68%±0.05%	{Var_mov_flux, Stda_long, Mov_Ratio}
4	1.12%±0.07%	{Var_mov_flux, Var_mov_subflux}
5	1.17%±0.07%	{Var_mov_flux, Var_mov_subflux, Stda_long}
6	1.81%±0.08%	{Var_mov_flux, Var_mov_subflux, Stda_long, Mov_Ratio, Var_Ratio}

Part of our major works is to propose a balance structure between the short-time delay and long-time delay features. In the first step, we divide the initial set of features into two groups according to the span of time-delay: one group is {*Var_Flux, Var_subflux, Ratio, Stda_short, ZCR*}, and the other is {*Var_mov_flux, Var_mov_subflux, Stda_long, Mov_ Ratio, Var_Ratio*}. Secondly, we selected them by means of the FSS. The FSS can be considered as a black box with induction algorithm inside running on the dataset with different sets of features removed from the data, choosing the highest-evaluation feature subset, and then evaluating on an independent test set. At last, we set a hang-over strategy, described in the next section, to balance the two kinds of discrimination possibilities.

As shown in Table 2 and 3, the short-time and long-time delay optimal feature subsets are {*Var_Flux, Var_subflux*} and {*Var_mov_flux, Var_mov_subflux, Mov_Ratio*}. However, both subsets reveal merely standard music-to-speech misjudgment rates, not as excellent as we expected. Considering the *Hss* is an effective way to detect the music, and the correlation coefficients between *Hss* and other features listed in Table 1 are only 0.15 at maximum, we add *Hss* to both of the subsets. And the experiment results shown in Section 3 indicate its validity.

2.3. Different classifier in making decision tree

Decision trees (DTs) are a useful tool for classification by a sequence of simple, easy to understand tests. The semantics meaning of the trees is intuitively clear to do further

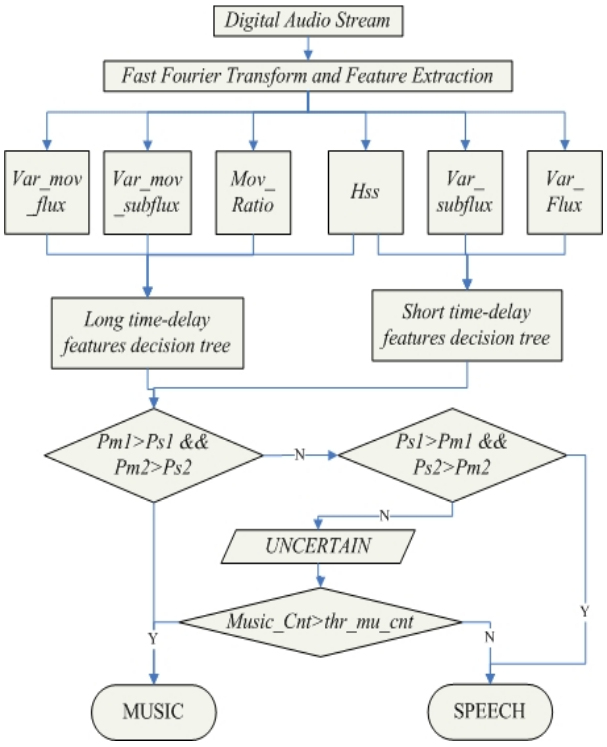

Figure. 1: *The system diagram of speech/music discrimination with hierarchical DT*

processing. In further division, oblique decision trees produce polygonal partitioning of the attribute space, while axis-parallel trees produce partitioning in the form of hyper-rectangles that are parallel to the feature axes.

For oblique method, the OC1 [9] can find a good oblique split in the form of a hyper-plane at each node of the DT. And the overall hyper-plane takes the form:

$$\sum_{i=1}^{d} a_i x_i + a_{d+1} > 0$$

where $a_1,\ldots,a_{d+1}$ are real-value coefficients and x_i are real-value attributes. Compared with the binary division of Classification and Regression Trees (CART) [10], which every single feature has relatively more overlap distribution, we assume that the OC1 could outperform CART by classifying in a multiple-dimension feature vector space, and the results in Section 3 confirm this.

