SCALING OF WAVEFORM SEGMENTS ALONG THE TIME AXIS FOR CONCATENATIVE SPEECH SYNTHESIS

Nobuyuki Nishizawa and Hisashi Kawai

ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288 Japan
{nobuyuki.nishizawa, hisashi.kawai}@atr.jp

ABSTRACT

Waveform scaling along the time axis is introduced as a pitch and duration conversion method for concatenative speech synthesis. With this method, although not only F_0 and duration but also spectrum are affected, no degradation of naturalness is caused when the scaling ratio is nearly 1. In corpus-based concatenative speech synthesis, when there are many segment candidates with various F_0 values or durations, excessive scaling may be unnecessary. The result of experiments indicated that the difference in F_0 and duration between the target and a selected segment became smaller. However, it also showed that the conventional cost function in selection cannot represent the degradation of naturalness by spectral distortion, and that scaling range without the degradation may not be enough for the pitch conversion required in our synthesizer. These problems should be improved by wider range scaling with a new cost function that also considers the degradation.

1. INTRODUCTION

In a concatenative speech synthesizer, the naturalness of synthetic speech sounds is determined by the difference between a target sound and a selected sound from a speech corpus. For more natural speech sound, optimizing the cost, which is a criterion of the segment selection, was studied. In our synthesizer, by relating the cost to perceptual characteristics, naturalness of speech was improved[1][2]. However, since sounds that are not included in the corpus are impossible to generate by a simple concatenative synthesizer (not including complicated signal processing like pitch conversion), the improvement of naturalness by optimizing the design of the cost has a limitation. Although naturalness of synthetic sounds is improved more by using a larger corpus, the improvement becomes smaller as the size of the corpus becomes larger; the size of the corpus that is required for truly natural speech sounds is unknown.

Therefore, the adoption of waveform conversion techniques is desired. However, most speech sound conversion techniques tend to cause degradation of naturalness. For example, the PSOLA (pitch-synchronous overlap and add)[3] method can cause serious degradation of naturalness if the extracted pitch is inaccurate. Robust pitch extraction is difficult especially in sounds with unstable vibrations. If the number of waveform segments in the database is small, manual correction of the pitch extraction results may be possible. But in a corpus-based speech synthesis, manual correction of all segments is impractical. For that reason, no PSOLA-based prosodic conversion techniques have been adopted to our synthesizer. On the other hand, STRAIGHT[4]-based speech conversion is known as a relatively high-quality and robust sound conversion method. But a little degradation of quality is actually caused through its vocoder-like signal processing even when the conversion ratio is exactly 1. By the degradation, the naturalness of synthetic speech in Japanese by concatenative synthesis with STRAIGHT-based prosodic conversion was found to be inferior to that by simple concatenative synthesis when a corpus of over about 20 hours was used[5].

In contrast, waveform scaling along the time axis is the simplest speech conversion method related to prosody. However, waveform scaling affects not only fundamental frequency (F_0) and duration, but also spectrum. In contrast to PSOLA or STRAIGHT, though, this method is robust and causes no degradation in synthetic sound quality when the conversion ratio is nearly 1. However, F_0 and duration cannot be changed separately. Moreover, a little conversion makes speech sounds unnatural because the spectrum is also scaled linearly. Although a pitch conversion method using waveform scaling has been proposed by Higashi *et al.*[6], in their proposal, restoration of the spectral distortion and the changed duration based on the LSEE-MSTFTM (least squares error estimation from modified short time Fourier transform magnitude) algorithm proposed by Griffin *et al.*[7] was also introduced for wide range conversion. However, in concatenative synthesis with a large corpus, a slight conversion may be enough because there are many segment candidates with various F_0 values or durations in the corpus. Therefore, in our case, the restoration of spectrum and duration is not introduced for simplicity. Waveform scaling is used only in ranges without degradation of naturalness.

The rest of this paper is constructed as follows: Section 2 explains segment selection in our concatenative synthesizer for later discussion. In Section 3, a flexible sampling frequency converter for the proposed waveform scaling is introduced. In Section 4, the proposed method is integrated with segment selection and the result of an experiment is shown and discussed. Section 5 estimates the required range of the scaling by the proposed method with a large corpus. Finally, Section 6 concludes the paper.

