9th International Conference on Music Perception and Cognition

Alma Mater Studiorum University of Bologna, August 22-26 2006

On human capability and acoustic cues for discriminating singing and speaking voices

Yasunori Ohishi
Graduate School of Information Science, Nagoya University
Nagoya, Aichi, Japan
ohishi@sp.m.is.nagoya-u.ac.jp

Masataka Goto
National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, Ibaraki, Japan
m.goto@aist.go.jp

Katunobu Itou
Faculty of Computer and Information Sciences, Hosei University
Koganei, Tokyo, Japan
itou@k.hosei.ac.jp

Kazuya Takeda
Graduate School of Information Science, Nagoya University
Nagoya, Aichi, Japan
kazuya.takeda@nagoya-u.jp

ABSTRACT

In this paper, acoustic cues and human capability for discriminating singing and speaking voices are discussed to develop an automatic discrimination system for singing and speaking voices. Based on the results of preliminary subjective experiments, listeners discriminate between singing and speaking voices with 70.0% accuracy for 200-ms signals and 99.7% for one-second signals. Since even short stimuli of 200 ms can be correctly discriminated, not only temporal characteristics but also short-time spectral features can be cues for discrimination. To examine how listeners distinguish between these two voices, we conducted subjective experiments with singing and speaking voice stimuli whose voice quality and prosody were systematically distorted by using signal processing techniques. The experimental results suggest that spectral and prosodic cues complementarily contributed to perceptual judgments. Furthermore, a software system that can automatically discriminate between singing and speaking voices and such performances is also reported.

In: M. Baroni, A. R. Addessi, R. Caterina, M. Costa (2006) Proceedings of the 9th International Conference on Music Perception & Cognition (ICMPC9), Bologna/Italy, August 22-26 2006.

Keywords

Singing voice, Speaking voice, Perception, Discrimination, Voice quality, Prosody

INTRODUCTION

Sounds from the human mouth include such acoustic events as speaking, singing, laughing, coughing, whistling, and lip noises. Humans communicate by creatively using these acoustic events because they can instantaneously discriminate between such sounds by perceiving the various features that characterize them. The purpose of our research is to clarify how humans discriminate between these voices.

Among such acoustic events, this paper focuses on the discrimination between singing and speaking voices. When humans sing, the vocal style can vary from the speaking voice to some degree. Furthermore, singing voice is a vocal style to which various emotions are added based on a song's key and its lyrics; that is, vocal style represents various emotional voices in an abstract form. Therefore, revealing the characteristics that influence the perception of the singing voice creates the possibility of applications that discriminate between other vocal styles, such as irate or whispery voices.

Many research results have reported the characteristics of singing voices, whose typical characteristics include the fundamental frequency (F0, perceived as pitch) and intensity that vary widely; the spectral envelope of the singing voice has additional resonance at a medium frequency range known as the *singing formant* (Sundberg, 1974). Although the *singing formant* is observed in the voices of opera singers, it is not necessarily observed in amateurs.

However, humans can discriminate a singing from a speaking voice in daily conversation even if these voices are produced by an amateur.

Also, previous work related to the singing voice includes a control model of fundamental frequency trajectory (Saito et al., 2002; Saito et al., 2004), general characteristics (Kawahara and Katayose, 2001; Edmund Kim, 2003), acoustic differences between trained and untrained singers' voices (Omori et al., 1996; Brown et al., 2000; Watts et al., 2006), the subjective evaluation of common singing skills (Nakano et al., 2006), and singing voice morphing between expressions (Yonezawa et al., 2005). On the other hand, previous work related to the discrimination between singing and speaking voices includes a holomorphic model of the differences in glottal air flow (Rothenberg, 1981; Alku et al., 1992; Alku et al., 1996) and the dynamic characteristics of F0 trajectory (Shin et al., 2001). Therefore, most previous work has focused on either the singing or the speaking voice.

Table 1. Listening samples based on signal length in investigation of signal length necessary for discrimination.