For comparison, CART is also fed with identical feature sets. This type of decision tree gives several branches with ending nodes, which export respective possibilities of music and speech.

In this work, for either method, we use the audio files of pure and 20 signal-to-noise ratio (SNR) separately to induce two groups of trees, which are constructed by short-time delay and long-time delay feature subsets. A double-layers hierarchical DT structure is also deployed for each group, as shown in Figure. 1. *Pm1*, *Ps1*, *Pm2*, and *Ps2* are the possibilities of music or speech classification in training of DT1 or DT2 respectively. The hierarchical DT framework outputs three types of indicators: *MUSIC*, *SPEECH*, and *UNCERTAIN*. The classification is then refined with a counter *music_Cnt*, which calculates how many continuous frames are classified as *MUSIC* at the time being. If the real-time output is *UNCERTAIN* and *music_Cnt* is larger than the threshold *thr_mu_cnt*, which is experimentally set as 10, then the current frame is set as *MUSIC*, otherwise is set as *SPEECH*.

Table 4. *Testing performance of different algorithms in pure environment*

Testing algorithm	Accuracy of speech	Accuracy of music	Average
HODT	98.9%	96.3%	97.6%
HODT_1	99.9%	93.6%	96.8%
ODT	97.8%	94.5%	96.2%
CART	99.6%	90.0%	94.8%
AMR_WB+	93.3%	92.8%	93.1%

HODT stands for the Hierarchical Oblique Decision Tree. HODT_1 is the HODT without Hss parameter. ODT is the Oblique Decision Tree, building without the features selected by FSS. AMR_WB+ is chosen to be reference, and the result listed above is the performance of its open-loop function.

3. Experiment and discussion

3.1. Database description

The corpus of monophonic speech and music samples, which are strictly made in accordance with the ITU-T proposed standard [11], is band-limited to 16 kHz and sampled at 100Hz rate built by our lab. These audio files are obtained from National 863 Chinese Speech database and ITU-T codec test sets. Music types varies from jazz, piano, sax, folk, symphony, concerto to Chinese folk music. The speech data covers about 20 multiple speakers in English, French and Chinese from both genders. The training set contains pure and 20 SNR audios without silence and 228512 music frames and 237671 speech frames in total.

Besides, another 450 music and speech files are selected for the independent test. These files are chosen from the same database but entirely different from the files in training set. These test files are at three SNR levels: pure, 30dB and 20dB by above mentioned procedure and labeled at every 10 ms. The test files are then divided to three genres: speech only, music only and speech-and-music mixed. In the experiments, all the discrimination accuracies are calculated according to the standard label files, which are generated semi-automatically by first marking those frames that exceeded a power threshold [12], and then evaluated twice by expert listeners to adjust the speech or music frames and then check them.

3.2. Experiments evaluation

We run all the proposed structures in the pure audio environment to clarify which kind of feature subsets are preferable to the discrimination. After comparing three diverse decision-making algorithms in high SNR environment, it is clarified that the Hierarchical Oblique DT (HODT) can provide a better discernment performance. Note that the open-loop mode selection in AMR-WB+ [8], which is a standard codec in ITU-T and outputs result in every 20 ms, is chosen as benchmark algorithm. We then expand each of the outputs twice to meet the standard label files.

Table 5. *Testing performance of different algorithms in high SNR environment*

Testing algorithm	SNR (dB)	Accuracy of speech	Accuracy of music	Average
HODT	30	96.7%	94.4%	95.6%
	20	95.0%	95.4%	95.2%
CART	30	98.4%	86.6%	92.5%
	20	97.4%	83.0%	90.2%
AMR_WB+	30	89.8%	93.8%	91.8%
	20	84.9%	91.3%	91.6%

From test results in Table 4 and 5, it can be concluded that:

- By the comparison of HODT and HODT_1, harmonic structure stability (Hss) reveals improvement of music discrimination.
- By the comparison of HODT and ODT, two optimal feature subsets with double layers hierarchical structure can improve the average performances.
- All of the three oblique DTs perform better in both accuracy of music and speech than the CART algorithm and AMR_WB+.
- HODT classifier gives a frame-level accuracy of 97.6% in pure environment, which performs the best of all the tested frameworks in pure condition.
- In the high SNR condition, the HODT gives the frame-level accuracy of 95.6% in 30dB and 95.2% in 20dB, which are the best performances and slowest decrease with the SNR decreasing among the three algorithms.