2. SEGMENT SELECTION FOR CONCATENATIVE SPEECH SYNTHESIS

In concatenative speech synthesis, a criterion of segment selection is often called a cost. In many cases, the cost is calculated from some sub-cost functions, each of which corresponds to a factor of sound. Since the cost and the sub-cost functions were designed to correspond to the naturalness of synthetic sounds, the naturalness of sounds can be discussed by referring to the cost or sub-cost values.

In this section, for a discussion of influences by waveform

scaling in each sub-cost, our design of the cost and the sub-cost functions is introduced.

2.1. Sub-cost functions

In our synthesizer, six kinds of sub-costs, C_{F_0}, C_{dur}, C_{cen}, C_{env}, C_{spg} and C_{F_0c} are considered.

C_{F_0}, and C_{dur} correspond to the degradation of naturalness by the difference in $\log F_0$ and duration, respectively. C_{cen} corresponds to the degradation of naturalness by the difference of cepstral mean between a candidate and the target. In our text-to-speech system, since all of the target parameters are generated using an HMM-based speech synthesis framework[8], cepstrum is also usable as a target feature. C_{F_0}, C_{dur} and C_{cen} compose a target cost C_T.

C_{env} corresponds to the degradation of naturalness by a mismatch of the phonetic environment between a candidate and the target. C_{env} also includes influences by concatenation. When two segments are not connected in the corpus, C_{env} is not 0 even if the segments match in the phonetic environment. The C_{spg} corresponds to the degradation of naturalness by spectral discontinuities at segment boundaries. This is calculated from the weighted sum of mel-cepstral distances between the proceeding segment and the following segment around the boundary. C_{F_0c} corresponds to the degradation of naturalness by the discontinuity in $\log F_0$. C_{env}, C_{spg} and C_{F_0c} compose a concatenation cost function C_C. When two segments are connected in the corpus, C_{env}, C_{spg} and C_{F_0c} are set to 0.

Each design of these sub-cost functions was optimized based on the results of perceptual experiments[2].

2.2. Calculation of cost

Segment selection is done based on minimization of an integrated cost. The integrated cost function C coresponds to the degradation of naturalness when segment sequences $\{u_i\}$ are used. C is given by

$$C(u_0,u_1,\cdots,u_{N-1}) = w_T \cdot \left(\frac{1}{N}\sum_{i=0}^{N-1} C_T(u_i,t_i)^{p_T}\right)^{\frac{1}{p_T}} + w_C \cdot \left(\frac{1}{N-1}\sum_{j=1}^{N-1} C_C(u_{j-1},u_j)^{p_C}\right)^{\frac{1}{p_C}} \quad (1)$$

$$w_T + w_C = 1 \quad (2)$$

where C_T, C_C and $\{t_i\}$ denote a target cost, a concatenation cost and sequence of target segments, respectively. Also, w_T, p_T, w_C and p_C denote the weight and the power coefficient of the target cost and the concatenation cost, respectively.

C_T and C_C are given by

$$C_T(u_i,t_i) = w_{F_0}\cdot C_{F_0}(u_i,t_i) + w_{dur}\cdot C_{dur}(u_i,t_i) + w_{cen}\cdot C_{cen}(u_i,t_i) \quad (3)$$

$$w_{F_0} + w_{dur} + w_{cen} = 1 \quad (4)$$

$$C_C(u_{j-1},u_j) = w_{env}\cdot C_{env}(u_{j-1},u_j) + w_{spg}\cdot C_{spg}(u_{j-1},u_j) + w_{F_0c}\cdot C_{F_0c}(u_{j-1},u_j) \quad (5)$$

$$w_{env} + w_{spg} + w_{F_0c} = 1 \quad (6)$$

relatively, where w_{F_0}, w_{dur}, w_{cen}, w_{env}, w_{spg} and w_{F_0c} denote the weights of each sub-cost, respectively. These coefficients were also estimated by the results of perceptual experiments.

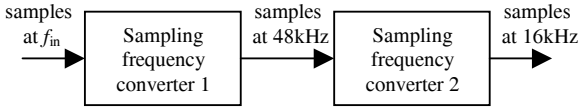

Fig. 1. Structure of the proposed sampling frequency converter.

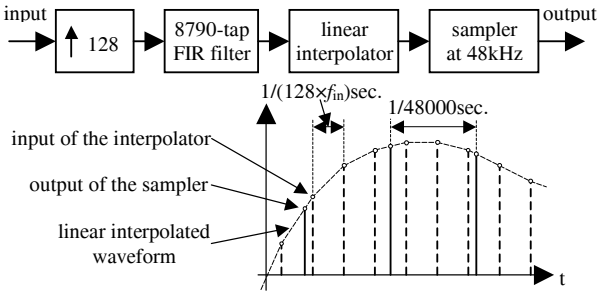

Fig. 2. Schematic diagram of sampling frequency conversion in the first stage of the converter.

