14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

SPEECH STYLE CONVERSION BASED ON THE STATISTICS OF VOWEL SPECTROGRAMS AND NONLINEAR FREQUENCY MAPPING

Toru Takahashi, Hideki Banno, Toshio Irino, and Hideki Kawahara

Faculty of Systems Engineering, Wakayama University
930 Sakaedani Wakayama, Japan
phone: +(81)73-457-8530, fax: +(81)73-457-8109,
email: {tall,irino,kawahara}@sys.wakayama-u.ac.jp
web: www.sys.wakayama-u.ac.jp/˜tall/indexe

Department of Information Engineering, Meijo University
1-501, Shiogamaguchi, Tempaku-ku, Nagoya, Japan
phone: +(81)52-838-2088,
email: banno@ccmfs.meijo-u.ac.jp
web: www-is.meijo-u.ac.jp/teacher/banno

ABSTRACT

A simple, efficient, and high-quality speech style conversion algorithm is proposed based on STRAIGHT. A very high-quality VOCODER STRAIGHT consists of instantaneous-frequency based F0 and source information extraction part and F0-adaptive time-frequency smoothing part to eliminate preriodicity interferences. The proposed method uses only vowel information to design the desired conversion functions and parameters. So, it is possible to reduce the amount of training data required for conversion. The processing of the proposed method is : 1) to produce abstract spectra that is represented on the perceptual frequency axis and is derived as average spectrum for each vowel and each style; 2) to decompose the original spectrum into the abstract spectrum and the residual, fine structure; 3) to replace the abstract spectrum from the original to the target style; 4) to map the fine structure with nonlinear frequency warping for adapting the target style fine structure; 5) then to add them together to produce target speech. An efficient algorithm for this conversion was developed using an orthogonal transformation referred to as warped-DCT. An informal listening test indicated that the proposed method yields more natural and high-quality speech style conversion than the previous methods.

1. INTRODUCTION

Speech style plays important roles in speech communications by conveying emotions and other non and para-linguistic information. Therefore, it is crucial to provide means to control the speech styles of synthetic speech both in scientific and practical applications [1, 2]. However, so far few systems have successfully provided high-quality speech style conversion, perhaps because there still remains a huge gap in the speech quality between natural and synthetic speeches, as demonstrated in the Blizzard Challenge 2005 [3, 4].

Speech morphing [5, 6] based on STRAIGHT [7] is one prospective alternative to provide an answer to control speech style. It was reported that the naturalness of morphed speech sounds synthesized by a method equivalent to the original natural speech samples. It is also worth notice that STRAIGHT provides fine spectral representation for the speech synthesis system won in the Blizzard Challenge.

The systems are based on linear, frame-wise spectra which are not necessarily representative in terms of human speech production and perception. For example, there would be supra-segmental information for each speech style.

The proposed method is designed to extract the differences in supra-segmental aspects due to speech style differences from their instantiation as deviations in syllabic information. It is also intended to preserve other aspects to prevent possible degradations caused by averaging and/or other manipulations. To attain these design objectives, a set of assumptions and approximations was introduced.

The first assumption is that the degree of degradations can be measured as spectral distance on the perceptual frequency axis like ERB_N [8]. For the sake of computational efficiency and theoretical simplicity, an orthonormal basis, namely warped-DCT [9], was used to implement the measure. The second assumption is that the differences between the speech styles are represented as the deviations in the abstract spectra of individual vowels, and that the deviations remain stable within a provided sentence spoken in one speech style. The abstract spectrum for each style and each vowel is calculated as the average spectrum of vowel segments in the speech database. The third assumption is that the fine structures susceptible to degradation are approximated by the residual spectra. With these assumptions, the style conversion is established as a procedure involving the replacement of abstract spectra, frequency mapping of residual spectra, and adding them together. In the following sections, we describe the detail of the procedure.

