Discrimination between Singing and Speaking Voices

Yasunori Ohishi[†], *Masataka Goto*[‡], *Katunobu Itou*[†], *and Kazuya Takeda*[†]

[†]Graduate School of Information Science , Nagoya University
ohishi@sp.m.is.nagoya-u.ac.jp, itou@is.nagoya-u.ac.jp, kazuya.takeda@nagoya-u.jp
[‡]National Institute of Advanced Industrial Science and Technology (AIST)
m.goto@aist.go.jp

Abstract

Discriminating between singing and speaking voices by using the local and global characteristics of voice signals is discussed. From the results of subjective experiments, we show that human beings can discriminate singing and speaking voices with more than 70% and 95% accuracy from 300 ms and one second long signals, respectively. From the subjective experiment results, assuming that different features are effective for short-term and long-term signals, we designed two measures using a spectral envelope (MFCC) and the fundamental frequency (F0, perceived as pitch) contour. Experimental results show that the F0 measure performs better than the spectral envelope measure when the input voice signals are longer than one second. Particularly, it can discriminate singing and speaking voices with more than 80% accuracy with two-second signals. On the other hand, when the input signals are shorter than one second, the spectral envelope measure performs better than the F0 measure. Finally, by simply combining the two measures, more than 90% accuracy is obtained for two-second signals.

1. Introduction

Sounds from the human mouth include various acoustic events such as speaking, singing, laughing, coughing, whistling, and lip noises. The discrimination of these sounds contributes to avoiding spoken dialogue system errors and to understanding human speech communication.

Among such varieties of acoustic events, this paper focuses on the discrimination between singing and speaking voices. Many research results have reported the differences between singing and speaking voices [1, 2, 3, 4, 5]. Typical characteristics of the singing voice include: F0 and intensity vary widely; F0 is constrained by the equal temperament; the spectral envelope of the singing voice has extra formant [6]; the F0 of a singing voice has greater power than a speaking voice, etc.

Although some algorithms that discriminate "music" and "speech" were reported [7, 8, 9, 10, 11, 12, 13], they had difficulty discriminating singing and speaking voices because they dealt with the "music" category consisting of only instrumental sounds and singing with accompaniment sounds and depended on their spectral characteristics. The characteristics of the singing voice without accompaniments have not been fully discussed yet. Therefore, the goal of this study is to characterise the nature of the singing voice and build a measure to discriminate it from the speaking voice.

The rest of the paper consists of the following sections. In Section 2, the human performance of discrimination between singing and speaking is discussed based on a subjective experiment. In Section 3, some signal measures for discriminating singing and speaking voices are proposed. After introducing the test samples in Section 4, experimental evaluations are shown in Section 5. Section 6 discusses the results.

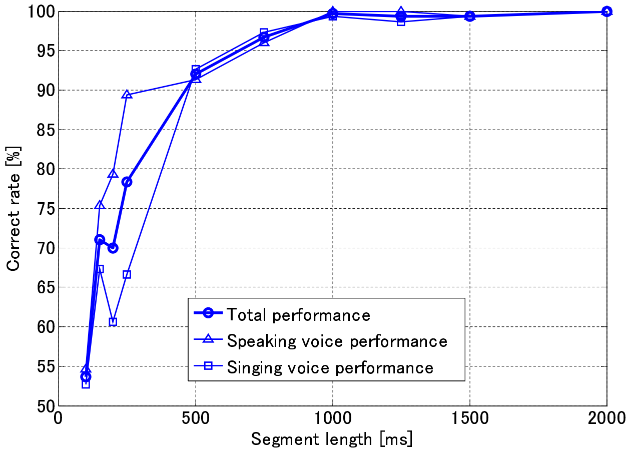

Figure 1: Human discrimination performance between singing and speaking voices.

2. Human Performance of Discriminating Singing and Speaking Voices

We first investigate the segment length necessary for human listeners to discriminate singing and speaking voices by conducting subjective experiments using 50,000 voice signals (25 male and 25 female speakers, extracted from 25 different songs of 10 different lengths). Ten subjects listened to 500 signals (250 singing voices and 250 speaking voices) randomly extracted from those 50,000 voice signals and answered whether the voice was singing or speaking.