3. WAVEFORM SCALING ALONG THE TIME AXIS

In a discrete-time system, waveform scaling along the time axis can be handled as a sampling frequency conversion. For example, contraction of a waveform is equal to raising its sampling frequency. For more precise waveform scaling, non-integer-ratio conversion should be possible. Moreover, to change the sampling frequency dynamically, the designs of all filters should be fixed. In this section, a flexible sampling frequency converter that satisfies these requirements is introduced.

3.1. Configuration of a flexible sapling frequency converter

Fig. 1 shows the configuration of the introduced converter. The converter consists of two main stages. In this research, the converter is only used as a decimator because the sampling frequency of the output is 16 kHz although that of the corpus is 48 kHz.

Fig. 2 shows the first stage of the converter. In the first stage, the sampling frequency is converted from an arbitrary input frequency to 48 kHz using the following steps:

1. Up-sampling to $128 \times f_{in}$
2. Low-pass filtering by an 8790-tap FIR filter to cut all components over $\frac{f_{in}}{3}$. To lower the order of the filter, the upper edge of the pass band of the filter is set to $\frac{f_{in}}{4}$.
3. Conversion to a continuous-time system signal with linear interpolation.
4. Conversion to a discrete-time system signal by re-sampling at 48kHz

This method is equivalent to a method proposed by Katsumata *et al.*[9]. The cut-off frequency of the output depends on the sampling frequency of the input when the design of the converter filter is fixed. In this configuration, the filter in step 2 should cut all components over 24 kHz; i.e., f_{in} must be set below 72kHz.

In the second stage, waveforms are decimated to the final output frequency by the following steps:

5. Low-pass filtering by a 696-tap FIR filter to cut all components over 8kHz.

6. Down-sampling to 16kHz.

Through this process, components above 8kHz are removed. To preserve all components below 8kHz, f_{in} must be above 32kHz due to the design of the filter in step 2.

4. INTEGRATION OF THE SCALING TO SEGMENT SELECTION

When a segment waveform is scaled, all sub-costs related to the segments except C_{env} is changed. Therefore, the optimal conversion ratio cannot be estimated before segment selection and should be estimated simultaneously with segment selection.

4.1. Estimation of optimal conversion ratio at segment selection

Ideally, the scaling ratio should be decided online to reduce the number of candidates in segment selection. However, the method for effectively searching for the optimal ratio is not obvious. Therefore, for simplicity, converted waveforms were made and added to the corpus beforehand in this research. Using this method, the size of the corpus is enlarged virtually. For example, in the following evaluation, waveforms were converted beforehand from -200 cents (1200 cents = 12 semitones = 1 octave) to 200 cents with 25-cent separation in $\log F_0$; i.e., in this case, the size of the corpus was virtually made about 17 times as large as the original corpus.

4.2. Evaluation on sub-costs

For evaluation of waveform scaling, the changes in the sub-costs were investigated. Using the method described in section 4.1, waveform scaling was introduced into segment selection. The conversion was from -200 cents to 200 cents with 25-cent separation in $\log F_0$. Although the sampling frequency of the corpus is 48 kHz, each segment was down-sampled to 16 kHz with waveform scaling by the proposed sampling frequency converter. All of the features of each segment were extracted from the 16-kHz-sampling sounds. The speech corpus used for synthesis was a 450 sentences corpus pronounced by a male speaker. The sentences were the same sentences in ATR's 503-sentence corpus[10] except set J. The synthetic targets were set J sentences in the ATR's corpus (53 sentences). All of the targets of the synthetic sounds were extracted from the sounds of set J sentences by the same speaker.

For comparison, the maximum conversion range was changed. In this experiment, ± 0, ± 25, ± 50, ± 75, ± 100, ± 150 and ± 200 cents in $\log F_0$ were selected as the range.

