Discrimination of Speech, Musical Instruments and Singing Voices Using the Temporal Patterns of Sinusoidal Segments in Audio Signals

Toru Taniguchi†, Akishige Adachi†, Shigeki Okawa‡, Masaaki Honda¶, Katsuhiko Shirai†*

†Dept. of Computer Science, Waseda University, Japan
* NEC Corporation, Japan
‡Dept. of Information and Network Science, Chiba Institute of Technology, Japan
¶School of Sport Sciences, Waseda University, Japan
† {tani,adachi,shirai}@shirai.cs.waseda.ac.jp, ‡ okawa@net.it-chiba.ac.jp, ¶hon@waseda.jp

Abstract

We developed a method for discriminating speech, musical instruments and singing voices based on sinusoidal decomposition of audio signals. Although many studies have been conducted, few have worked on the problem of the temporal overlapping of the categories of sounds. In order to cope with such problems, we used sinusoidal segments with variable lengths as the discrimination units, although most of traditional work has used fixed-length units. The discrimination is based on the temporal characteristics of the sinusoidal segments. We achieved an average discrimination rate of 71.56% in classifying sinusoidal segments in non-mixed audio data. In the time segments, the accuracy 87.9% in non-mixed-category audio data and 66.4% in 2-mixed-category are achieved. In the comparison of the proposed and the MFCC methods, the effectiveness of temporal features and the importance of the use of both the spectral and temporal characteristics were proved.

1. INTRODUCTION

Advances in computer network technology have enabled people to access a huge amount of various data. The efficient retrieval of multimedia contents requires extracting useful information from these contents and then attaching it as meta-data to the original contents automatically or semi-automatically. The meta-data depends on the category of the original data. For example, if the original data is speech, the transcription (linguistic information) of the speech is useful as its meta-data; if music, the score of it is useful. The media processing necessary for attaching the meta-data to a given piece of audio data that includes speech, a singing voice, a musical instrument, and background noise, requires the discrimination of the categories of the audio contents as well as recognition of the meta-data of the audio data. In this paper, we focus on the categorical discrimination of audio data.

Over the past few years, several studies have been made on speech and music discrimination (SMD). Many studies on SMD have focused on the acoustic features used for categorical discrimination. They are classified into frequency-domain and time-domain features, and a third combined one. Frequency-domain features characterize the spectral envelope or the harmonic structure like the spectral centroid[1], MFCC (Mel-frequency cepstral coefficients), or Harmonic Coefficients[2] of speech/music signals. Time-domain features represent the temporal characteristics of speech/music signals like the zero-crossing rate[3]. The combination of the frequency and time domain features such as Spectral 'Flux' (it was later extended as Cepstrum 'Flux'[4]) and 4-Hz modulation energy[1] has also been suggested.

One of the problems of these conventional features is that they do not cope with the temporal overlapping of sounds. The overlapping problem involves two aspects. One is the overlapping of the different categories: speech and music, and the other is the overlapping of sound events in the same category, such as multiple instruments in a music segment and multiple speakers in a speech segment. The sinusoidal decomposition of the audio signals is a promising approach for coping with these problems. Because the sinusoidal components are a connected temporal sequence of the sinusoids (segment in the time-frequency domain), each sinusoidal segment can be discriminated into each sound category based on its temporal characteristics. For instance, the length of utterance of a singing voice tends to be longer than that of the speaking voices (speech). Thus, the sinusoidal segments with variable lengths can be used as the discrimination unit while, until now, most traditional features were extracted from a fixed length segment.

In this paper, we describe a method of discriminating between speech, musical instruments and singing voices based on variable length sinusoidal segments. The discrimination procedures are described in section 2 and in section 3, we evaluate and discuss the experiments and results.