2. SPEECH STYLE CONVERSION

We use STRAIGHT in the analysis and synthesis parts of the proposed system for speech style conversion. The physical parameters derived in the analysis are fundamental frequency, F0, smoothed spectrum, and aperiodicity. Fig. 1 shows the spectral conversion part which consists of two main stages. In stage 1, abstract spectra for style *A* are selected in accordance with the phonetic labels and then replaced with abstract spectra for style *B*. In stage 2, the fine structures are derived as the difference between the input

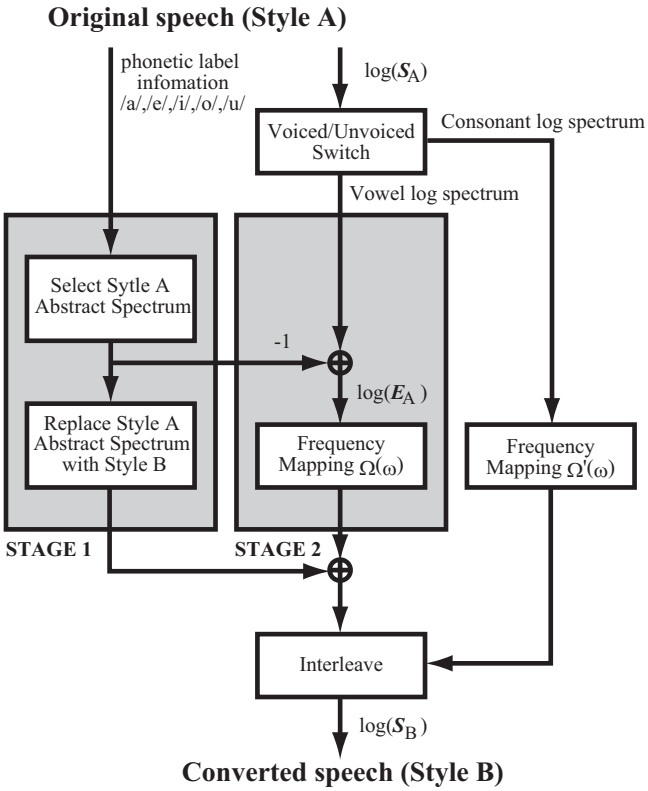

Figure 1: Overview of two-stage speech style conversion

spectra and the abstract spectra and then nonlinear frequency mapping are applied to compensate the positions of spectral peaks. This procedure is also applied to the aperiodicity parameters. F0 conversion is performed in the same manner proposed in the previous method[10]. In the following subsections, spectral conversion is described in detail.

2.1 Perceptual orthogonal transfrom for spectral representation

In the previous system[10], spectral parameters are represented as the logarithmic magnitude of the STRAIGHT spectrum that has equal weights on a linear frequency axis. For producing perceptually natural sounds, it would be desirable to introduce a knowledge of the human auditory system. It is well known that the frequency axis of the human peripheral analysis is represented as the ERB_N scale[8] which is similar to mel scale commonly used in the speech processing systems.

In this paper, we used the warped-frequency version of the Discrete Cosine Transformed coefficients (warped-DCT) [9] to represents the spectral information. The warped-DCT is useful since it simultaneously performs the frequency warping from linear to ERB_N scale and the orthogonal transform for smoothing the representation on ERB_N scale.

The m-th warped-DCT coefficient $c(m)$ of the STRAIGHT spectrum $S(e^{-j\omega})$ is defined as:

$$c(m) = \frac{1}{\sqrt{\pi}} \int_{-\pi}^{\pi} Re[\Psi_m(e^{-j\tilde{\omega}})] \cdot S(e^{-j\omega}) d\omega, \quad (1)$$

$$\Psi_m(z) = \begin{cases} \frac{\sqrt{1-\alpha^2}z^{-1}}{1-\alpha z^{-1}} \left(\frac{z^{-1}-\alpha}{1-\alpha z^{-1}} \right)^{m-1}, & (m>0), \\ \frac{1}{\sqrt{2}}, & (m=0). \end{cases} \quad (2)$$

The real part of the frequency response of $\Psi_m(e^{-j\tilde{\omega}})$ is a normalized orthogonal function, when $\{\omega | 0 \le \omega \le \pi\}$. A coefficient that determines the degree of frequency warping is α. When α is zero, $Re[\Psi_m(e^{-j\tilde{\omega}})] = \cos(m\omega)$. When α is between 0 and 1, $\Psi_m(e^{-j\tilde{\omega}})$ corresponds to a cosine component defined on warped frequency $\tilde{\omega}$,

$$\tilde{\omega} = \omega + 2\tan^{-1}\{\frac{\alpha \sin\omega}{1-\alpha\cos(\omega)}\}. \quad (3)$$

At the 44.1 kHz sampling frequency, the warped frequency $\tilde{\omega}$, is close to the ERB_N scale when $\alpha = 0.68$.