Signal length	Singing voice	Speaking voice
100, 150, 200, 250, 500, 750, 1,000 ms	25 signals	25 signals
1,250 ms	20 signals	20 signals
1,500, 2,000 ms	10 signals	10 signals
Total	215 signals	215 signals

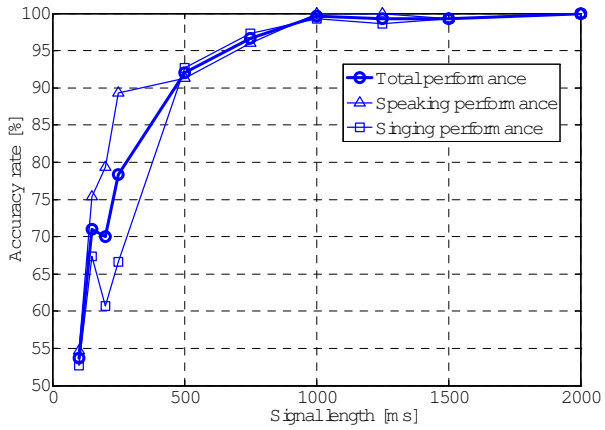

Figure 1. Human discrimination performance between singing and speaking voices as a function of signal length.

None of these works has presented knowledge based on subjective and objective evaluations of acoustic features that influence discrimination between voices. The goal of this study is to characterize the nature of singing and speaking voices based on subjective experiments and build measures that automatically discriminate between them.

HUMAN PERFORMANCE OF DISCRIMINATING SINGING AND SPEAKING VOICES

We investigated the human capability to discriminate between singing and speaking voices by conducting a subjective experiment. First, we introduce the voice database that we used. Second, we show subjective experimental conditions and results.

Voice database

We used 7,500 sound samples excerpted from an original voice database called the ``AIST Humming Database'' (Goto and Nishimura, 2005) developed at the National Institute of Advanced Industrial Science and Technology (AIST). Those samples, each about 7.0 to 12.0 seconds long, consist of 3,750 samples of singing voices and 3,750 samples of speaking voices recorded from 75 subjects (37 males, 38 females). At an arbitrary tempo without musical accompaniment, each subject sang two excerpts from the chorus and the first verses of 25 songs in different genres

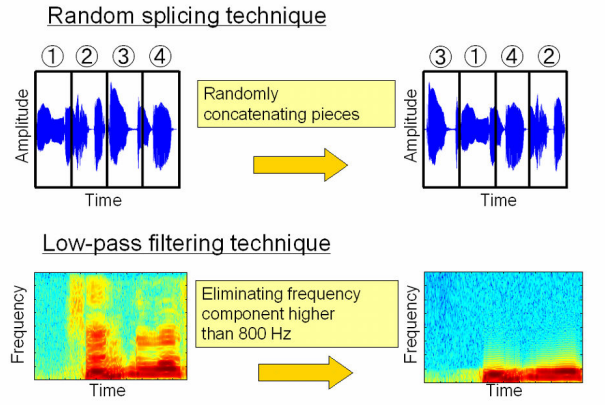

Figure 2. Random splicing and low-pass filtering techniques.

(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 a popular music database, *"RWC Music Database: Popular Music"* (RWC-MDB-P-2001) (Goto et al., 2002), which is an original database available to researchers around the world.

Investigation of signal length necessary for discrimination

We investigated the signal length necessary for human listeners to discriminate between singing and speaking voices by conducting a subjective experiment. In the experiment, we used 5,000 voice signals (2,500 singing and 2,500 speaking voices) recorded from 50 subjects (25 males, 25 females) randomly selected from the voice database, and cut them into 50,000 voice signals of 10 different lengths (from 100 to 2,000 ms). 10 subjects listened to 430 signals (215 singing and 215 speaking voices) randomly extracted from those 50,000 voice signals (Table 1) and determined

ISBN 88-7395-155-4

whether the voice signal is singing, speaking, or impossible to discriminate.

Figure 1 shows that approximately one second is enough for humans to discriminate between singing and speaking voices. Even with a 200-ms signal, discrimination accuracy is more than 70%. This suggests that not only temporal characteristics corresponding to rhythm and melody but also such short-term features as spectral envelopes carry discriminative cues between singing and speaking voices.

