A trial to improve the intelligibility of spontaneous concatenative speech synthesis

Kei Fujii
Department of Information and Computer Sciences
Kumamoto National College of Technology
2659-2, Suya, Kohshi-city, Kumamoto 861-1115
Email: fujii @ cs.knct.ac.jp

ABSTRACT

Realization of spontaneous speech synthesis is a one of recently focused research purposes. In previous research, authors reported about a trial to apply concatenative speech synthesis to spontaneous speech, especially, some evaluations of unit selection performance in concatenative speech synthesis. From these results, we obtained the limit of the conventional method. In this paper, TD-PSOLA is introduced to obtain the synthetic speech improving about spectral and prosodic discontinuity which causes degradation of intelligibility. As the result, intelligibility score of synthetic speech has been improved by 4%.

Keywords: Spontaneous speech synthesis, Concatenative speech synthesis, Unit selection, Intelligibility, TD-PSOLA

1 INTRODUCTION

Techniques about spontaneous speech processing are a one of recently focused research areas, and are expected to apply to machines that used for our life such as robots. Corpora of Japanese spontaneous speech have constructed in Japan these days [1][2] . The realization of spontaneous speech synthesis is the one of these purposes. Akagawa et al have investigated about the possibility to realize HMM-based spontaneous speech synthesis [3]. As same purpose, authors have tried to apply concatenative speech synthesis to spontaneous speech, and have obtained 77.4% of sentence intelligibility (94.2% in case of natural speech) from the evaluation experiment about spontaneous synthetic speech by concatenative speech synthesis with about 100-minutes speech corpus [4]. Discontinuity between adjacent speech wave segments is a one of the reasons of this degradation about intelligibility. In this paper, author introduces TD-PSOLA as post processing of concatenative speech synthesis for intelligibility improvement.

1.1 Concatenative speech synthesis

Concatenative speech synthesis is the method that makes speech sound well-keeping naturalness and voice quality of a speaker by introducing large speech corpus [5]. Speech corpus for this method accumulates various speech wave segments of a speaker and their features. Processing flow of this method for reading speech synthesis is shown in ***Fig. 1***. The target feature generation part makes a series of appropriate target prosodic features from parsed input text information. According to these target feature series, the unit selection part chooses the most preferable wave segment series from the corpus. Output synthetic speech is made by connecting these wave segments, thus, is recycling of a speaker's natural voices.

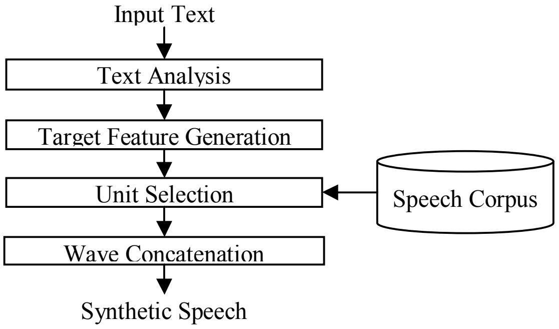

Fig.1: *Processing flow of concatenative Text-to-Speech synthesis*

In unit selection, quality degradation of a synthetic speech is expressed as two kinds of costs shown in ***Fig.2***.

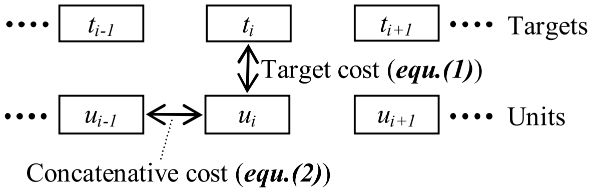
t_{i-1}
t_i
t_{i+1}
Targets
Target cost (***equ.(1)***)
u_{i-1}
u_i
u_{i+1}
Units
Concatenative cost (***equ.(2)***)

Fig.2: *Cost calculation in unit selection*

The target cost which expresses the quality degradation caused by the difference between *i*-th candidate unit u_i and *i*-th target t_i is defined as the norm of vector of sub-costs as follows

$$C_{tgt}(u_i,t_i)=\sqrt{\sum_{j=1}^{N_{tgt}}\left(SC_{tgt}(j,u_i,t_i)\right)^2}\ ,\qquad (1)$$

where N_{tgt} is the number of sub-costs (N_{tgt} = 3 in this paper: difference of F_0, difference of duration and difference of power). SC_{tgt} means the j-th sub-cost function that is calculated according to the weighted distance between a feature of candidate unit and a feature of target.

The concatenative cost that expresses the difference between two candidate units which are stuck mutually is defined as

$$C_{con}(u_i,u_{i-1}) = \sqrt{\sum_{j=1}^{N_{con}} \left(SC_{con}(j,u_i,u_{i-1})\right)^2} \text{ , (2)}$$

where N_{con} is the number of sub-costs (N_{con} = 3 in this paper: F_0, power and MFCC). SC_{con} is the j-th sub-cost function that is calculated as the weighted distance about prosodic features (F_0 and power) and phonetic features (MFCC) on the boundary.

These costs are integrated by follows

$$C(u_i,u_{i-1},t_i) = \sqrt{C_{tgt}(u_{i,}t_i)^2 + C_{con}(u_i,u_{i-1})^2} \text{ . (3)}$$

When a target series are given as $(t_1 \cdots t_M)$, the series of u which have the minimum value of the sum of ***equ. (3)*** are the most preferable. This series of u is searched through the corpus by using Viterbi algorithm.