2.2 Style conversion by replacing and mapping

The spectral shape of speech changes in accordance with the variation of speech style. The change is particularly remarkable in vowel segments which are also perceptually dominant. So, we have developed a style conversion method based on the manipulation of the vowel segments[10].

In stage 1 of Fig. 1, the abstract spectrum of style A is replaced with that of style B for each vowel segment. $S(t,\omega)$ is a STRAIGHT spectrum at time t segment and at frequency ω. Subscripts A and B denote original and target styles, respectively. The abstract spectrum is defined as a STRAIGHT spectrum averaged over the vowel segments in the speech database and is represented as smoothed spectrum derived from the warped-DCT coefficients truncated by M-th order. As Japanese has 5 vowels (/a/, /e/, /i/, /o/, and /u/) and there are two styles (A and B), it is necessary to prepare 10 abstract spectra. The abstract spectra of styles A and B for vowel /ξ/ are denoted by $S_A^{/\xi/}$ and $S_B^{/\xi/}$, respectively. $\xi \in \{\xi | a,e,i,o,u\}$. For each vowel, /$\xi$/, the abstract spectra are derived from the vowel spectra in the training database as follows:

$$S_A^{/\xi/}(t,\omega) = \frac{1}{|T|} \sum_{t \in T} S_A(t,\omega), \quad (4)$$

$$S_B^{/\xi/}(t,\omega) = \frac{1}{|T|} \sum_{t \in T} S_B(t,\omega), \quad (5)$$

where $T = \{t | Label(t) = /\xi/\}$, in which the function $Label(t)$ gives the phoneme corresponding to time t, and notation $|\cdot|$ represents a number of vowel segments in the database. The block processing similar to the codebook mapping are applied for replacing the abstract spectrum.

The resultant abstract spectra for vowel /a/ for neutral and angry styles are shown in Fig. 2. Vertical and horizontal axes show normalized level and frequency, respectively. The difference between neutral and angry style appears as level and the peak frequency differences.

In stage 2 of Fig. 1, the STRAIGHT spectrum $S(t,\omega)$ for vowel /ξ/ segment can be decomposed into the abstract spectrum $S^{/\xi/}(t,\omega)$ and residual spectrum, or fine structure, $E^{/\xi/}(t,\omega)$. The relationships between these variables are:

$$S_A(t,\omega) = S_A^{/\xi/}(t,\omega) \cdot E_A^{/\xi/}(t,\omega), \quad (6)$$

$$S_B(t,\omega) = S_B^{/\xi/}(t,\omega) \cdot E_B^{/\xi/}(t,\omega). \quad (7)$$

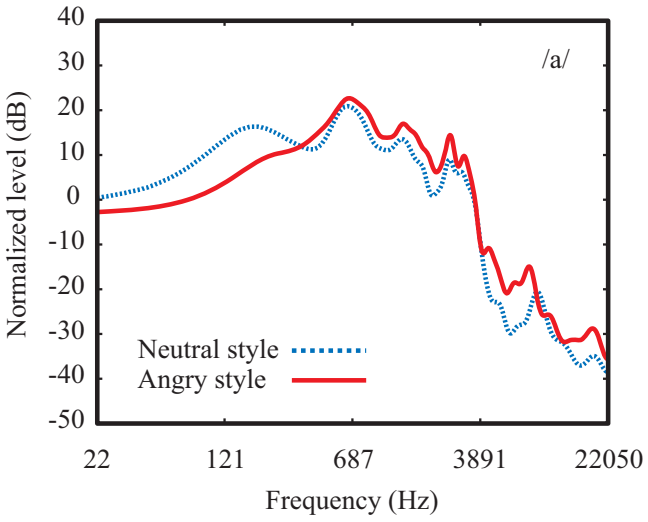

Figure 2: Abstract spectra for vowel /a/

Since the peak frequencies are difference between the styles, they are conpensated with nonlinear frequency mapping. The mapping function for each vowel $/\xi/$, $\Omega^{/\xi/}(\omega)$, is carefully designed to minimize the spectral distance between the spectrum for style *A* after the mapping and the spectrum for style *B*. Fig. 3 shows the frequency mappinig function between neutral and angry styles. A bilinear model is used as frequency mapping function $\Omega^{/\xi/}(\omega)$

