World Academy of Science, Engineering and Technology
International Journal of Electrical and Computer Engineering
Vol:1, No:11, 2007

High-Individuality Voice Conversion Based on Concatenative Speech Synthesis

Kei Fujii, Jun Okawa, and Kaori Suigetsu

***Abstract*—**Concatenative speech synthesis is a method that can make speech sound which has naturalness and high-individuality of a speaker by introducing a large speech corpus. Based on this method, in this paper, we propose a voice conversion method whose conversion speech has high-individuality and naturalness. The authors also have two subjective evaluation experiments for evaluating individuality and sound quality of conversion speech. From the results, following three facts have be confirmed: (a) the proposal method can convert the individuality of speakers well, (b) employing the framework of unit selection (especially join cost) of concatenative speech synthesis into conventional voice conversion improves the sound quality of conversion speech, and (c) the proposal method is robust against the difference of genders between a source speaker and a target speaker.

***Keywords*—**concatenative speech synthesis, join cost, speaker individuality, unit selection, voice conversion

I. Introduction

VOICE conversion is a technique that converts a source speaker's voice into another voice as if another speaker had uttered it [1]–[5]. The framework of this technique is generally divided into two stages: training stage and conversion stage. In the training stage which is done as an offline processing, system decides a conversion rule by using source speaker's voice resources and target speaker's voice resources that have collected beforehand. Using this conversion rule, in the conversion stage which is done as an online processing, an input source speech is transformed into the output target speech. The quality of conversion speech is mainly measured in terms of individuality and sound quality. Moreover, genders of speakers are also considered because the quality degradation of conversion speech is generally proportional to the degree of the difference between a source and a target speaker's voice characteristics.

Text-to-speech (TTS) is a technique in which system reads any input texts by generating the appropriate synthetic speech automatically. In recent years, corpus-based time-domain approach has become widely used for realizing high-quality speech synthesis [6]. Concatenative speech synthesis, which is the one of corpus-based TTS, is a method that can make synthetic speech which have the high-individuality and naturalness of speaker [7]–[9]. To realize these features, in this method, the synthetic speech is made by joining a speaker's natural short term waveform segments which have accumulated in a large speech corpus beforehand. In other words, the synthetic speech, generated by this method, is recycling of a speaker's natural voices that preserve naturalness and individuality. The most preferable waveform segments are searched from the corpus. The search is based on minimization of following two kinds of distortions: the distortion between candidates and target criteria, and the distortion caused by discontinuity of waveform boundaries which are stuck mutually. Although this methodology is simple compared with other conventional TTS methods such as LPC synthesis, larger amount of computation and wider memory space are needed. However, the advancement in recent years on computer resource is overcoming this computer specification problem.

In general, the degree of quality of synthetic speech is proportional to the corpus size. This is clear because a larger corpus can have more appropriate waveform candidates and more precise parameters. By researches in recent years, it was confirmed that the high-quality synthetic speech can be realized by using concatenative speech synthesis with extra large speech corpus [9].

The data size in voice conversion system, on the other hand, is generally smaller than the corpus of concatenative speech synthesis. Thus, it is thought that it becomes difficult to make a conversion rule accurately enough in case of less data condition. In addition, it is also difficult to achieve high-quality enough conversion speech, because speech has degraded by through "decomposition and re-synthesis" process that is employed by conventional methods.

From the above-mentioned, the authors have thought there is a possibility that the individuality and quality of conversion speech are improved by introducing the essence of concatenative speech synthesis into conventional voice conversion. This also means that the disadvantages of concatenative speech synthesis such as the necessity of abundant computer resources are introduced. That is, these disadvantages restrict the application area of proposal method. However, the proposal method can work well enough under

Manuscript received October 15, 2007.

Kei Fujii is with the Department of Information and Computer Sciences, Kumamoto National College of Technology, 2659-2, Suya, Kohshi-city, Kumamoto, 861-1102 JAPAN (e-mail: fujii @ cs.knct.ac.jp).