2. METHOD

2.1. System Overview

Figure 1 illustrates the procedures for discriminating musical instruments, singing voices and speech. First, an audio signal is analyzed with STFT (Short-time Fourier transform), and then the spectral peaks that are dominant in power are selected from each analysis frame. Next, the peaks are connected using constraint of temporal continuity of power and frequency values in adjacent analysis frames. The connecting operation produces sinusoidal segments that are a series of joined peaks and correspond to the temporal trajectories of the fundamental frequency(F_0) or its harmonic of voices or music instruments. The temporal features that represent the shape of a sinusoidal segment are then extracted. They represent temporal changes of audio signal, especially temporal changes of F_0s or harmonics. Finally, the sinusoidal segments are integrated into time segments, and classified into three categories, musical instrument, singing voices and speech. When time segments are classified, mixed-categories, such as instruments and singing, is considered.

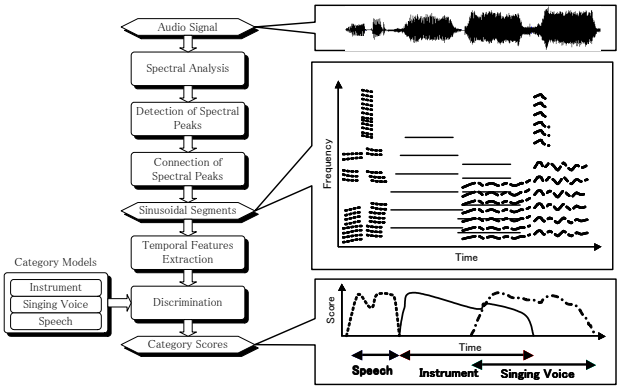

Figure 1: Overview of the discrimination system.

2.2. Detection of Spectral Peaks

The first procedure of the system seeks to choose candidates for sinusoidal points in each analysis frame for use in the next step. It consists of three processes: spectral analysis, smoothing, and choosing spectral peaks.

Spectral analysis(STFT) was performed on an input audio signal with a sampling rate of 16 kHz, Hamming windows with widths of 32 ms (512 points) and frame rates of 16 ms (256 points). In order to attain high-frequency resolution of about 4Hz, STFT was performed for the signal of 4096 samples by adding zero samples. The power spectrum obtained with STFT was then smoothed by using the Gaussian window (17 cent standard deviation) on a cent frequency scale[6]. Cent is defined as

$$f_{\mathrm{cent}} = 1200 \cdot \log_2 \frac{f_{\mathrm{Hz}}}{440 \cdot 2^{\frac{3}{12}-5}} \tag{1}$$

where f_{hz} is frequency in hertz and f_{cent} is the converted frequency in cent.

The spectral peak was determined as a local peak of the smoothed spectrum. These peaks are candidates for sinusoidal segments.

2.3. Connection of Spectral Peaks

A sinusoidal segment is determined as a sequence of spectral peaks. We connect the peaks in a frame next to each other on the basis of the temporal continuity of magnitude and frequency values, and determine sinusoidal components by finding the optimum solution of the cost function that represents temporal continuity in terms of the magnitude and the frequency values of the spectral peaks. The connecting algorithm is a DP-based method[5] suggested by Sakakibara et al.

The cost function of the continuity is defined as

$$d(t) = \sqrt{\left(\frac{f_t - f_{t-1}}{C_f}\right)^2 + \left(\frac{p_t - p_{t-1}}{C_p}\right)^2} \tag{2}$$

where f_t and p_t denote the peak frequency and the peak magnitude at the frame t. If $d > d_{th}$, the spectral peaks are not connected. C_f, C_p and d_{th} are experimentally determined as $C_f = 100$, $C_p = 3$ and $d_{th} = 1$. For this continuity criterion, the spectral peaks are not connected if the difference of adjacent frequency values exceeds 100 cent, or if the difference of the adjacent peak magnitudes exceeds 3.

2.4. Extraction of Temporal Features

Various temporal features are extracted from the sinusoidal segments. These features are as follows.