$$d\Big(\frac{1}{|T_1|}\sum_{t\in T_1} S_A^{/\xi/}(t,\Omega^{/\xi/}(\omega)), \frac{1}{|T_2|}\sum_{t\in T_2} S_B^{/\xi/}(t,\omega)\Big), \quad (8)$$

where $T_1 = \{t|Label(t) = /\xi/\text{in set1}\}$ and $T_2 = \{t|Label(t) = /\xi/\text{in set2}\}$, in which the function $Label(t)$ gives phonemes corresponding to time t. The notation $|\cdot|$ represents the number of elements in a set. Fig. 3 shows the frequency warping function $\Omega^{/a/}(\omega)$ between neutral and angry styles. The degree of goodness in stage 2 of Fig. 1 is mainly determined by the mapping functions.

It is assumed that $E_A^{/\xi/}(t,\Omega(\omega))$ is sufficiently close to $E_B^{/\xi/}(t,\omega)$ in the proposed speech style conversion. Converted spectrum $S_B(t,\omega)$ is defined as follows:

If $S_A(t,\omega)$ is in vowel ξ segments,

$$\begin{aligned} S_B(t,\omega) &= S_B^{/\xi/}(t,\omega)\cdot E_A^{/\xi/}(t,\Omega(\omega))), \\ &= S_B^{/\xi/}(t,\omega)\cdot S_A(t,\Omega(\omega))/S_A^{/\xi/}(t,\Omega(\omega)). \end{aligned} \quad (9)$$

If $S_A(t,\omega)$ is not in vowel ξ segments,

$$S_B(t,\omega) = S_A(t,\Omega'(\omega)). \quad (10)$$

The frequency mapping in the vowel segments can be observed as formant position changes along with time. On the other hands, it is hard to observe frequency mapping in the segments except for vowel segments. However, it is considered that a frequency mapping in the segments except for vowel segments is caused by a frequency mapping

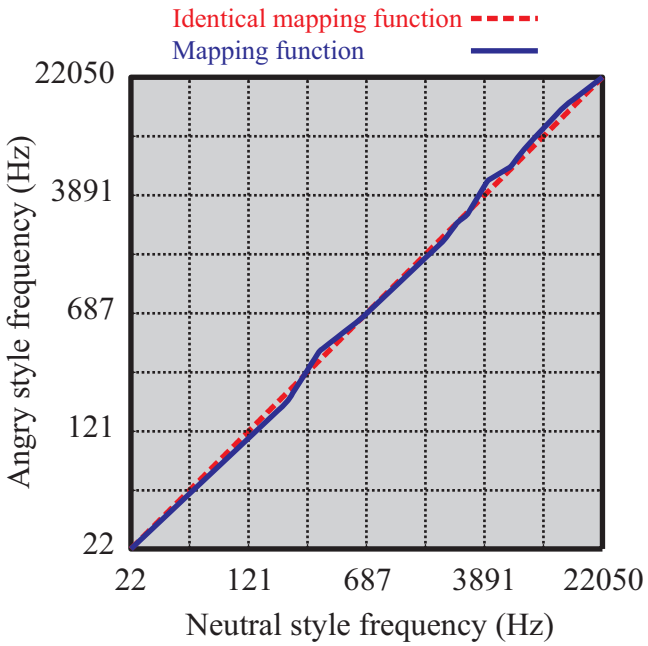

Figure 3: Mapping function with nonlinear frequency mapping for vowel /a/

in the vowel segments. Therefore frequency mapping function $\Omega'(\omega)$ is given by linear interpolation between the frequency mapping functions $\Omega^{/\xi/}(\omega)$ of the previous and the next nearest neighbor vowel segments. The method of interpolation between warping functions is described in [6].

2.3 Spectral variation model related to intensity

Intensity parameters randomly fluctuate and systematically change simultaneously while singing [11]. Systematic spectral variation due to changes in intensity was found. The intensity of various speech styles, such as emotional speech, resembles a singing voice. For modeling spectral variation related to intensity, Eqs. (6) and (7) are extended as follows:

$$S_A(t,\omega) = S_A^{/\xi/}(t,\omega)\cdot F_A^{/\xi/}(t,\omega)\cdot G_A^{/\xi/}(t,\omega), \quad (11)$$

$$S_B(t,\omega) = S_B^{/\xi/}(t,\omega)\cdot F_B^{/\xi/}(t,\omega)\cdot G_B^{/\xi/}(t,\omega), \quad (12)$$

where $F_A^{/\xi/}(t,\omega)$ and $F_B^{/\xi/}(t,\omega)$ are spectral variations, respectively. $G_A^{/\xi/}(t,\omega)$ and $G_B^{/\xi/}(t,\omega)$ are residual spectra removed from the spectral variations related to intensity.