Jun Okawa was with the Department of Information and Computer Sciences, Kumamoto National College of Technology. He is now with the Japan Software Engineering Co. Ltd., 4-4-20, Nihonbashi-motoishi-cho, Chuo-ku, Tokyo, 103-0021 JAPAN (e-mail: jokawa17 @ pr.cs.knct.ac.jp).

Kaori Suigetsu was with the Department of Information and Computer Sciences, Kumamoto National College of Technology. She is now with the Advanced Course of Control and Information Systems Engineering, Kumamoto National College of Technology, 2659-2, Suya, Kohshi-city, Kumamoto, 861-1102 JAPAN (e-mail: suige @ pr.cs.knct.ac.jp).

specific conditions. Furthermore, the voices of specific person are very valuable for specific persons; e.g., the voice of famous actor for his fan, the voice of mother for her children, and the own voices for the person who lost own voice by getting serious illness. Therefore, the authors think that there are more merits in these cases.

In this paper, section II and section III describe the overviews of conventional voice conversion and concatenative speech synthesis. In section IV, the proposal voice conversion based on unit selection of concatenative speech synthesis is presented. The performance of the proposal method with about 40-minutes reading speech corpus is evaluated and discussed in section V. Finally, we conclude in section VI.

II. Overview of Voice Conversion

Voice conversion is defined as a technique that converts a speaker's voice into another voice as if uttered by another speaker.

Stylianou *et al.* proposed a statistical voice conversion method using Gaussian mixture model (GMM) [1], and Kain *et al.* improved this method for increasing the variety of speaker's voices without expansion of database in TTS application [2]. Toda *et al.* also improved [1] by introducing Dynamic Frequency Warping for solving the over-smoothing problem of spectral envelope of conversion speech [3]. In these statistical approaches, source speech is decomposed to spectral and prosodic parameters frame by frame. These parameters are modified according to spectral stochastic tendency, and then they are re-synthesized as output speech. This approach is effective especially in case of less speech data condition.

On the other hand, segment-based approach has also attempted. Abe *et al.* proposed a method that converts source speech to target speech phoneme-segment by phoneme-segment [4]. At the offline procedure of this method, source and target speakers read same texts because of construction of the mapping table which provides the correspondence information between a source speaker's waveform segments and a target speaker's waveform segments. Each segment is decomposed into LPC parameters, and then is accumulated in each corpus. The conversion procedure is as follows.

1. An input speech of the source speaker is dictated and divided into phoneme segments.
2. The system searches through the source speaker's corpus, and finds the optimal segment which has the minimum DTW (Dynamic Time Warping) score that is calculated from the input segment and itself.
3. According to the mapping table, the optimal segment of the source speaker is replaced with the target speaker's segment.
4. The output speech is re-synthesized from LPC parameters of the target speaker's segments.

This segmental approach can preserve the dynamic characteristics of natural speech within each segment. Abe *et al.* obtained better result than their former "frame by frame" approach in terms of speaker individuality. Furthermore, DTW score can be regarded as a target cost of unit selection in concatenative speech synthesis described later.

Sündermann *et al.* proposed text-independent voice conversion based on unit selection without employing linguistic knowledge [5]. This technique is developed for speech-to-speech translation.

In General, voice conversion is done with small training data because the necessity of large speech data restricts its application area. However, on the other hand, it is difficult to make a sufficient accurate conversion rule and to generate a high-quality conversion voice in case of less data condition. That is, there is a relation of trade-off between data size and conversion quality. Furthermore, the excessive decomposition and re-synthesis to natural speech cause the lacks of naturalness and individuality.

The quality of conversion speech is generally influenced by gender, that is, (a) the individuality of speech which has converted from a male speaker to a female is more deteriorated compared with the speech which has converted from a male speaker to another male, and (b) the sound quality of speech which has converted from a male speaker to a female is inferior compared with the speech which has converted from a male speaker to another male.

III. Overview of Concatenative Speech Synthesis

Concatenative speech synthesis, which is the one of corpus-based time-domain TTS, employs a large amount of calculation and a large speech corpus which accumulates many natural speech waveform segments [7]–[9]. Synthetic speech by this method has high-naturalness and high-individuality of a speaker because of recycling of natural speech waveform segments involving these characteristics. The processing flow is shown in Fig.1.