3.3. Further discussion

First of all, according to the tree-quality measurements, classification or prediction accuracy, the number of leaf nodes of the decision tree and the maximum distance from the root to the farthest leaf node are all critical evaluating standards.

Although CART could do some pruning on the end of each node, it cannot provide an overall advantageous solution to the feature set, and worse still, it produces a bigger decision tree with several branches and more than 15 ending nodes to meet satisfying performance. In contrast, especially to the Speech/Music classification problem, OC1 could obtain only a polynomial to represent a specific hyper-plane for one feature subset.

Besides, a hyper-plane could adapt the system much more easily to different use scenarios than CART, for the reason that, in different level SNR circumstances, a hyper-plane just need to adjust its polynomial coefficients $a_1,\ldots,a_{d+1}$, while the fitted axis-parallel trees differ in both the shapes and each node's logics.

Furthermore, at SNR of 20dB, because the features' distributions in musical environment are all concentrated in a lower value region, which is overlapped with minor parts of speech features due to added noise, the HODT shows better performances in discriminating music than speech. On the other hand, at a given pruning level, the CART always extends much more leaf nodes to differentiate speech from music. Accordingly, it can divide speech with few mistakes at the cost of aggravating music classification.

4. Conclusions

In the paper, we propose a Speech/Music classifier based on the feature subset selection (FSS) tool and oblique decision tree induced by the algorithm OC1. The experiment results indicate that both methods above and the hierarchical structure are all effective to Speech/Music discrimination. The proposed system outperforms AMR-WB+ by a frame level accuracy of 97.6% in pure environment and over 95.2% in high SNR (20dB) environment. Furthermore, complexity of the system code is lower than 1 Weighted Million Operations per Second (WMOPS), which is the standard unit to evaluate the code complexity in 3GPP. Accordingly, it can be easily utilized into diverse application scenarios where short delay and low complexity are essential. Finally the proposed method gives output every 10ms, which can be effortlessly adapted to segment-level-based applications.

A further direction of this study will be focus on improving the accuracy in lower SNR levels and extending the algorithm to music genre categorization.

5. References

[1] 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.

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

[3] K. El-Maleh, M. Klein, G. Petrucci, P. Kabal, "Speech/Music discrimination for multimedia application," Int. Conf. Acoustic, Speech, and Signal Processing, pp. 2445–2448, 2000.

[4] S. S. Chen, P. S. Gopalkrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," IBM Tech. J, 1998.

[5] T. Zhang, J. Kuo, "Hierarchical classification of audio data, for archiving and retrieving," Int. Conf. Acoustic, Speech, and Signal Processing, pp. 3001–3004, 1999.

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

[7] Ji-Soo Keum, Hyon-Soo Lee, "Speech/Music Discrimination using Spectral Peak Feature for Speaker Indexing," Intelligent Signal Processing and Communications, ISPACS '06, Yonago Convention Center, Tottori, Japan, pp. 323-326, Dec 2006.

[8] 3GPP, "Technical Specification Group Service and System Aspects; Extended Adaptive Multi-Rate-Wideband (AMR-WB+) codec; Transcoding functions," 3rd Generation Partnership Project, Jun. 2005.

[9] 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.

[10] Kamil A. Grajski, L. Breiman, Gonzalo V. D. Prisco, Walter J. Friedman, "Classification of EEG Spatial Patterns with a Tree-Structured Methodology: CART," IEEE Transactions on Biomedical Engineering, vol. 33, NO. 12, pp1076-1086, Dec 1986.

[11] Shigeaki Sasaki, Hitoshi Ohmuro, Yusuke Hiwasaki, "Draft Processing Test Plan for ITU-T G.711 Wideband Extension Qualification Phase," ITU-T Document AC-0703-Q10-20, Geneva, 22 March 2007.

[12] TIA, "TDMA Third Generation Wireless-Minimum Performance Standards for ACELP Voice Activity Detection," Telecommunications Industry Association, Adopted proposal TIA/EIA-136-250, April 2001.