It is assumed that $G_A^{/\xi/}(t,\Omega(\omega))$ is similar enough to be approximated to $G_B^{/\xi/}(t,\omega)$ in the proposed speech style conversion model. Converted spectrum $S'_B(t,\omega)$ is defined as follows:

If $S_A(t,\omega)$ is in vowel ξ segments,

$$S'_B(t,\omega) = S_B^{/\xi/}(t,\omega)\cdot F_B^{/\xi/}(t,\omega)\cdot G_A^{/\xi/}(t,\Omega(\omega)). \quad (13)$$

If $S_A(t,\omega)$ is not in vowel ξ segments,

$$S'_B(t,\omega) = S_A(t,\Omega'(\omega)). \quad (14)$$

For removing spectral variation related to intensity from the residual spectrum, residual spectrum parameters are decomposed into an eigen matrix and a principal component score matrix by principal component analysis (PCA). It is possible to completely reconstruct the residual spectrum parameters from the eigen matrix and principal component scores. As the first principal component score relates to intensity, it is considered that a product of the first principal component score $P(t|\xi)$ and the inverse matrix of eigen matrix $inv(M^{/\xi/})$ represents spectral variation. It is assumed that $P_A^{/\xi/}(t,\omega)$ is similar enough to be approximated to $P_B^{/\xi/}(t,\omega)$. Eq. (13) can be rewritten as

$$S'_B(t,\omega) \simeq S_B^{/\xi/}(t,\omega) \cdot P_A^{/\xi/}(t,\omega) \cdot inv(M_B^{/\xi/}) \cdot G_A^{/\xi/}(t,\Omega(\omega)). \quad (15)$$

3. EXPERIMENTS

Experiments were conducted that evaluated the effectiveness of spectrum mapping, replacing the abstract spectra, and the spectrum variation model. Three spectral conversion methods were compared. The baseline system conversion method uses mapping model. The conversion method showed in previous work uses mapping model and replacing model. The proposed conversion method in this article uses mapping model, replacing model and variation model. Experimental settings of three methods are shown in Table 1. Although target style speech sounds are not required in speech style conversion, they are prepared for evaluation by actually requiring the actor to pronounce the desired target style speeches. The sounds have the same phonetic labels as those of 'neutral' speech sounds. The spectral distortion between converted and target speech corresponding to the vowel parts is shown in Fig. 4. Vertical and horizontal axes show spectral distortion and Japanese vowels (/a/,/e/,/i/,/o/, and /u/), respectively. White, gray, black bars show the proposed, and conventional, and baseline methods, respectively. The distortions of proposed method are the lowest for each vowel. However, the distortion of the proposed method is almost same as the conventional one, Fig. 5 shows that the spectrum converted by the proposed method is different from the spectrum converted by the previous one. Vertical and horizontal axes show spectral level and frequency. The results are considered to reflect averaging for overall spectra.

4. CONCLUSION

A simple and efficient speech style conversion procedure was introduced based on STRAIGHT. The proposed method only uses averaged vowel information for designing entire transformation functions. This feature enables a significant reduction of the size of the training data for speech style conversions. Preliminary test results encouraged formal listening tests in the near future. The informal tests indicated that the resynthesized sounds preserve the natural quality of the original speech and provide impressions of intended speech styles.

Table 1: Experimental settings

System Type	Baseline	Previous	Proposed
Mapping model	on	on	on
Replacing model	off	on	on
Variation model	off	off	on