Fig. 3 and Fig. 4 show the weighted sub-costs included in the target cost and the concatenation cost, respectively. Each point indicates the average of 53 sentences in each weighted sub-cost per segment. From the results, apparently, the sub-costs of F_0 and duration declined as the maximum range became wider, due to the increase of candidates. In contrast, the sub-cost of the phonetic environment increased when the range was narrower than ± 100 cents. Since C_{env} also represents the degradation of naturalness by the concatenation of segments that are not connected in the corpus, an increase of concatenation for decreasing the prosodic cost made C_{env} increase. On the other hand, when the range was wider than ± 100 cents, decrease of mismatch in the phonetic environment by increase of candidates caused C_{env} to decrease.

The problem is that C_{cen} hardly changed as C_{F_0} and C_{dur} decreased. According to preliminary listening, the synthetic speech

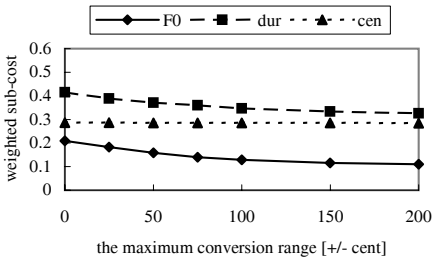

Fig. 3. Weighted sub-costs per segment in a target cost. F0, dur and cen denote $w_{F_0} \cdot C_{F_0}$, $w_{dur} \cdot C_{dur}$ and $w_{cen} \cdot C_{cen}$, respectively.

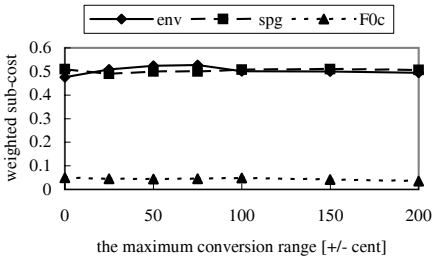

Fig. 4. Weighted sub-costs per segment in a concatenation cost. env, spg and F0c denote $w_{env} \cdot C_{env}$, $w_{spg} \cdot C_{spg}$ and $w_{F_0c} \cdot C_{F_0c}$, respectively.

with segments converted over ± 50 cents seemed unnatural in many cases. This implies that C_{cen} cannot represent the degradation of naturalness by waveform scaling appropriately. Therefore, it is necessary to make the conversion range narrower or to add another sub-cost to the target cost function in order to represent the degradation of naturalness by waveform scaling.

5. DISCUSSION OF THE REQUIRED CONVERSION RANGE

The result in Section 4.2 indicates that the current design of the cost function cannot manage degradation of naturalness by waveform scaling appropriately. Although a conversion range without degradation of naturalness should be estimated by perceptual experiments, the required conversion range with a large corpus was not clear. In this section, the required conversion range is estimated.

5.1. Estimation of the required conversion range with a large corpus

For optimal results, selection against pre-converted segments like the method in Section 4 is necessary. However, when a large corpus is used, segment selection with an enlarged segment database is very difficult. Therefore, in this evaluation, waveform scaling based on minimization of F_0 error was applied to selected segments by the former method. Since the result of the sub-cost evaluation in Section 4.2 implies that error of duration is less critical than error of F_0, and change in C_{cep} by the waveform scaling hardly affects to result of segment selection, the difference between sounds of the pre-converted and optimally selected segments and the post-converted sounds based on minimization of F_0 error may be small.

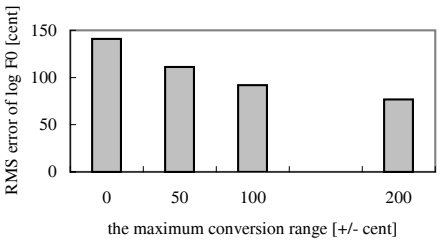

Fig. 5. RMS error per segment of $\log F_0$ in the synthetic speech sounds

The speech corpus that was used for speech synthesis includes sentences in dialogues and newspapers pronounced by a male speaker. The total number of sentences is 45,032 (about 59.3 hours in all). The sampling frequency of the corpus and the output of the synthesizer are 48 kHz and 16 kHz, respectively. Set J in ATR's 503 sentences was used as target sentences. The targets of synthesis were extracted from speech sounds by the same speaker. In this experiment, ± 0, ± 50, ± 100 and ± 200 cents in $\log F_0$ were selected as the maximum conversion range. In the estimation, segment selection was first done. Next, waveform scaling was applied to each segment to decrease the error of $\log F_0$ in the predefined range. The conversion ratio was fixed in each segment for simplicity. Finally, the error in $\log F_0$ was calculated.