Investigation of acoustic cues necessary for discrimination

To compare the importance of temporal and spectral cues for discrimination, we conducted subjective experiments using two sets of stimuli whose voice quality and prosody were distorted by using signal processing techniques, as shown in Figure 2.

Table 2. Listening samples in investigation of acoustic cues necessary for discrimination.

Random Splicing technique		
Length of pieces	**Singing voice**	**Speaking voice**
125 ms	40 signals	40 signals
200 ms	40 signals	40 signals
250 ms	20 signals	20 signals
Total	100 signals	100 signals
Low-pass filtering technique		
	Singing voice	**Speaking voice**
Total	100 signals	100 signals

The first set of stimuli was generated by randomly splicing the waveform, i.e., dividing a signal into small pieces and randomly concatenating them. In the set of stimuli, the temporal structure of the signal is distorted whereas short-time spectral features are maintained (Scherer, 1985; Friend et al., 1996).

The second set of stimuli was generated by low-pass filtering, i.e., eliminating frequency component higher than 800 Hz. This set of stimuli maintains the temporal structure of the original signal although short-time spectral features are distorted (Scherer, 1985).

In the experiment, we used 5,000 voice signals (2,500 singing and 2,500 speaking voices) recorded from the 50 subjects (25 males, 25 females) used above, and obtained 15,000 voice signals (7,500 singing and 7,500 speaking voices) by random splicing, which cut one-second signals into small pieces of three types (125, 200, and 250 ms) and 5,000 voice signals (2,500 singing and 2,500 speaking voices) generated by low-pass filtering. 10 subjects listened to 200 signals (100 singing and 100 speaking voices) randomly extracted from 15,000 voice signals by random splicing and 200 signals (100 singing and 100 speaking voices) randomly extracted from 5,000 voice signals by low-pass filtering (Table 2), and determined whether the voice signal is singing or speaking.

Discrimination results by random splicing technique

In Figure 3, the discrimination results of singing and speaking voices are shown for one-second signals that were not distorted at all. They are 99.3% and 100%, respectively. However, the accuracy rate declines by random splicing. The accuracy rate of singing voices especially declines as the length of the pieces shortens from 250 to 125 ms. When the length of the pieces is 125 ms, the accuracy rate of the singing voice is 70.6%, which is 28.7% lower than the results of original voices. On the other hand, when the length of the pieces is 125 ms, the accuracy rate of speaking voices is 95.0%, which is only 5.0% lower than the results of original voices. We obtained the following comments from listeners after this experiment:

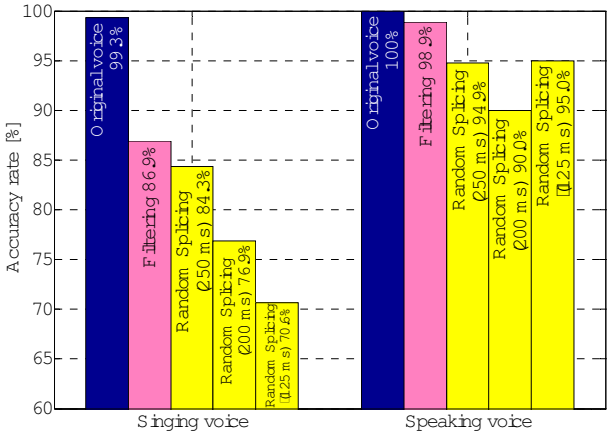

Figure 3. Accuracy rate of one-second signals by random splicing and low-pass filtering techniques.

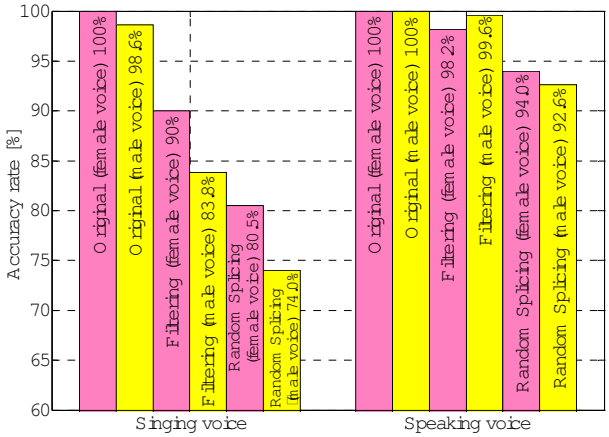

Figure 4. Accuracy rate of one-second signals by random splicing and low-pass filtering techniques as a function of vocal person's gender.