- Duration of a sinusoidal segment (#1)
- Standard deviation of cent frequencies(#2), log powers(#3) and its proportions in a frame(#4)
- Mean(#5) and standard deviation(#6) of differences in cent between frequency and the nearest note in equal temperament:

  The difference can be calculated as

  $$\mathit{diff} = (f + 50) \pmod{100} - 50 \tag{3}$$

  where diff is the difference and f is the sinusoid frequency in cent. (The interval of neighboring notes is always 100cent in equal temperament.)
- Mean and Standard deviation of the difference between two sinusoids in neighboring or two neighboring frames in cent frequency, in log power and in its proportion in a frame (#7-18)
- Symbolic representation of a sinusoidal segment in cent frequency(#19) and in log power(#20):

  Each mean is calculated for the 5 frames of the head, the middle and the tail of a sinusoidal segment in cent frequency and in log power. If the mean of the 5 frames is more than 10 cent higher than the mean of the former 5 frames, the symbol 'H' is used. If lower than 10 cent, 'L' is used. In the other cases, 'M' is given. For example, 'HM', 'HL' and 'MM' are representations of components. In the case of log power, 0.2 is the threshold value.

2.5. Classification

A time segment of an audio signal comprises multiple sinusoidal segments. We devised a method to discriminate the time segment by integrating the sinusoidal segments statistically.

We defined the problem as selecting a category C_{mix} to maximize the likelihood $p(X(t)|C_{mix})$ for the given audio segment $X(t)$. Since we assume mixed-category audio signals, a category C_{mix} may be a mixture of two or three non-mixed categories. A time segment $X(t)$ comprises the sinusoidal segments derived from the multiple categories. Then, we decompose the likelihood $p(X(t)|C_{mix})$ of mixed category C_{mix} as follows:

$$p(X(t)|C_{mix}) = p(s_1(t), s_2(t), \cdots, s_n(t)|C_{mix}) \tag{4}$$

where $s_k(t)$ is a sinusoidal segments and n is the number of the segments at the time t. Assuming that the sinusoidal segments are independent of each other, the likelihood is calculated using equation 5 by finding a category c_k to maximize the likelihood of each segment as

$$p(s_1(t), s_2(t), \cdots, s_n(t)|C_{mix}) = \prod_{k=1}^{n} \max_{c_k \in C} p(s_k(t)|c_k) \tag{5}$$

where $C = \{instruments, singing, speech\}$ is the category set. The likelihood $p(s_k(t)|c_k)$ is represented by the Gaussian mixture model (GMM) that is trained by the category-labeled sinusoidal segments data.

Table 1: Experimental data sets: I (**I**nstruments), V (singing **V**oices) , S (**S**peech).

	samples	# sinusoidal segments	duration (s)
Closed, cross-validation test data / Training data	Instruments	10570	460
	Singing	8785	600
	Speech	10074	230
Non-overlapped test data	Instruments	135	10
	Singing	166	10
	Speech	491	10
Overlapped test data	I & V	225	10
	I & S	301	10
	V & S	263	10

3. EXPERIMENTS

3.1. Data Sets

We used the data sets listed in Table 1 in the experiments described in the following sections, which were obtained from the databases listed below.

- RWC Music Database of Popular, Classical, and Jazz Music [9]: It contains music pieces that achieve quality as high as that of commercially distributed music. We selected instrumental pieces.
- Corpus of Spontaneous Japanese [11]: It contains spontaneous speech in Japanese. We selected speech pieces from it.
- Originally recorded music data: The former databases contain only a few singing voice pieces that are not overlapped with instrumental sounds; Therefore, we recorded original music containing only singing voices.

3.2. Classification of Sinusoidal Segments

We conducted an experiment to classify the extracted sinusoidal segments into three categories. We adopted two classifiers to classify the segments; the binary decision tree method and the GMM method. The two classifiers were trained by labeled training data using the C4.5 algorithm [7] and the EM algorithm[8], respectively. The GMM tests were examined by varying the number of mixtures from 1 to 16. The two classifiers were evaluated in both closed data test and 10-fold cross-validation test. In the closed data test, all the labeled samples listed in Table 1 were used to build the classifiers of the three categories and to discriminate themselves. In the 10-fold cross-validation test, all the labeled samples of each category were randomly divided into 10 sets and the 10 tests were carried out. One set selected at each test was used as the test data, and the remains were used to train the classifier. The rates of classification have been calculated for each tests, and aggregated.