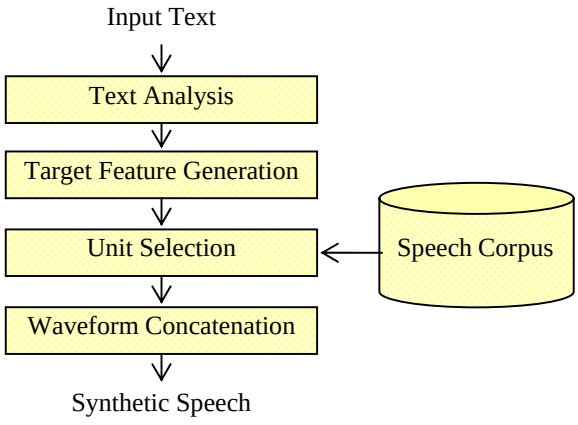

Fig. 1 The procedure of concatenative speech synthesis

Text information such as pronunciation and accent is extracted from input texts at the text analysis part. At the target feature generation part, ideal prosodic parameters of synthetic speech are estimated from the text information. The most preferable waveform segments sequence is decided at the unit

selection part, and then the segments are connected mutually.

Unit selection is defined as the minimum cost path searching problem of the network constructed by waveform segments in speech corpus. Dynamic programming algorithm is the one of typical solutions of this problem. Two kinds of costs are introduced for ranking each segment candidate: target cost and join cost. As shown in Fig. 2, Target cost $C_{tgt}(t_i, u_i)$, which expresses the degree of quality degradation caused by the difference between *i*-th candidate unit u_i and *i*-th target t_i, is defined as Euclidean distance of feature vectors. Join cost $C_{join}(u_i, u_{i-1})$, which expresses the degree of distortion between *i*-th candidate unit u_i and preceding candidate unit u_{i-1} which are stuck mutually, is also defined as Euclidean distance of feature vectors. Total cost of *i*-th candidate unit u_i is obtained by integrating these two costs. Not only spectral features and phonemic context but also prosodic features are used the calculations.

There is a correlation between the quality of synthetic speech and the corpus size. By researches in recent years, it was clarified that the high-naturalness synthetic speech can be realized by concatenative speech synthesis with extra large speech corpus [9].

This method avoids the "decomposition and re-synthesis" process such as LPC synthesis. The system reuses a speaker's natural waveforms to make a synthetic speech. Accordingly, this approach can also achieve high-individuality. It can be thought that this merit is kept even if the corpus is not so large, because the speaker individuality is kept within each natural waveform segment. For instance, in case of CHATR [7] with a speech corpus of less than one hour, although synthetic speech has discontinuity of waveform concatenation and lack of naturalness on prosody, it still has high-individuality [10].

Therefore, there is a possibility that the individuality of conversion speech can be improved by introducing the essence of concatenative speech synthesis into voice conversion framework. In this paper, the authors clarify this possibility by using about 40 minutes reading corpus.

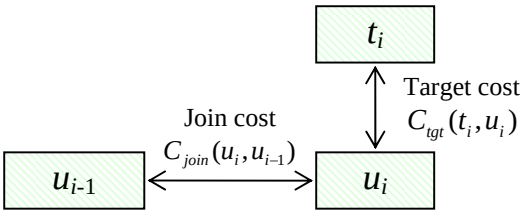

Fig. 2 Cost calculation in unit selection of concatenative speech synthesis

IV. Voice Conversion Based on Unit Selection

This section describes about the proposal voice conversion method based on concatenative speech synthesis. The procedure of the proposal is shown in Fig. 3. The method of Abe *et al.* [4] is the origin of this proposal framework because there are a lot of common parts between their method and concatenative speech synthesis. The main differences of them are as follows: (a) the "LPC decomposition and re-synthesis" process is not employed in the proposal method, and (b) join cost is introduced in the proposal method. The proposal method has two stages described following subsections: training stage and conversion stage. Three kinds of cost functions for unit