ISBN 88-7395-155-4

- When I listened to prolonged vowel production, I judged it to be a singing voice.
- I focused on the difference in voice quality between singing and speaking voices.
- When the amplitude fluctuation degree of a voice signal was great, I thought it was a singing voice.
- If the pitch varied widely, I thought it was a singing voice.
- It was easier to discriminate between singing and speaking voices in female voices than in male voices because the difference in pitch between them is wider for female voices.

Discrimination results by low-pass filtering technique

In Figure 3, the discrimination results of singing and speaking voices by low-pass filtering are 86.9% and 98.9%, respectively. As in the random splicing technique, the accuracy rate of singing voice declines more than the speaking voice.

Table 3. Analysis conditions of voice signals.

Sampling rate	16 kHz
Window	Hamming
Frame length	25 ms
Frame time	10 ms
Mel-filterbank	24

We obtained the following comments from listeners after this experiment:

- By focusing on differences in tempo, rate of speech, rhythm, and pitch fluctuation, I could discriminate between singing and speaking voices. When the amplitude fluctuation degree of a voice signal was great, I thought it was a singing voice.
- If a voice signal contained a constant location in pitch, I thought it was a singing voice.

Discrimination results for vocal person's gender

Figure 4 shows discrimination results by random splicing and low-pass filtering techniques for vocal person's gender. The accuracy rate of female singing voices by random splicing is 80.5%, which represents a mean accuracy rate of 125, 200 and 250 ms by length of the pieces. On the other hand, the accuracy rate of male singing voices is 74.0%, which is a decrease of 6.5% compared to female singing voices. The accuracy rate of male singing voices by low-pass filtering is 83.8%, a decrease of 6.2% compared to female singing voices. These results show that discrimination between male singing and speaking voices is harder than between female singing and speaking voices.

Discussion

Because the temporal structure of the original singing voices that correspond to rhythm and melody has been distorted to render them unavailable for discrimination, the accuracy rate of singing voices by random splicing technique declined. It is also considered that listeners confused singing voices with speaking voices because of the short vowels of singing voices divided by the random splicing technique. Based on the investigation of vowel length for a certain signal that confused singing with speaking voices, the vowel length of the original singing voice averaged 146.7 ms, and vowel length by random splicing averaged 73.3 ms: that is, half the vowel length of the original singing voice. On the other hand, in a signal that contained the same lyrics read by the same subject, vowel length by random splicing averaged 60.0 ms. This only slightly changed compared to the original average vowel length of 70 ms. Vowel length is clearly an important cue for discrimination. Consequently, the results clarified that a speaking voice generated by random splicing resembles a speaking voice; on the other hand, a singing voice generated by random splicing also resembles a speaking voice.

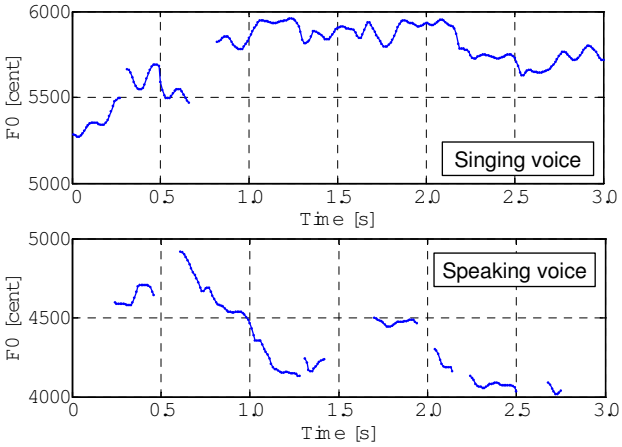

Figure 5. F0 contour of singing and speaking voices corresponding to identical lyrics.