The results shown in Figure 1 show that approximately one second is enough for a human to discriminate between singing and speaking. Even with a 300-millisecond signal, discrimination accuracy is more than 70%. This suggests that not only the long-term characteristics corresponding to rhythm and melody but also such short-term features as spectral envelopes carry the discriminative features between singing and speaking voices. Based on these observations, in the sections below, we will develop two measures for discriminating singing and speaking voices.

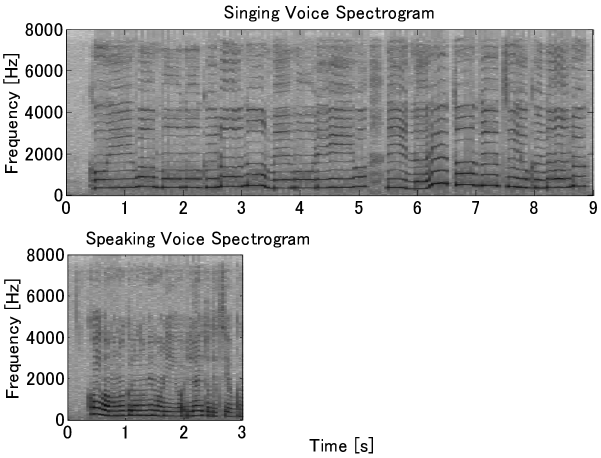

Figure 2: Spectrogram of singing and speaking voices corresponding to the same lyrics.

3. Discrimination Measures

In this section we propose two different measures — local and long-term feature measures — for discriminating singing and speaking voices. The local feature measure exploits the spectral envelope represented by using Mel-Frequency Cepstrum Coefficients (MFCC) and their derivatives (ΔMFCC). The long-term feature measure exploits the dynamics of F0 and prosody extracted from the voice signal.

3.1. Local (short-term) feature measure

It has been reported that the singing voice has additional resonance to the speaking voice at a medium frequency range known as *singer's formant* [6]. It is also known that the spectral shape of the *breathy voice* has steeper tilt than the speaking voice [14]. Therefore, we hypothesize that the spectral envelope has a discriminative cue of a singing voice that can be extracted from a short-term signal. Figure 2 shows the spectrogram of singing and speaking voices when a speaker sang and read the same phrase in the lyrics of a song. The difference of spectral envelope as well as harmonic structure can be observed. As for the measure for a spectral envelope, Mel-Frequency Cepstrum Coefficients (MFCC) and their derivatives (ΔMFCC), which are successfully used for envelope extraction in speech recognition applications, are used. Every 10 ms, the MFCC are calculated for a 100-ms hamming windowed frame[1] whereas ΔMFCC is calculated as regression parameters over five frames.

In this approach, the distributions of MFCC vectors are modeled by 16-mixture Gaussian Mixture Models (we used diagonal covariance matrices) for both singing and speaking voice signals, and discrimination is performed through the maximum likelihood principle:

$$\hat{d} = \underset{d=\text{sing,speak}}{\text{argmax}}\; f(\mathbf{x}; \Lambda_d),$$

where Λ_d, $(d = \text{sing, speak})$ are the GMM parameters for the distributions of MFCC vectors denoted by $\mathbf{x}$.

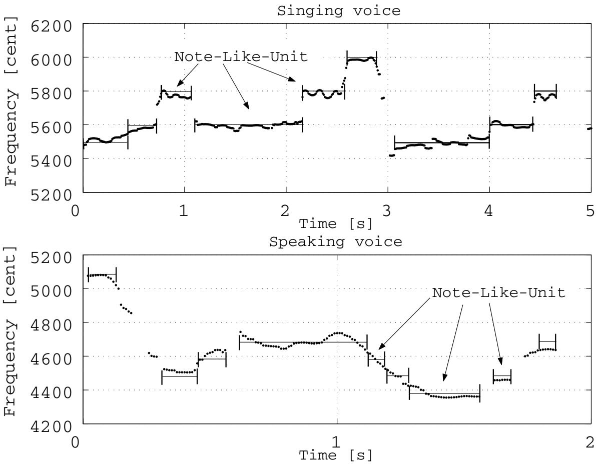

Figure 3: F0 contour and Note-Like-Unit extracted.

3.2. Long-term feature measure

Since the singing voice is generated under the constraint of melodic and rhythm patterns, the dynamics of prosody are different from the speaking voice. Therefore, dynamics of the prosody extracted from voice signals over several musical notes are expected to be cues for discriminating singing and speaking voices. To capture such features, we designed several measures based on F0 contour.