Table 2 shows the accuracy of the classification of sinusoidal segments. In the cross-validation test, the performance of GMMs is better than that using the C4.5 decision tree for every number of mixtures. Since the overall accuracy of the classification with the C4.5 tree in the cross-validation test is much lower than that in the closed test, over-estimation would happen. Increasing the number of mixtures of the GMM results the improvement of the performance. Therefore, it is expected that the distribution of the features of sinusoidal segments is complicated. The best accuracy is 74.41% in the closed test and 71.56% in the cross-validation test when both tests used 16-mixture GMM.

Table 2: The accuracy (%) of the classification of sinusoidal segments with the C4.5 decision trees and the GMMs with various number of mixtures.

	classifier	inst.	singing	speech	overall
Closed test	C4.5 tree	79.59	54.48	60.55	89.08
	GMM 1	88.35	39.23	68.24	66.80
	GMM 2	90.01	53.81	57.29	68.00
	GMM 4	82.14	61.45	62.53	69.25
	GMM 8	84.41	64.92	69.48	73.48
	GMM 16	83.80	71.71	66.90	74.41
10-fold cross-validation test	C4.5 tree	78.59	54.48	60.55	65.22
	GMM 1	88.33	38.95	68.14	66.68
	GMM 2	88.45	54.56	58.07	67.93
	GMM 4	80.93	58.82	67.03	69.57
	GMM 8	82.97	62.63	65.79	71.02
	GMM 16	82.90	64.08	66.18	71.56

3.3. Classification of Time Segments

3.3.1. Comparison of Methods

We evaluated the discrimination method for classifying the time segments described in section 2.5 by comparing to a spectral method using the MFCC. The MFCC represents the spectral envelope of a sound while our proposed method uses the temporal characteristics of a sound. The evaluation is performed in the 10-fold cross-validation test as described in section 3.2 with 10,000 samples for each category. The data set was composed of non-mixed sounds unlike the date sets described in section 3.1, and was obtained from the musical instrument database[10] (both instruments and singing voices) and the speech database[11]. Table 3 shows the accuracy of classification by the 5 methods. In all the *MFCC* methods, the GMM classifiers with 16 mixtures were used. In the *MFCC_E_D* method, MFCC, logarithmic energy and their time derivatives were used as the acoustic parameters. The *SS+MFCC* and the *SS+MFCC_E_D* are the statistical integration of two classifiers by adding the log likelihood calculated by the model in each of the methods. The classification was conducted in one analysis frame in all the method.

The accuracy in the sinusoidal method, *SS*, is 93.96%, and it is superior to those of *MFCC* and *MFCC_E_D*. In the integrated methods, *SS+MFCC* and *SS+MFCC_E_D*, the accuracy rises to 95.88% and 96.12% respectively. These results prove the independence of the spectral and the temporal acoustic features, and the importance of the use of both the spectral envelope and the temporal characteristics of sinusoidal segments for the classification of the speech and music.

3.3.2. Classification of Overlapped Data

We evaluated the proposed method on overlapped data. In advance, we classified the non-overlapped data derived from the same sources for comparison. Then, we classified three 2-mixed-category data; {*instruments and singing*}, {*instruments and speech*} and {*singing and speech*}, into the same three mixed categories. The test data sets used in the experiments were the 10 s audio segments of the three categories and the three 2-category mixtures whose volume of sound was normal-

Table 3: The accuracy (%) of the classification of time segments with the sinusoidal segment, the MFCC and their integrated features: MFCC_E_D means 12 MFCC coefficients, logarithmic speech energy and their time derivatives. SS means sinusoidal segment.

method	inst.	singing	speech	overall
MFCC	92.60	86.10	96.10	91.60
MFCC_E_D	89.62	92.25	98.50	93.46
SS	91.28	92.40	98.20	93.96
SS + MFCC	94.50	94.25	98.88	95.88
SS + MFCC_E_D	94.62	94.62	99.13	96.12

Table 4: Classification ratio (%) for the sinusoidal segments and the time segments of the non-mixed data.

samples		classified into			overall accuracy
		inst.	singing	speech	
Sinusoidal segments	inst.	**91.9**	3.0	5.2	69.1
	singing	6.6	**65.1**	28.3	
	speech	10.0	25.9	**64.2**	
Time segments	inst.	**100.0**	0.0	0.0	87.9
	singing	0.3	**84.6**	15.0	
	speech	0.3	20.5	**79.2**	

ized with each other when mixed. The classification was examined at every 16 ms time segments, which is of the same length as that of the STFT analysis frame rate in section 2.2. The GMM classifier with 16 mixtures was used, and it was trained with the data set shown in table 1 as "Training data".