Despite eliminating frequency component higher than 800 Hz, a speaking voice can be distinguished from a singing voice by perceiving the remaining prosody and tempo. However, a singing voice by low-pass filtering is not always easy to distinguish from a speaking voice because the distinction requires short-time spectral features. Although the cut-off frequency of the filter is 800 Hz in this experiment, by varying this value, which frequency bands are important for discrimination remains a matter of future research.

DISCRIMINATION MEASURES

From subjective experiments, human listeners distinguished between singing and speaking voices with 100% accuracy for one-second signals. On the other hand, even if

the signal length was as short-term as 200 ms, the discrimination rate was 70.0%. Moreover, it was found that not only temporal characteristics but also short-term spectral features are important for discrimination.

Therefore, to objectively clarify how these features contribute to the discrimination of the two styles, we propose an automatic vocal style discriminator that can discriminate between singing and speaking voices by using two different measures: short-term and long-term feature measures. The short-term 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 extracted from voice signals.

Short-term spectral feature measure

To measure a spectral envelope, Mel-Frequency Cepstrum Coefficients (MFCC) and their derivatives (ΔMFCC), which are successfully used for envelope extraction in speech recognition applications, were used. As shown in Table 3, every 10 ms, MFCC are calculated for 25-ms hamming windowed frames; ΔMFCC is calculated as regression parameters over five frames.

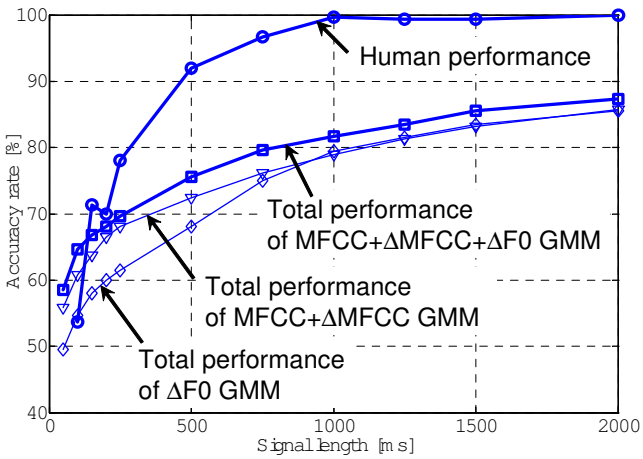

Figure 6. Comparing and integrating two measures using a spectral envelope (MFCC) and ΔF0.

Long-term feature measure

Since the singing voice is generated under the constraints of melodic and rhythm patterns, the dynamics of prosody differ from the speaking voice. Therefore, the dynamics of prosody extracted from voice signals are expected to be cues for automatically discriminating between singing and speaking voices (Figure 5).

F0 is estimated by using the predominant-F0 estimation method of Goto et al. (Goto et al., 1999) that estimates the relative dominance of every possible harmonic structure in the sound mixture and determines the F0 of the most predominant one. Relative dominance is obtained by treating the mixture as if it contains all possible harmonic structures with different weights, which are calculated by Maximum A *Posteriori* Probability (MAP) estimation.

Using the method, we determined 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 five-point regression, as in the MFCC case.

Training the discriminative model

In this approach, the distribution of MFCC vectors or ΔF0 values are represented by 16-mixture Gaussian Mixture Models (GMM) trained on the training set using the expectation maximization algorithm for both singing and speaking voice signals. The variances of distributions were modeled by a diagonal covariance matrix. Discrimination was performed through the maximum likelihood principle:

$$\hat{d} = \arg\max_{d=\text{singing, speaking}} \frac{1}{N}\sum_{n=1}^{N} \log f(\mathbf{x}_n; \Lambda_d), \quad (1)$$

where $\mathbf{x}_n$ is the nth feature vector, N is input signal length and Λ_d $(d = \text{singing, speaking})$ are the GMM parameters for the distribution of MFCC vectors. Function f calculates posterior probability by using all GMM parameters for both singing and speaking voices.