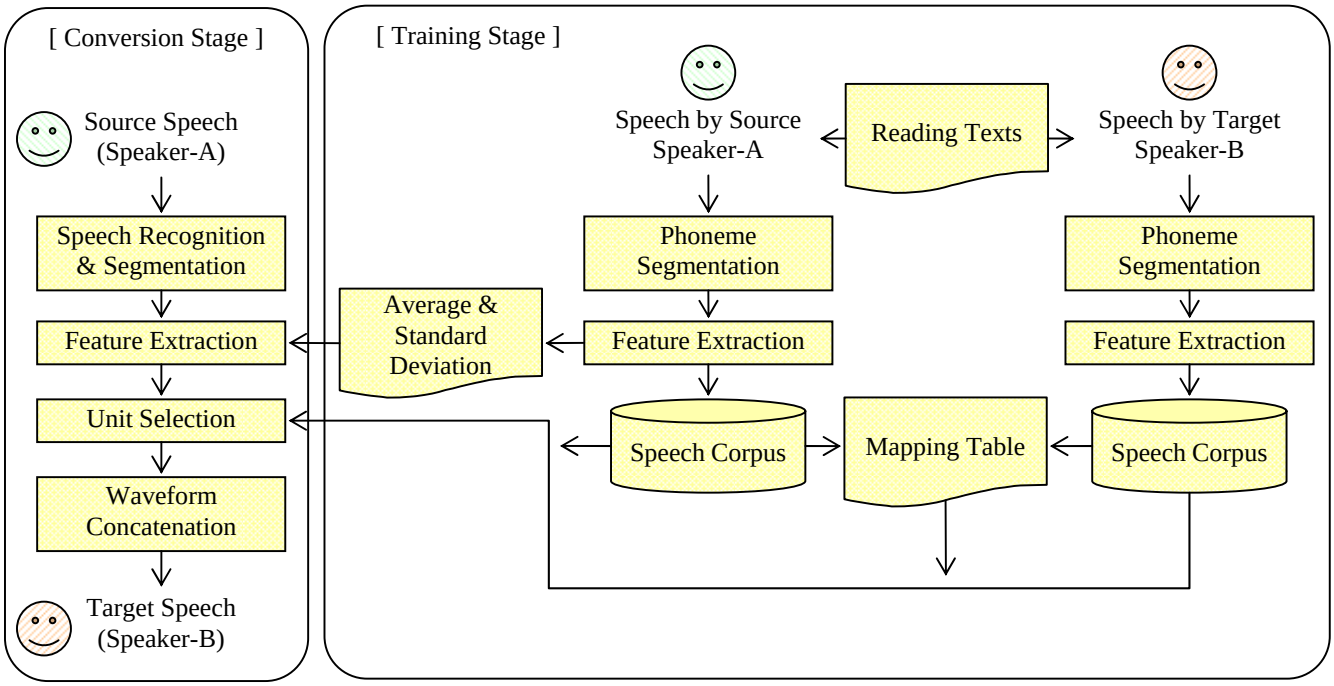

Fig. 3 The procedure of concatenative voice conversion

selection are employed in this paper.

A. Training Stage

Training procedure is defined as an offline process that associates a source speaker's waveform segments and a target speaker's waveform segments for building speech corpora which are used at the conversion stage.

The training procedure is as follows.

1. **Recording:** source speaker and target speaker read same texts to collect their natural waveform data.
2. **Phoneme segmentation:** the recorded speech data is divided into phoneme segments. In this paper, this processing is done by an open-source HMM speech recognizer for Japanese spoken language called as Julius [11].
3. **Feature extraction:** prosodic features (F_0, duration, and power) and spectral feature (Mel-cepstrum) are extracted from each segment. Twelve-dimensional Mel-cepstrum is obtained by SPTK [12]. Prosodic features are obtained by Snack [13], and are transformed into z-score for normalizing distribution of arbitrary speaker's characteristics. Average and standard deviation of prosodic features of source speaker are preserved for taking z-score at the conversion stage.
4. **Corpus construction:** speech corpus, which accumulates source speaker's and target speaker's waveform segments and their features, is built.
5. **Mapping of segments:** mapping table, which has correspondence between source and target speaker's waveform segments, is prepared.