3.2.1. F0 Extraction

F0 is estimated by using a predominant-F0 estimation method called *PreFEst* [15], originally designed for estimating the melody and bass lines in polyphonic audio signals. This method estimates the relative dominance of every possible harmonic structure in the sound mixture and determines the F0 of the most predominant one. The relative dominance is obtained by treating the mixture as if it contains all possible harmonic structures with different weights and estimating their weights by Maximum *A Posteriori* Probability (MAP) estimation.

Using the *PreFEst* method, we calculated the F0 value for every 10 ms, and then a F0 trajectory was smoothed by a median filter of a 100 ms moving window. Furthermore, ΔF0 is calculated by the five point regression, as in the MFCC case.

3.2.2. Trigram of the Note-Like-Unit (NLU) sequence

To capture the long-term characteristics of the singing voice, we first define a prosodic unit called *Note-Like-Unit* (NLU), which roughly corresponds to a musical note. As depicted in Figure 3, NLU is an interval within which F0 remains a chromatic semitone, i.e., 100 cents. Conversion from the frequency in [Hz], f_{Hz}, to [cent], f_{cent}, is given by

$$f_{\text{cent}} = 1200 \log_2 \frac{f_{\text{Hz}}}{440 \times 2^{\frac{3}{12}-5}}.$$

Once the NLU sequence is obtained for the signal, we then give two different prosodic labels to each NLU with respect to the duration and relative F0 value of the unit. As for the duration label, one of the five labels (L1 to L5) is assigned to each NLU as listed in Table 1. For the relative F0 label, one of the three labels (UP, DN, and CN) is assigned to each NLU as listed in

[1] To avoid inappropriate dependence of the spectral envelope on phonetic identity, a longer frame length is required. A 100-ms frame performed best in our preliminary experiments. The sampling frequency of the signal is 16 kHz.

Table 1: Duration and Relative F0 Label

Label	Duration
L1	50-150 ms
L2	150-300 ms
L3	300-600 ms
L4	600-1200 ms
L5	1200 ms-

Label	Relative F0
UP	higher
DN	lower
CN	same

relative position
to the preceding NLU

Table 1. Finally, trigrams of each label sequence are trained for both singing and speaking voices and used for the discrimination. In the discrimination stage, the two NLU-based measures are integrated into a likelihood measure through a weighting sum of log likelihood as follows:

$$\hat{d} = \underset{d=\text{sing,speak}}{\text{argmax}} \left[\alpha \log P(\mathbf{O}_{F0}; \Theta_{d,F0}) + \right.$$
$$\left. (1-\alpha) \log P(\mathbf{O}_{D}; \Theta_{d,D})\right],$$

where α is the weight for balancing two log probabilities, $\Theta_{d,F0}$ and $\Theta_{d,D}$, $(d = \text{sing, speak})$ are the trigram model parameters for the relative F0 and duration labels of NLU sequences denoted by $\mathbf{O}_{F0}$ and $\mathbf{O}_{D}$.

3.2.3. Distribution of ΔF0

Since Japanese intonation is characterised by a falling F0 contour, we can use the distribution of Δ F0 calculated over a long-term period as a measure for discriminating the singing and speaking voices. This approach contrasts the method discussed above where we tried to capture melodic constraints in a singing voice using NLU.

In this approach, the distributions of ΔF0 values are modeled by 16-mixture GMMs for both singing and speaking voice signals, and discrimination can be performed by a maximum likelihood principle:

$$\hat{d} = \underset{d=\text{sing,speak}}{\text{argmax}} \; f(\mathbf{y}; \Omega_d),$$

where Ω_d, $(d = \text{sing, speak})$ are the GMM parameters for the distributions of ΔF0 values denoted by $\mathbf{y}$.

4. Voice Database

The method was tested on an original voice database developed at the National Institute of Advanced Industrial Science and Technology. The database includes 7500 sound samples about five to eight seconds long that consist of 3750 samples of singing voice and 3750 samples of speaking voice recorded from 75 subjects (38 male, 37 female). At an arbitrary tempo without musical accompaniment, each subject sang two excerpts from chorus and "verse A" sections of twenty-five songs (50 sound samples), and read the lyrics of those excerpts (50 sound samples), resulting in a total of 100 samples per subject. The songs were selected from the popular music database *"RWC Music Database: Popular Music"* (RWC-MDB-P-2001) [16], which is an original database available to researchers around the world.

