Robust Singing Detection in Speech/Music Discriminator Design

Wu Chou and *Liang Gu**

Bell Laboratories, Lucent Technologies
600 Mountain Ave., Murry Hill, NJ 07974, USA

ABSTRACT

In this paper, an approach for robust signing signal detection in speech/music discrimination is proposed and applied to applications of audio indexing. Conventional approaches in speech/music discrimination can provide reasonable performance with regular music signals but often perform poorly with singing segments. This is due mainly to the fact that speech and singing signals are extremely close and traditional features used in speech recognition do not provide a reliable cue for speech and singing signal discrimination. In order to improve the robustness of speech/music discrimination, a new set of features derived from harmonic coefficient and its 4Hz modulation values are developed in this paper, and these new features provide additional and reliable cues to separate speech from singing. In addition, a rule-based post-filtering scheme is also described which leads to further improvements in speech/music discrimination. Source-independent audio indexing experiments on PBS Skills database indicate that the proposed approach can greatly reduce the classification error rate on singing segments in the audio stream. Comparing with existing approaches, the overall segmentation error rate is reduced by more than 30%, averaged over all shows in the database.

1. INTRODUCTION

Distinguishing speech signal from music signals is one of the critical problems in intelligent multimedia information processing and extraction (IMIPE). Among various components in IMIPE system, speech/music discrimination (SMD) is of particular importance, especially for the application of audio indexing. SMD partitions the audio stream into homogeneous segments, and it removes non-speech signals from original speech/music-mixed signals for accurate automatic speech recognition and text alignment. These applications motivate the research and development of robust, high performance SMD techniques.

Statistical modeling and feature analysis are two main units in most SMD systems. Previous work on statistical modeling in SMD was reviewed and evaluated in [1]. Based on the experimental results, Gaussian mixture model (GMM) and k-nearest-neighbor classifier could be two effective approaches in SMD design.

Previous work on feature analysis in SMD can be categorized into three major categories. One is based on time-domain features (such as zero-crossing rate) [2]. Another category of approaches adopts frequency-domain features (such as mel-frequency cepstral coefficients - MFCC) popularly used in automatic speech recognition [3]. The third category of approaches exploits both time-domain and frequency-domain features (such as 4-Hz energy) to enhance accuracy and robustness of SMD [1][4]. Due to the high diversity of audio signals, combining features with different characteristics was found to improve SMD accuracy [1][3], leading to various research efforts in searching for better features for SMD applications.

While the performance of SMD with current feature analysis and statistical modeling techniques can provide good performance for simple audio signals, its accuracy on complex audio signals may degrade sharply. This is often due mainly to the high classification error rate on signing signals. Singing signal should be classified as non-speech signal or music, and be separated from normal speech signals. However, with current SMD design techniques, singing signals will mostly be recognized as speech signals. This kind of errors could have severe impact on speech/music segmentation and deteriorate the performance of caption alignment via automatic speech recognition. Therefore, robust singing signal detection is a challenge in SMD design for complex audio streams with singing segments.

In this paper, a two-stage algorithm is proposed to realize SMD with robust singing detection. The input audio streams are classified into singing and non-singing segments in the first stage, followed by conventional speech/music discrimination in the second stage. To enhance the performance of singing detection, two new features of harmonic coefficient and 4-Hz modulation value of the harmonic coefficient are developed to characterize the difference in harmonic structures between singing and non-singing segments. These features are combined with conventional 4-Hz modulation energies in our system for best classification performance. The preliminary discrimination results from the above two-stage SMD are further smoothed by a rule-based post-filtering technique. SMD experimental results based on source-independent audio indexing task indicated that the singing detection accuracy was improved significantly with the proposed new harmonic features. The application scope and SMD accuracy for complex audio signals are therefore greatly enhanced.

2. HARMONIC COEFFICIENTS FOR ROBUST SINGING DETECTION

A. Problem Statement

Singing is a special kind of audio signal that needs different treatment compared with other signals such as speech or music. The conventional short-term features are not useful, because there is no significant difference existing between normal speech and singing signals from short-term spectral analysis. Speech/music classifiers based on these features will most probably recognize singing signals as speech signals.

* The author works at Lucent as a summer intern student

	Effectiveness for Singing detection
Mel-Frequency Cepstral Coefficients (MFCC)	Low
First and second differential MFCC	Medium
Log energy	Low
First and second differential log energy	Medium
4-Hz modulation energy	High
Zero-crossing rate	Low

Figure 1. Effectiveness of conventional feature analysis methods for signing detection

Meanwhile, experiments show that the difference between singing and speech signals exists in long-term acoustic features (as shown in Figure 1). One of such features is 4-Hz modulation energy [1]. Speech has a characteristic energy modulation peak around 4-Hz syllabic rate, which is lacked in most singing signals. Hence, when the normalized energy is band-pass filtered at 4-Hz, speech signal tends to have much higher output value compared with singing signals [1]. Although this attribute is very attractive, the experimental results of singing detection based on 4-Hz modulation energy solely was not satisfactory. The main reason is that this feature is not very robust, as some speech may have low 4 Hz energy (such as fast speech) and some singing signals may have high 4 Hz energy (such as songs with 4 Hz beat).