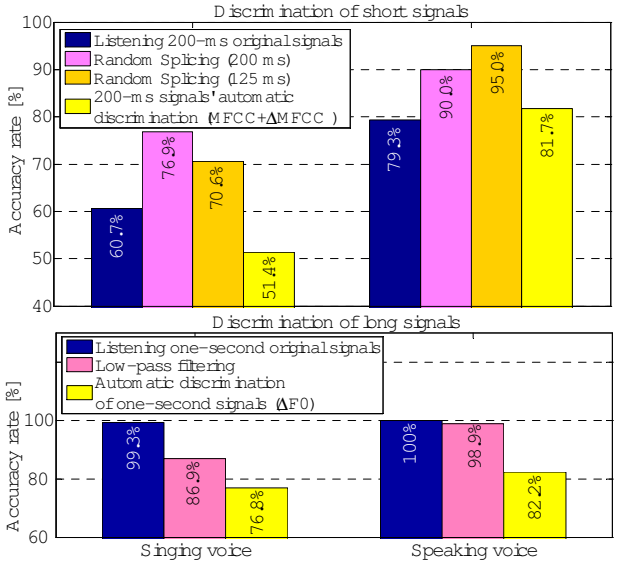

Figure 7. Comparing automatic discrimination performance with results of subjective experiments.

EVALUATION OF PROPOSED METHOD

In this section, we show experimental evaluations for automatic discrimination between singing and speaking voices. In evaluating the discrimination performance using the spectral envelope and the dynamics of F0, 7,500 sound samples of singing and speaking voices from 75 subjects were used to train the GMMs of the feature vectors and to test the method. A fifteen-fold cross-validation approach

was used for evaluation. First, sound samples from 75 subjects were divided into fifteen groups. Eight of the fifteen groups were used for GMM training, and the rest were used as a test. An average discrimination rate was obtained from the fifteen cross-validation tests.

In Figure 6, discrimination results using MFCC+ΔMFCC and ΔF0 are plotted. MFCC was used up to the 12th coefficients. In both measures 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 measures were integrated into a 25-dimensional vector. It can be seen from Figure 6 that discrimination performance is improved by 2.6% for two-second signals.

DISCUSSION

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

Furthermore, we compared automatic discrimination performances with the results of subjective experiments. When the temporal structure of the signal is distorted by a random splicing technique, human capability for discriminating between singing and speaking voices decreased because the vowel length is shorter than the original singing voice signals. However, when the length of the pieces in the random splicing technique is 125 ms, human capability is 70.6%. Based on this result, the short-term spectral features of signals affect discrimination. When comparing the automatic discrimination results of 200-ms singing voices using MFCC and ΔMFCC with the human capability of 200-ms singing voices, although the automatic discrimination result decreased by 9.3% compared with human capability, this human capability is more similar to the automatic discrimination result using MFCC and ΔMFCC than any other automatic discrimination results in the above chart of Figure 7. Consequently, this result shows the importance of spectral features for automatically discriminating between singing and speaking voices. MFCC is successfully used to represent the phoneme structure in speech recognition applications; however, to discriminate between those voices, we need to focus on the features which can not be represented by MFCC. In the future, we plan to propose new measures to improve the automatic discrimination performance.

Even though short-term spectral features are distorted by eliminating frequency component higher than 800 Hz, humans can distinguish between those voices by perceiving such temporal features of signals as melody and rhythm patterns. In other words, the temporal features included in long-term signals are important for discriminating between those voices. When comparing the automatic discrimination results using ΔF0 with the subjective experimental results, the discrimination results are low, as shown in Figure 7. In this paper, ΔF0 is calculated as regression parameters over five frames (50 ms) of F0 that are estimated continuously. However, from the subjective experimental results, humans distinguish between those voices by perceiving continuous changes of F0 longer than 50 ms. Therefore, a longer ΔF0 calculation method that considers the F0 interpolation of unvoiced sounds is needed to further improve the performance.

CONCLUSION

In this paper, we discussed acoustic cues and human capability for discriminating singing and speaking voices. When investigating the signal length necessary for singing and speaking voice discrimination, we showed that humans can discriminate singing and speaking voices 200-ms long and one-second long with 70.0% and 99.7% accuracy, respectively. By conducting subjective experiments with voice signals whose voice quality and prosody were systematically distorted by signal processing techniques, we showed that spectral and prosodic cues complementarily contributed to perceptual judgments.