Table 4 lists the non-mixed data classification results from the first experiment. For comparison, the classification ratio of the sinusoidal segments from the same data was presented. The overall accuracy for the time segments was much higher than that for sinusoidal segments. With regard to the individual categories, the instruments achieved very high accuracy although there was a considerable classifying confusion between singing and speech. This result indicates that the classification of the time segments inherited the classifying manner from that of the sinusoidal segments. However, the statistical integration of sinusoidal segments into time segments improves the performance of the discrimination system.

Table 5 lists the 2-mixed-category data classification results from the second experiment. The overall accuracy is 66.4%. The accuracy of *I & V* is 72.8% and is the highest one of all the categories. This result indicates that there was a lot of mis-discrimination between the categories of singing voice and speech. It was observed that errors in the step of the analysis of sinusoidal segment caused the errors of discrimination in mixed data.

Table 5: The 2-mixed-category data classification ratio (%) of the time segments using the GMM with 16 mixtures: I (**I**nstruments), V (singing **V**oices), S (**S**peech).

samples	classified into			overall accuracy
	I & V	I & S	V & S	
I & V	**72.8**	19.5	7.7	66.4
I & S	29.0	**64.3**	6.7	
V & S	26.2	11.7	**62.1**	

4. CONCLUSIONS

We proposed a method of discriminating sounds of musical instruments, singing and speech based on sinusoidal decomposition using temporal characteristics of each category. We achieved the discrimination accuracy of 71.56% in classifying the sinusoidal segments into the three categories using the GMM classifier, 87.9% in classifying of the time segments into non-mixed-categories and 66.4% in classifying of the time segments into 2-mixed-categories. In the comparison of the proposed and the MFCC methods, the effectiveness of temporal features and the importance of the use of both the spectral and temporal characteristics were proved.

5. ACKNOWLEDGMENTS

This study was supported by the Advanced Research Institute for Science and Engineering of Waseda University under the project "A Study on Integrated Human Interface for Multimodal Information Space."

6. REFERENCES

[1] E. Scheirer, M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator.", Proc. of ICASSP'97, vol. II, pp.1331-1334, Apr. 1997.

[2] W. Chou, L. Gu, "Robust Singing Detection in Speech/Music Discriminator Design.", Proc. of ICASSP 2001, vol. II, pp.865–868, May 2001.

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

[4] S. Takeuchi, M. Yamashita, T. Uchida, M. Sugiyama, "Optimization of Voice/Music Detection in Sound Data.", Consistent & Reliable Acoustic Cues for sound analysis(CRAC workshop), Sep. 2001.

[5] K. Sakakibara, N. Osaka, "On Concatenation of Musical Sounds using a Sinusoidal Model.", Technical Report of IEICE, SP97-108, pp.1-6, Feb. 1998.(in japanese)

[6] M. Goto, "A Predominant-F0 Estimation Method for CD Recordings: MAP Estimation using EM Algorithm for Adaptive Tone Models.", Proc. of ICASSP 2001, vol.V, pp.3365-3368, May 2001.

[7] I. H. Witten, E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques With Java Implementations.", Morgan Kaufmann Publishers, 6.1, pp.159–169, 2000.

[8] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm.", Journal of the Royal Statistical Society, B39 (1), pp 1–38, 1977. Morgan Kaufmann Publishers, 2000.

[9] M. Goto, H. Hashiguchi, T. Nishimura, R. Oka, "RWC Music Database: Popular, Classical, and Jazz Music Databases.", Proc. of ISMIR 2002, pp.287-288, Oct. 2002.

[10] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC Music Database: Music Genre Database and Musical Instrument Sound Database", Proc. of ISMIR 2003, pp.229–230, Oct. 2003.

[11] K. Maekawa, H. Koiso, S. Furui, H. Isahara, "Spontaneous speech corpus of Japanese.", Proc. of LREC2000, 947-952, 2000.