B. Harmonic Coefficients

To improve the accuracy of singing detection, we developed a new feature called *Harmonic Coefficient* to represent the characteristics of harmonic structures of voiced speech, which is calculated by the average maximum auto-correlation value in time-domain and frequency-domain. This method was first used in speech coding area for accurate pitch estimation [5].

Given a speech signal $s_t(n)$, the temporal auto-correlation (TA) for candidate pitch t is defined as

$$R^T(t) = \frac{\sum_{n=0}^{N-t-1} [\tilde{s}_t(n) \cdot \tilde{s}_t(n+t)]}{\sqrt{\sum_{n=0}^{N-t-1} \tilde{s}_t^2(n) \cdot \sum_{n=0}^{N-t-1} \tilde{s}_t^2(n+t)}}$$

where $\tilde{s}_t(n)$ is the zero-mean version of $s_t(n)$, and *N* is the number of samples for feature analysis.

The spectral auto-correlation (SA) is defined as

$$R^S(t) = \frac{\int_0^{p-w_t} \tilde{S}_f(w)\tilde{S}_f(w+w_t)dw}{\sqrt{\int_0^{p-w_t} \tilde{S}_f^2(w)\int_0^{p-w_t} \tilde{S}_f^2(w+w_t)dw}}$$

where $w_t = 2p/t$, $S_f(w)$ is the magnitude spectrum of $s_t(n)$, and $\tilde{S}_f(w)$ is the zero-mean version of $S_f(w)$.

To improve robustness, spectral-temporal auto-correlation (STA) is defined as:

$$R(t) = b \cdot R^T(t) + (1-b) \cdot R^S(t)$$

where $b = 0.5$ was found to yield good results in practice [5].

The Harmonic coefficient H_a in this paper is defined as

$$H_a = \max_t R(t) .$$

H_a is much higher for voiced speech than that for unvoiced speech. This attribute can be used to improve the performance of SMD. While the application of harmonic coefficient can enhance the performance of normal speech/music discrimination, greater benefit of the new feature appears in the process of singing detection. As what we have discussed before, traditional singing detection is not satisfactory because of the low robustness of 4-Hz modulation energy. The main problem involved is the serious overlap of feature distribution between speech signals and singing signals, which greatly reduces the efficiency of statistical modeling. The feature of harmonic coefficient is proposed to ease this difficulty. When H_a is low, the discrimination between singing and non-singing signals will largely depend on the value of 4-Hz energy. When H_a is high, we can loose the threshold of 4-Hz energy as the input tends to be a signing signal. This process can be automated via GMM training technique, which will be discussed in next section.

C. 4-Hz Modulation Harmonic Coefficients

The accuracy of singing detection can be further enhanced by 4 Hz modulation harmonic coefficients. Similar to 4-Hz modulation energy, speech signals have a characteristic harmonic feature modulation peak around 4-Hz syllabic rate, as the signals periodically change between voiced speech (high H_a) and unvoiced speech (low H_a). Singing segments usually have long duration of consonants with very high H_a and low 4-Hz modulation value. As illustrated in Figure 2, from which we can see that 4-Hz modulation value of harmonic coefficient can be used as a complementary feature with full-band harmonic coefficient and 4-Hz modulation energy in singing detection.

The 4-Hz modulation harmonic coefficient can be calculated as follows:

Audio Type	4 Hz Energy	Harmonic Coefficients	4 Hz Harmonic Coefficients
Singing	Low	High	Very Low
Speech	High	Medium	High
Music	Low - Medium	Low	Low

Figure 2. Illustration of feature value distributions for different audio signals

- Computation of harmonic coefficient H_a for every 10ms with 25ms Hamming window;
- Frequency analysis by discrete cosine transform (DCT) for every 500 ms;
- 4-Hz band-pass digital filtering based on DCT output.