Furthermore, by hypothesizing that listeners depend on different cues based on the length of signals, we proposed an automatic vocal style discriminator that can distinguish between singing and speaking voices by using two measures: spectral envelope (MFCC) and F0 derivative. In our experimental results, when voice signals longer than one second are discriminated, the F0-based measure outperforms the MFCC-based measure. On the other hand, when voice signals shorter than one second are discriminated, the MFCC-based measure outperforms the F0-based measure. While discrimination accuracy with the F0-based measure is 85.0% for two-second signals, a simple combination of two measures improves it by 2.3% 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 clarify the differences of spectral features between singing and speaking voices and to discuss a longer F0 contour modeling method.

REFERENCES

Sundberg, J. (1974). Articulatory interpretation of the 'singing formant'. *J. Acoust. Soc. Amer.*, Vol.55, pp. 838-844.

Saito, T., Unoki, M. and Akagi, M. (2004). Development of the F0 control method for singing-voices synthesis. *Proc. SP 2004*, pp. 491–494.

Saito, T., Unoki, M. and Akagi, M. (2002). Extraction of F0 dynamic characteristics and development of F0 control model in singing voice. *Proc. ICAD 2002*, pp. 275–278.

Kawahara, H. and Katayose, H. (2001). Scat singing generation using a versatile speech manipulation system, STRAIGHT. *J. Acoust. Soc. Amer*., Vol. 109, pp. 2425–2426.

Edmund Kim, Y. (2003). Singing voice analysis/Synthesis. PhD Thesis, MIT.

Omori, K., Kacker, A., Carroll, L., Riley, W. and Blaugrund, S. (1996). Singing Power Ratio: Quantitative Evaluation of Singing Voice Quality. *Journal of Voice*, Vol. 10, No. 3, pp. 228–235.

Brown, W. S. J., Rothman, H. B. and Sapienza, C. (2000). Perceptual and Acoustic Study of Professionally Trained Versus Untrained Voices. *Journal of Voice*, Vol. 14, No. 3, pp. 301–309.

Watts, C., Barnes-Burroughs, K., Estis, J. and Blanton, D. (2006). The Singing Power Ratio as an Objective Measure of Singing Voice Quality in Untrained Talented and Nontalented Singers. *Journal of Voice*, Vol. 20, No. 1, pp. 82–88.

Nakano, T., Goto, M. and Hiraga, Y. (2006). Subjective Evaluation of Common Singing Skills Using the Rank Ordering Method. *Proc. ICMPC2006*. (accepted).

Yonezawa, T., Suzuki, N., Mase, K. and Kogure, K. (2005). Gradually Changing Expression of Singing Voice based on Morphing. *Proc. Eurospeech 2005*, pp. 541–544.

Rothenberg, M. (1981). The Voice Source in Singing. *Research Aspects of Singing, Pub.*, No. 33, pp. 15–31.

Alku, P. (1992). Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering. *Speech Communication*, No. 11, pp. 109–118.

Alku, P. and Vilkman, E. (1996). Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering. *Speech Communication*, No. 18, pp. 131–138.

Shih, C. and Kochanski, G. (2001). Prosody control for speaking and singing styles. *Proc. Eurospeech 2001*, pp. 669–672.

Goto, M. and Nishimura, T. (2005). AIST Humming Database:Music Database for Singing Research. *The Special Interest Group Notes of IPSJ (MUS)*, Vol. 2005, No. 82, pp. 7–12. (in Japanese).

Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R. (2002). RWC Music Database: Popular, Classical, and Jazz Music Databases. *Proc. ISMIR 2002*, pp. 287–288.

Scherer, K. R. (1985). Vocal cues to deception: A comparative channel approach. *Journal of Psycholinguistic Research*, Vol. 14, No. 4, pp. 409–425.

Friend, M. and Farrar, M. J. (1996). A comparison of contentmasking procedures for obtaining judgments of discrete affective states. *J. Acoust. Soc. Amer.*, Vol. 96, No. 3, pp. 1283–1290.

Goto, M., Itou, K. and Hayamizu, S. (1999). A Real-time Filled Pause Detection System for Spontaneous Speech Recognition. *Proc. Eurospeech 1999*, pp. 227–230.

ISBN 88-7395-155-4