5. Evaluation of the Proposed Method

In this section, we show experimental evaluations. First, we evaluate the discrimination performance using spectral envelope. Second, we evaluate the method using the long-term features, i.e., the NLU sequence and F0. Finally, we compare the long-term and short-term features.

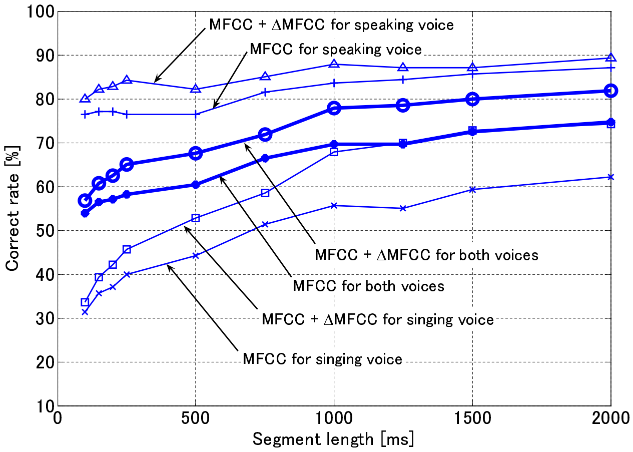

Figure 4: Discrimination accuracy using short-term features (GMMs of MFCC and ΔMFCC).

5.1. Performance of short-term features (MFCC)

In evaluating the discrimination performance using spectral envelope, 2500 sound samples of the singing and speaking voices by 25 subjects were used for training the GMMs of the MFCC and ΔMFCC, and 480 sound samples of 50 subjects were used for testing the method. The MFCC was used up to the 12th coefficients. Figure 4 shows the discrimination accuracy as a function of the input voice length. As seen from the figure, discrimination accuracy is almost monotonically improved as the length of the test set increases. Moreover, combining with ΔMFCC, the discrimination accuracy is improved at most by 10%. With test data of two seconds length, the total performance is 81.8%; however, it is 18.2% lower compared to the human listener result.

5.2. Performance of long-term features

Methods using long-term features are also evaluated using the same test samples as above. In Figure 5, performances using the trigrams of NLU label sequence and the GMM of ΔF0 are compared. For the NLU trigram, $\alpha = 0.5$ is used for integrating duration and F0 probabilities.

As shown in the figure, NLU trigram performs better than ΔF0 GMM for detecting speaking voices, and ΔF0 GMM works better for singing voices. ΔF0 GMM performed better than NLU trigram in total.

5.3. Comparing long-term and short-term features

In Figure 6, the discrimination results using MFCC+ΔMFCC and ΔF0 are plotted. It can be seen in both measures that the absolute performance improved when a longer signal was available. For input signals shorter than one second, MFCC performed better, whereas ΔF0 performed better for signals longer than one second.

Finally, two GMM measures are integrated into a likelihood measure, where the integrating weight was assumed to be 0.5. It can be seen from Figure 6 that the discrimination performance

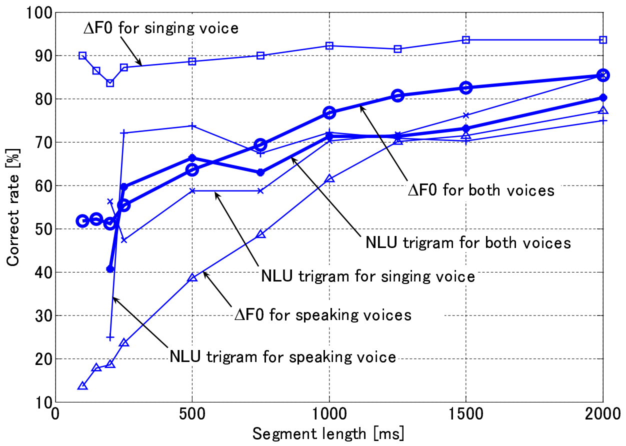

Figure 5: Discrimination accuracy using long-term features (NLU trigram and ΔF0 GMM).