3. TWO-STAGE SPEECH/MUSIC DISCRIMINATION

A. Two-stage SMD

The above three features (4-Hz modulation energy, harmonic coefficient and 4-Hz modulation value of harmonic coefficient) can be applied along with conventional features (such as MFCC and log energy) in a GMM-based SMD system. However, from Figure 1. and Figure 2., the effectiveness of these features differs significantly for the discrimination between speech and music, and between singing and non-singing signals. We devised a two-stage SMD method to exploit the merits of new features for singing detection, as shown in Figure 3. In the first stage, discrimination is carried out for singing and non-singing signals, based on the proposed features of harmonic coefficient and its 4-Hz modulation value in addition to other features. The GMM models used in the first stage are built from the labeled singing and non-singing audio samples. In the second stage, conventional discrimination between speech and music is implemented over pre-filtered non-singing segments, where more features, such as MFCC and log energies, are used. The GMM models are trained on pure music and clean speech samples. The result speech, music or singing segments are further smoothed by a rule-based post-filtering technique, which will be discussed in next sub-section.

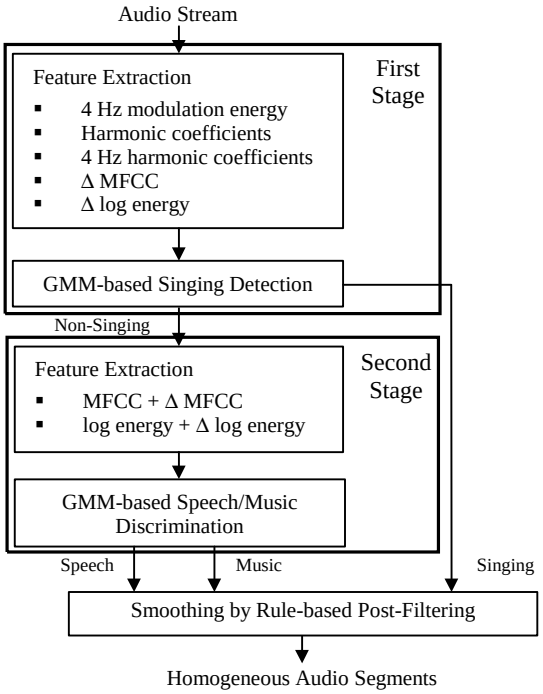

Figure 3. Two-stage speech/music discriminator with signing detection

Note that since the two 4 Hz modulation features are long-term features (500ms) compared with other features (25ms), the algorithm delay in the first-stage discrimination is increased, and so does the total algorithm delay of the two-stage SMD compared with SMD with no 4 Hz features. This is reasonable considering the fact that even human ear may not be able to recognize singing signals from normal speech within a very short time duration. On the other hand, we think that half-second delay is short enough for applications in audio indexing or caption alignment.

B. Smoothing via Rule-based Post-Filtering

The result after audio type classification is a set of segments labeled as speech, music, singing or noise. In this segmentation stage, all these classes are treated separately in order to achieve near-real-time performance. These segmentation results can be further smoothed into homogeneous segments according to the class-correlation information in time-domain.

In our work, we proposed a rule-based post-filtering method to fulfill the smoothing task. If we represent speech as "S", music as "M", singing as "A", noise as "N", and let "_" represent any audio type except noise, the smoothing rules can be defined as:

- N_N → NNN
- SSMSS → SSSSS, SSASS → SSSSS
- MMSMM → MMMMM, MMAMM → MMMMM
- AAMAA → AAAAA, AASAA → AAAAA
- NMSSS → NSSSS, NASSS → NSSSS, SSSMN → SSSSN, SSSAN → SSSSN
- NMAAA → NAAAA, NSAAA → NAAAA, AAAMN → AAAAN, AAASN → AAAAN
- NN__NN → NNNNNN
- SSMMSSS → SSSSSSS, SSSMMSS → SSSSSSS
- NNAASSS → NNSSSSS, NNMMSSS → NNSSSSS,
- NNAMSSS → NNSSSSS, NNMASSS → NNSSSSS,
- SSSAANN → SSSSSNN, SSSMMNN → SSSSSNN,
- SSSAMNN → SSSSSNN, SSSMANN → SSSSSNN,
- ASA → AAA

Application of a rule is repeated as long as the segmentation changes, before the next rule is used. After a complete loop over all rules, the loop is repeated, until the segmentation remains unchanged. In practice, these rules can be implemented over a 10s span for off-line indexing applications.