Fig. 5 shows the RMS error per segment of $\log F_0$. The result shows that the RMS error in the case without conversion is about 140 cents. Although the error decreased as the maximum conversion range was set wider, the extent of the decrease is not enough. This indicates that the conversion ratio should be variable in the interior of each segment.

Preliminary listening of the converted sounds was also conducted. From the listening, sounds converted over ± 50 cents seemed unnatural, similar to the result in Section 4.2. On the other hand, the difference between sounds without conversion and sounds within ± 50 cent conversion was not apparent even in F_0. If the possible conversion range is statically decided only by whether the converted sound is natural or unnatural, the conversion range may be insufficient to perceptually reduce error in pitch. Therefore, for large conversions to appropriately reduce error as a whole, the extent of the degradation of naturalness by the conversion should be represented in the cost function.

6. CONCLUSIONS

A speech waveform conversion method based on waveform scaling along the time axis was introduced for concatenative speech synthesis. The scaling is based on sampling frequency conversion. Using this method, sounds can be converted without degradation of naturalness when the conversion ratio is nearly 1. For precise and flexible conversion, a non-integer-ratio sampling frequency converter was introduced.

Since some factors of a segment are affected by scaling, the optimal sequence of selection is not decided before segment selection. Therefore, in order to integrate the effects of scaling into the cost calculation, the waveforms were scaled beforehand and the size of the corpus was enlarged virtually. The result of a segment selection experiment indicated a decrease of cost related to F_0 and duration. However, it was shown that the former design of cost could not represent the degradation of naturalness by waveform scaling.

For a discussion of the required range of the conversion for concatenative synthesis, error of $\log F_0$ in synthetic sounds was also estimated. The result showed that conversion range should be variable in the interior of each segment. On the other hand, preliminary listening of the sounds showed a wider range than the range without the degradation of naturalness may be required. Therefore, for large conversions to reduce error as a whole, the extent of the degradation of naturalness by waveform scaling should be represented in the cost function.

In future work, the cost function will be re-designed by the result of perceptual experiments to represent the degradation of naturalness by waveform scaling.

Acknowledgements: This research was supported in part by the Telecommunication Advancement Organization of Japan.

7. REFERENCES

[1] T. Toda, H. Kawai, and M. Tsuzaki, "Optimizing Integrated Cost Function for Segment Selection in Concatenative Speech Synthesis Based on Perceptual Evaluations," Proc. EUROSPEECH, Geneva, Switzerland, pp. 297–300, Sep. 2003.

[2] T. Toda, H. Kawai, and M. Tsuzaki, "Optimizing Sub-Cost Functions for Segment Selection Based on Perceptual Evaluations in Concatenative Speech Synthesis," Tech. Report of IEICE, SP2003-81, pp. 43–48, Aug. 2003 (in Japanese).

[3] E. Moulines, and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Communication, vol. 9, issues 5–6, pp. 453–467, Dec. 1990.

[4] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring Speech Representations Using a Pitch-Adaptive Time-Frequency Smoothing and an Instantaneous-Frequency-Based F0 Extraction: Possible Role of a Repetitive Structure in Sounds," Speech Communication, vol. 27, issues 3–4, pp. 187–207, Apr. 1999.

[5] T. Toda, H. Kawai, and M. Tsuzaki, "Effectiveness of Prosodic Modification in Concatenative Text-to-Speech Synthesis," Proc. ASJ 2003 Autumn Meeting, Japan, 1-8-10, vol. 1, pp. 201–202, Sep. 2003 (in Japanese).

[6] H. Higashi, and M. Kawamata, "A Method of Pitch Modification by Speech Synthesis from Short-Time Fourier Transform Magnitude," Tech. Report of IEICE, SP99-89, pp. 25–30, Oct. 1999 (in Japanese).

[7] D. W. Griffin, and J. S. Lim, "Signal Estimation from Modified Short-Time Fourier Transform," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.

[8] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, "Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis," Proc. ICASSP 2000, Istanbul, Turkey, vol.3, pp.1315–1318, June 2000.

[9] Y. Katsumata, and O. Hamada, "An Audio Sampling Frequency Conversion Using Digital Signal Processors," Proc. ICASSP 86, Tokyo, Japan, vol. 1, pp. 33–36, Apr. 1986.

[10] M. Abe, Y. Sagisaka, T. Umeda, and H. Kuwabara, Speech Database User's Manual, ATR Interpreting Telephony Research Laboratories Technical Report, TR-I-0166, Japan, Aug. 1990 (in Japanese).