In above procedure, there is necessity that source and target speakers read same texts. Although this requirement is indispensability in most conventional methods except the methods such as [5], the satisfaction of this condition becomes difficult when the corpus size is expanded. Therefore, in this paper, another method which avoids this requirement is also proposed. In this case, the training of the method is done as follows.

1. **Recording:** source and target speaker's reading speech are collected. The texts need not be same in this case.
2. **Phoneme segmentation:** each recorded speech is divided into phoneme segments.
3. **Feature extraction:** prosodic and spectral features are extracted. Target speaker's prosodic feature takes z-score to normalize. Average and standard deviation are extracted from prosodic features of source speaker.
4. **Corpus construction:** speech corpus, which accumulates target speaker's waveform segments and their features, is built (source speaker's each waveform segment is unnecessary).

That is, in this latter method, the statistics of source speaker and the speech corpus of target speaker are prepared. The processing for each speaker is independently done. This approach is used by the proposal method (III) described next subsection.

B. Conversion Stage

Conversion procedure is defined as an online process that arbitrary input speech of the source speaker is changed to the other speech as if uttered by the target speaker. The processing flow is as follows.

1. **Speech recognition and segmentation:** input speech is dictated and divided into phoneme segments.
2. **Feature extraction:** prosodic and spectral features are extracted from the phoneme segments. Each prosodic feature takes z-score to normalize. These features are used as criteria at succeeding unit selection.
3. **Unit selection:** The optimal waveform segments are selected from the target speaker's corpus (described below).
4. **Waveform concatenation:** selected waveform segments are stuck mutually.

The proposal cost calculation diagram of unit selection is depicted in Fig. 4. The source speaker-A's *i*-th input segment is denoted by $c_a(i)$. Segment $u_b(i)$ and $u_b(i-1)$ represent *i*-th and its preceding candidate units of the target speaker-B. The *i*-th unit of source speaker-A, corresponding to unit $u_b(i)$, is denoted by $u_a(i)$. Following three kinds of cost functions are employed in this paper.

(a) **Prosodic target cost:** this calculation between $c_a(i)$ and $u_b(i)$ is the conventional target cost of concatenative speech synthesis described in Section III.

(b) **Join cost:** this is also the conventional join cost between $u_b(i)$ and $u_b(i\text{-}1)$.

(c) **DTW score:** DTW score between $c_a(i)$ and $u_a(i)$ is used as a target cost. This cost is same as the method by Abe *et al*. described in Section II.

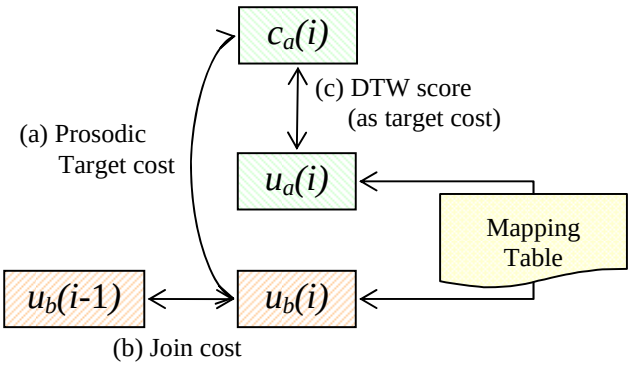

Fig. 4 The proposal cost functions

To examine the behavior of these costs, the authors have prepared following three types of unit selection.

(I) **Employing cost (b) and (c):** this type can regard as introducing the join cost of concatenative speech synthesis into the conventional conversion method of Abe *et al*.

(II) **Employing all costs:** this type can regard as adding the prosodic target cost and join cost of concatenative speech synthesis into the conventional conversion method of Abe *et al*.

(III) **Employing cost (a) and (b):** This is considered the unit selection of concatenative speech synthesis with using normalized features. In this case, the mapping table is unnecessary because the source speaker's unit $u_a(\cdot)$ is unused for this cost calculation. That is, the training of the source speaker is unnecessary except obtaining a few statistics (average and standard deviation only) to calculate z-score.