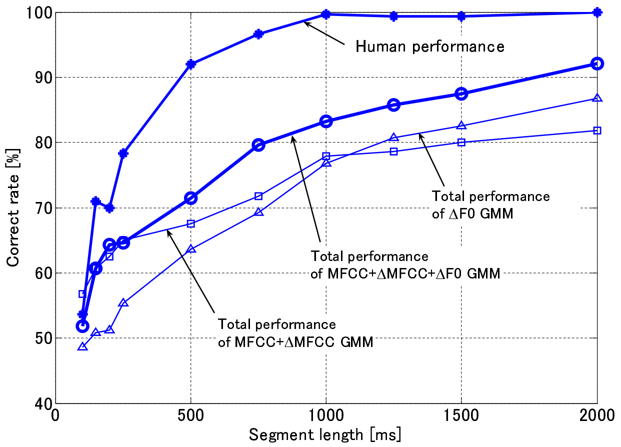

Figure 6: Comparing long-term and short-term features.

is improved by integrating two measures.

6. Discussion

The results clarified that the two measures can effectively and complementarily capture the signal features that discriminate singing and speaking voices. The discrimination using the MFCC and ΔMFCC is effective for less than one second signals. The difference between the spectrum envelopes of the singing and speaking voices is a dominant cue for the discrimination of short signals. On the other hand, the discrimination using the ΔF0 is effective for the signals of one second or longer. The GMM of ΔF0 appropriately deals with the difference of global F0 contours of singing and speaking voices by modeling local changes of the F0.

7. Summary

In this paper, we discussed the discrimination of singing and speaking voices by modeling two different aspects of voice signals in singing and speaking. Our discrimination method based on MFCC attained approximately 65% accuracy even with 300 ms signals. On the other hand, when voice signals longer than one second are available, more than 80% accuracy is obtained by the ΔF0 measure. Finally, by simply combining the two measures, more than 90% accuracy is obtained for two second signals. However, compared with human capability, discrimination performance is low, especially when the test signal is shorter than one second. In the future, we plan to optimize the integration method for the two complementary measures to further improve the total performance.

8. References

[1] T. Saitou, M. Unoki, and M. Akagi, "Extraction of F0 dynamic characteristics and development of F0 control model in singing voice," in *Proc. ICAD 2002*, July 2002, pp. 275–278.

[2] T. Saito, M. Unoki, and M. Akagi, "Development of the F0 control method for singing-voices synthesis," in *Proc. SP 2004*, Mar. 2004, pp. 491–494.

[3] C. Shih and G. Kochanski, "Prosody control for speaking and singing styles," in *Proc. EUROSPEECH2001*, Sept. 2001, pp. 669–672.

[4] H. Kawahara and H. Katayose, "Scat singing generation using a versatile speech manipulation system, STRAIGHT," *JASA*, vol. 109, pp. 2425–2426, 2001.

[5] Y. Edmund Kim, "Singing voice analysis/synthesis," Ph.D. dissertation, MIT, Sept. 2003.

[6] S. Johan, "The acoustics of the singing voice," *Scientific American*, p. 82, Mar. 1977.

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

[8] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal, "Speech/music discrimination for multimedia application," in *Proc. ICASSP 2000*, June 2000, pp. 2445–2448.

[9] J. Ajmera, I. McCowan, and H. Bourland, "Robust HMM-based speech/music segmentation," in *Proc. ICASSP 2002*, May 2002, pp. 297–300.

[10] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. ICASSP 1997*, Apr. 1997, pp. 1331–1334.

[11] G. Williams and D. Ellis, "Speech/music discrimination based on posterior probability features," in *Proc. EUROSPEECH 1999*, Sept. 1999, pp. 687–690.

[12] E. S. Parris, M. J. Carey, and H. Lloyd-Thomas, "Feature fusion for music detection," in *Proc. EUROSPEECH 1999*, Sept. 1999, pp. 2191–2194.

[13] D. B. Gerhard, "Perceptual features for a fuzzy speech-song classification," in *Proc. ICASSP 2002*, May 2002, pp. 4160–4163.

[14] D. Childers and C. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *JASA*, vol. 90, no. 5, pp. 2394–2410, 1991.

[15] M. Goto, "A real-time music scene description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311–329, 2004.

[16] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," in *Proc. ISMIR 2002*, Oct. 2002, pp. 287–288.