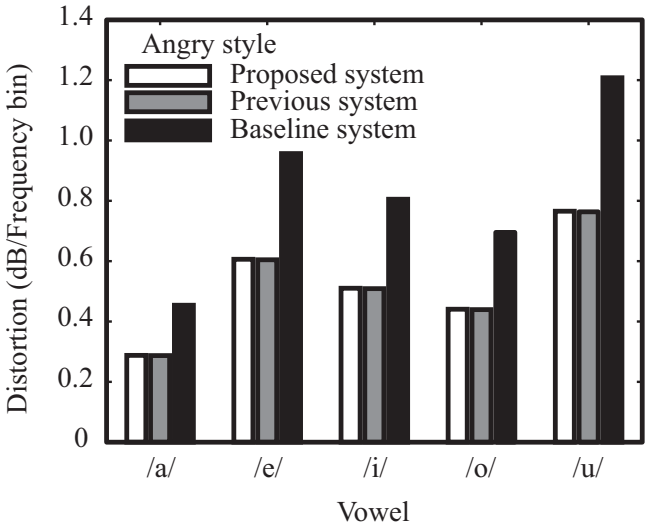

Figure 4: Speech style conversion results

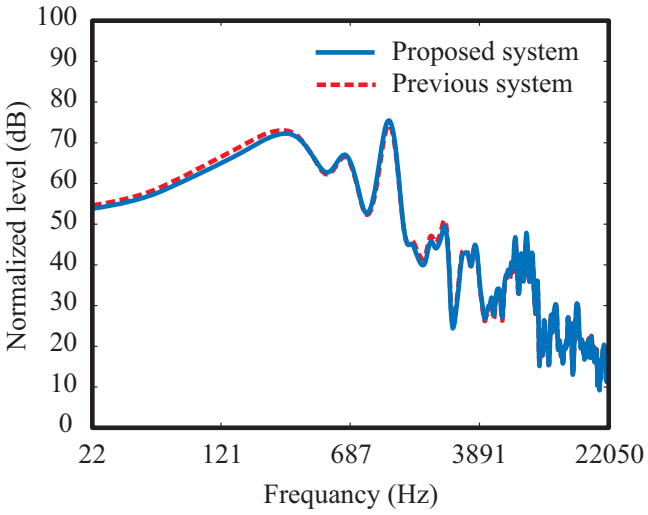

Figure 5: Converted spectrum examples

5. ACKNOWLEDGMENT

This work was partly supported by the Ministry Education, Culture, Sports, Science and Technology e-Society leading project.

REFERENCES

[1] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion ", *Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, February 1993.

[2] M. Schröeder, "Emotional speech synthesis: A review ", in *Proc. Eurospeech 2001*, vol. 1, pp. 561–564, vol. 1, pp. 561–564, Scandinavia, Denmark, September 3–7, 2001.

[3] A. W. Black and K. Tokuda, "The blizzard challenge 2005: Evaluating corpus-based speech synthesis on common databases ", in *Proc. Interspeech 2005*, pp. 77–80, Lisbon, Portugal, September 4–8, 2005.

[4] H. Zen and T. Toda, "An overview of nitech hmm-based speech synthesis system for bilizzard challenge 2005 ", in *Proc. Interspeech 2005*, pp. 93–96, Lisbon, Portugal, September 4–8, 2005.

[5] H. Matsui and H. Kawahara, "Investigation of emotionally morphed speech perception and its structure using a high quality speech minimization system ", in *Proc. Eurospeech 2003*, pp. 2113–2116, Geneva, Switzerland, September 1–4, 2003.

[6] H. Kawahara and H. Matsui, "Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation ", in *Proc. ICASSP 2003*, vol. I, pp. 256–259, Hong Kong, April 6–10, 2003.

[7] H. Kawahara, I. M. Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction ", *Speech Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.

[8] B. C. J. Moore and B. R. Glasberg, "Suggested formula for calculating auditory-filter bandwidths and excitation patterns ", *Journal of the Acoustical Society of America*, vol. 74, no. 3, pp. 750–753, September 1983.

[9] T. Irino, R. D. Patterson, and H. Kawahara, "Auditory VOCODER: Speech resynthesis from an auditory Mellin representation ", in *Proc. ICASSP 2002*, vol. II, pp. 1921–1924, Orlando, Florida, USA, May 13–17, 2002.

[10] T. Takahashi, T. Fujii, M. Nishi, H. Banno, T. Irino, and H. Kawahara, "Voice and emotional expression transformation based on statistics of vowel parameters in an emotional speech database ", in *Proc. Interspeech 2005*, pp. 537–540, Lisbon, Portugal, September 4–8, 2005.

[11] K. Tahara, T. Takahashi, M. Morise, H. Banno, and H. Kawahara, "Principal component analysis of spectral variations due to vocal effort in singing voice; applications to singing synthesis ", *Technical Report of the Institute of Electronics, Information and Communication Engineers*, vol. 105, no. 198, pp. 19–24, July, 2005 (in Japanese).