V. Evaluation and Discussion

To investigate the performance of the proposal method, the authors had two subjective experiments about the individuality and naturalness of conversion speech. The text reading corpora, which were read by two male and two female speakers, were used in this paper. Each corpus size is about 40 minutes (29054 phonemes including pause labels). Incidentally, in this paper, phoneme information of input speech is made manually to evaluate the performance of the proposal unit selection.

A. ABX test

We had an ABX test for evaluation of individuality of the proposal voice conversion speech. "X" of ABX is conversion speech in this experiment. "A" and "B" are a source speaker's speech and a target speaker's speech (order of playback is exchanged randomly). Four patterns of voice conversion (i.e., from female to another female, female to male, male to female, and male to another male) were tested. Number of stimuli is 10 phrases for each pattern. These stimuli were made by using the conversion method (II) described in section IV. In addition, 10 natural speech phrases, which were assumed as the results of absolutely perfect conversion, were also contained in each pattern to get base line. Five subjects listened with headphone in our office (not in a soundproof chamber). Listening time was once.

The result is shown in Table 1. As shown in Table 1, all results except type (1) are perfect. Even in case of type (1), the number of false answers was only one. Thus, the authors think that this mistake would be a human error. Generally, when the gender of a source speaker is not same to a target speaker, the judgment is easy because there is a decisive difference of F_0. On the other hand, the judgment about the conversion between same gender's speakers is not as easy as the former case. However, from the result, it was confirmed that the proposal method converted individuality of speaker well in both cases.

TABLE I
THE RESULT OF ABX TEST

No.	Conversion Type "X"	Source	Target	Correct rate [%]
(1)	conversion speech	Female	Female	98
(2)			Male	100
(3)		Male	Female	100
(4)			Male	100
(5)	Natural speech (as base line)	Female	Female	100
(6)			Male	100
(7)		Male	Female	100
(8)			Male	100

The reason of this would be same as the feature of concatenative speech synthesis; i.e., the reusing of natural waveform segments containing individuality.

B. MOS test

The naturalness of conversion speech was evaluated by MOS test. In this experiment, not only the three proposal methods (I) to (III) described in section IV, but also the following three kinds of conversion speech were examined:

(IV) using the cost (c) only described in Section IV,
(V) conversion speech from a female speaker to herself,
(VI) natural speech.

Type (IV) is almost similar to the conventional method by Abe *et al*. In case of type (V), the conversion is done by using type (II) and it avoids the influence of the feature normalization of z-score. Type (VI) is assumed as the results of absolutely perfect conversion (it is used for obtaining base line). In terms of investigation of influences of gender's difference, we tested two kinds of conversion (i.e., from female to another female, and from male to female). Number of stimuli is 10 sentences for each pattern. In this experiment, 8 subjects listened with headphone in our office and gave the subjective evaluation score. The range of evaluation score is from 1 to 5 (bad, poor, fair, good, and excellent, respectively). Listening time was once.

The result is shown in Table 2. The difference between three proposals is unclear in these results. The reason of this would be that the weight for the join cost (b) described in section IV was set to comparatively large. On the other hand, since the average of type (IV) is the lowest, we can say that the introducing of join cost is effective well. Furthermore, the result, that type (III) has almost same performance as the other proposal types, shows the possibility (similar to [5]) that the proposal method can remove the restrictions of the training stage such as unity of reading texts.

The difference of genders generally causes the degradation of quality of conversion. However, in this result, the clear differences about gender were not observed. Hence, it has clarified that the proposal method is robust against gender.

Although the proposal method improved the naturalness of conversion speech, all conversion speech results are still lower than the natural speech (VI). Especially from the result of type (V), it is thought that synthesis method must be more improved in future works. Several future works are enumerated below:

1. investigation about the influence of speech recognition error and its solution (e.g., employing acoustical

TABLE II
THE RESULT OF MOS TEST

Kind of stimuli	Female → Female ave. [a]	Female → Female s.d. [b]	Male → Female ave.	Male → Female s.d.
(I) Proposal	2.78	0.79	3.05	0.63
(II) Proposal	3.06	0.80	2.94	0.75
(III) Proposal	3.03	0.88	2.75	0.81
(IV) Conventional	1.53	0.71	1.58	0.65
(V) Female to herself	3.08	0.76		
(VI) Natural speech	4.96	0.25		

[a]ave. = average, [b]s.d. = standard deviation.

clustering without linguistic knowledge, instead of employing phoneme segments),
2. employing and evaluating post-processing for prosodic modification of waveform segments, or corpus expanding,
3. applying this conversion method to the conversation speech processing which is our previous work [14].