4. EXPERIMENTAL RESULTS

The above new approaches are evaluated in both source-dependent and source-independent SMD experiments. In source-dependent experiment, PBS skills database is used with 26 minutes speech, music and singing signals (16-bit monophonic samples at 16 kHz sampling rate). A cross-validated testing framework is used to evaluate the source-dependent classification performance. In this method, 20% of the labeled samples, selected at random, are held back as test data, and a classifier trained on the remaining 80% of the data. This classifier is then used to classify the test data, and the results are

Audio Type	Error Rate (%)	MFCC + E	MFCC+ ΔMFCC + E+ΔE	MFCC+E +ΔMFCC+ΔE +ΔΔMFCC +ΔΔE
With Singing	Speech	7.4	6.7	16.3
	Music	55.2	30.0	15.4
	Average	31.3	18.3	15.9
Without Singing	Speech	7.4	6.7	16.3
	Music	35.0	10.0	5.0
	Average	21.2	8.4	10.7

Table 1. Performance of conventional SMD for audio streams with or without singing signals under varied feature configurations

Features	Singing segments missed (out of 44)	Error Rate
MFCC+ ΔMFCC+ E+ΔE	30	68 %
+ 4 Hz modulation energy	20	45 %
+ Harmonic coefficients	8	18 %
+ 4 Hz harmonic coefficients	6	14 %
+ Rule-based post-filtering	0	0 %

Table 2. Performance improvement of Singing Detection via new feature analysis methods

	Speech error rate	Music error rate	Average error rate
Before post-filtering	7.6 %	13.0 %	10.3 %
After post-filtering	6.3 %	1.5 %	3.9 %

Table 3. SMD Performance comparison (before and after post-filtering) based on PBS skills audio indexing task

compared to the labels to determine the accuracy of the classifier at frame levels. By iterating this process several times (300 times in our experiment) and evaluate the classifier based on the aggregate average, robust experimental results can be obtained without strong dependency on the particular test and training sets being used. In source-independent experiment, the classifier is trained on independent and very different database (i.e. 30-minute CNBS news) and tested on PBS Skill database.

In our first experiment, we tried to find out the main cause of speech/music discrimination error on source independent PBS skills audio indexing task. 12-dimension MFCC were used along with log energies, plus their first-order and second-order differential values. The analysis frame width was 25ms, the analysis frame step was 10ms, and a Hamming Window was used. GMM models with 8 Gaussian pdfs per model were adopted as statistic models. The experimental results are shown in Table 1. We found out that, under various configurations, if the high error rate on singing segments can be discarded, the overall classification error rate can be reduced significantly (32% ~ 54%), indicating that classification error rate reduction on signing segments is critical.

In our second experiment, we evaluated our new approaches for robust singing detection on PBS skills source-independent task. 8-Gaussian pdf GMM models were used in all these experiments. The baseline system adopted 26-dimension features of MFCC, log energies and their dynamic values. Due to the singing segments in PBS skills database, the error rate based on these features was 68%, which was far from satisfactory. When 4-Hz modulation energy feature was added, the error rate decreased to 45%, which was still very high. The performance of singing detection was greatly improved after the implementation of harmonic coefficient and its 4-Hz modulation value, from which the error rate was reduced to 14%. These errors could be finally removed via the rule-based post-filtering technique as shown in Table 2.

In our third experiment, the rule-based post-filtering smoothing technique was further evaluated on the PBS skill source-dependent audio indexing task. Experimental results in Table 3 illustrated that classification error rates on both speech and music segments are reduced significantly after smoothing, leading to about 62% error rate reduction. This improvement came from the appropriate usage of the quasi-stationary characteristics of speech signals embedded in the proposed scheme.

5. CONCLUSION

Conventional speech/music discrimination techniques used in audio indexing ignore the different characteristics of singing signals from other music signals, and SMD performance degrades sharply if singing signals are involved. In the proposed two-stage speech/music discrimination approach, singing signals are separated from other signals in the first stage, through a robust signing detection scheme using the new harmonic coefficient based features presented in this paper. The normal speech/music classification is then performed in the second stage. New features of harmonic coefficient and its 4-Hz modulation values significantly improve the singing detection accuracy. A rule-based post-filtering smoothing algorithm is designed to further remove classification errors. Source-dependent and source-independent audio indexing experiments indicated that the new approaches could greatly reduce the classification errors caused by singing signals. The performance and the application scope of SMD are, therefore, significantly enhanced for complex audio streams.

6. REFERENCES

[1] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator", Proc. ICASSP'97, pp.1331-1334, 1997.

[2] J. Saunders, "Real-time discrimination of broadcast speech/music", Proc. ICASSP'96, pp.993-996, 1996.

[3] T. Hain, S. Johnson, A. Tuerk, et. al., "Segment generation and clustering in the HTK broadcast news transcription system" Proc. 1998 Broadcast News Transcription and Understanding Workshop, pp.133-137.

[4] M. Carey, E. Parris and H. Lloyd-Thomas, "A comparison of features for speech/music discrimination" ", Proc. ICASSP'99, pp. 1432-1436..

[5] Y. D. Cho, M. Y. Kim and S. R. Kim, "A spectrally mixed excitation (SMX) vocoder with robust parameter determination", Proc. ICASSP'98, pp. 601-604, 1998.