VI. Conclusion

This paper proposed an approach for voice conversion based on concatenative speech synthesis in terms of especially improving the individuality of conversion speech. From two kinds of subjective experimental results, it has confirmed that (a) the proposal method converts the speaker individuality well, (b) the introducing of the join cost of concatenative speech synthesis into voice conversion improves the naturalness of conversion speech, and (c) the proposal method is robust against the difference of genders of speakers.

References

[1] Y. Stylianou, O. Cappé, and E. Moulines, "Statistical methods for voice quality transformation," *Proc. of EUROSPEECH*, pp.447–450, September 1995.
[2] A. Kain, and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," *Proc. of International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp.285–288, 1998.
[3] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of straight spectrum," *Proc. of International Conference on Acoustics, Speech and Signal Processing*, Vol. 2, pp.841–844, 2001.
[4] M. Abe, "A segment-based approach to voice conversion," *Proc. of International Conference on Acoustics, Speech and Signal Processing*, pp.765–768, 1991.
[5] D. Sündermann, H. Höge, A. Bonafante, H. Ney, A. Black, and S. Narayanan, "Text-independent voice conversion based on unit selection," *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 2006.
[6] E. Keller, G. Bailly, A. Monaghan, J. Terken, and M. Huckvale, *Improvements in Speech Synthesis*, John Wiley & Sons, 1st Ed. 2001, ch. 1.
[7] N. Campbell, "CHATR: A high-definition speech re-sequencing system," *Proc. of ASA/ASJ Joint Meeting*, pp.1223–1228, Honolulu, December 1996.
[8] N. Campbell, and A. W. Black, "Prosody and the selection of source units for concatenative synthesis," in Progress in Speech Synthesis, Springer Verlag, Inc., New York, 1995, ch. 22.
[9] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "Ximera: A New TTS from ATR Based on Corpus-Based Technologies,"' *Proc. of ISCA 5th Speech Synthesis Workshop*, pp.179–184, Pittsburgh, U.S.A., June 2004.
[10] Synthetic speech sample demonstration of CHATR. Available: http://feast.atr.jp/chatr/chatr/e_tour/synth_examples.html
[11] Open-Source Large Vocabulary CSR Engine Julius. Available: http://julius.sourceforge.jp/en_index.php?q=en/index.html
[12] Speech Signal Processing Toolkit (SPTK) Ver 3.0. Available: http://kt-lab.ics.nitech.ac.jp/%7Etokuda/SPTK/index.html
[13] The Snack Sound Toolkit. Available: http://www.speech.kth.se/snack/
[14] K. Fujii, R. Ueda, H. Kashioka and N. Campbell, "A trial to apply concatenative speech synthesis to spontaneous speech," *Proc. of International Technical Conference on Circuits/Systems, Computers and Communications*, Vol. 2, pp.653–656, 2006.

Kei Fujii received his B.E. degrees from the Yamaguchi University, Yamaguchi, Japan, in 1999. And in 2001, he received his M.E. degrees from Nara Institute of Science and Technology, Nara, Japan. From 2004, he is working at Kumamoto National College of Technology as an assistant professor. From 2006, he is also belonging to the doctorial course at Kumamoto University. He is currently engaged in the research of speech information processing, especially speech synthesis.

Jun Okawa received his Associate degree in engineering from Kumamoto National College of Technology, Kumamoto, Japan, in 2007. Currently, he is working at Japan Software Engineering Co., Ltd., Tokyo, Japan.

Kaori Suigetsu received her Associate degree in engineering from Kumamoto National College of Technology, Kumamoto, Japan, in 2007. And she is currently the B.E. course student at the Advanced Course of Control and Information Systems Engineering in Kumamoto National College of Technology. She is currently engaged in the research of the synchronizing of 3-D lip animation and utterance.