Authors applied concatenative speech synthesis to spontaneous speech in the previous research, and obtained 77.4% of sentence intelligibility (94.2% in case of natural speech). One of the reasons of this degradation is the discontinuity of adjacent speech wave segments. In this paper, TD-PSOLA (described later) is introduced as post processing to connect each segment smoothly. ***Fig. 3*** shows the processing flow for performance evaluation in this paper. The target features are extracted from input natural speech (it is assumed that they are ideal features obtained by perfect target feature generation in ***Fig. 1***). Prosody of each selected wave segment is modified by TD-PSOLA to the same as target prosody.

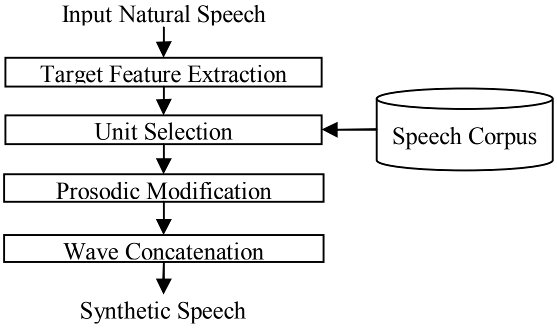

Fig.3: *Processing flow of synthetic speech generation for performance evaluation of unit selection and post processing*

1.2 TD-PSOLA

Time Domain Pitch Synchronous Overlap Add (TD-PSOLA) is a one of the techniques that modifies speech prosody [6]. Prosodic modification by TD-PSOLA is done according to the following procedures:

1. Each glottal closure position is marked on source speech waveform (giving the pitch mark to source speech waveform).
2. Each pitch waveform is cut out from source speech waveform with the window function corresponding pitch interval. The window function used in this paper is the hanning window whose length of the left side and the right side of the window are set to the same as the preceding and the following pitch mark intervals.
3. Each pitch waveform is placed to the nearest target pitch mark position.
4. Target speech waveform is obtained by adding each pitch waveform.

Fig. 4 and ***Fig. 5*** depict the procedure which modifies F_0 of speech. If F_0 is lowered, the target pitch mark interval is wider than the source pitch mark interval, hence some pitch waves are deleted (***Fig. 4***). Oppositely, if F_0 is raised, the target pitch mark interval is narrower than the source pitch mark interval, and some pitch waves are used two or more times (***Fig. 5***).

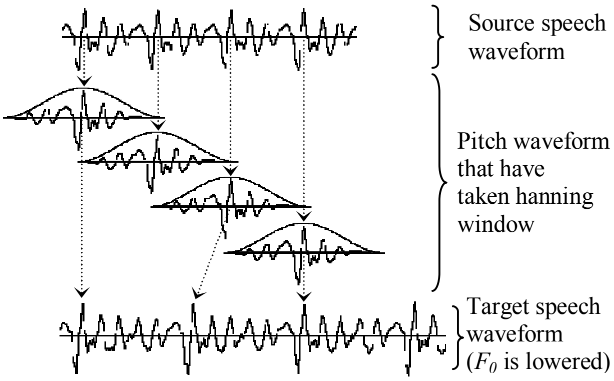
Source speech waveform
Pitch waveform that have taken hanning window
Target speech waveform (F_0 is lowered)

Fig.4: *Pitch modification (F_0 is lowered)*

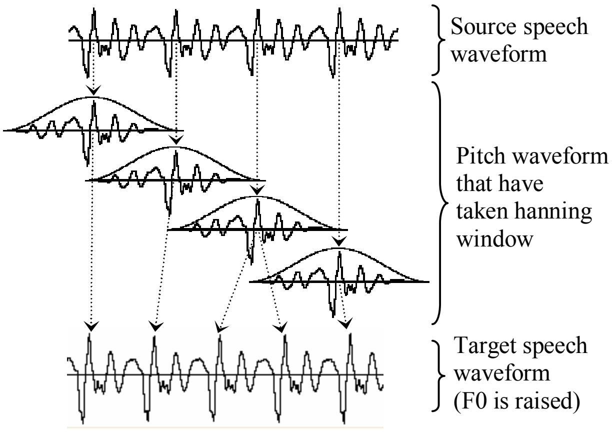
Source speech waveform
Pitch waveform that have taken hanning window
Target speech waveform (F0 is raised)

Fig.5: *Pitch modification (F_0 is raised)*

TD-PSOLA is simple processing and can modify speech prosody directly in time domain. If source pitch interval and target pitch interval are same, source waveform can be kept exactly. By introducing TD-PSOLA as post processing of unit selection, prosody of each selected wave segment can be corrected to the target prosody. This also means that the prosodic sub-costs become comparatively not important if TD-PSOLA is used. Thus, it is expected that the spectral discontinuity between adjacent speech wave segments also can be decreased by increasing the weight of sub-costs about spectral continuity, or by decreasing the weight of sub-costs about prosodic features.

2 CORPUS SPECIFICATION

This section is described about the spontaneous speech corpus which we constructed to use for concatenative spontaneous speech synthesis.

2.1 Recording conditions

We employed two female professional narrators to collect their conversation. To record each conversation voice without mixing of the other's voice, the soundproof chamber and recording equipments settings are used (depicted in ***Fig. 6***). The chamber has divided into two sub-rooms by soundproof wall with window. Each narrator put on the headphone and the lavalier microphone while spontaneous speech recording. Their voices were recorded by the digital recorder separately and sent to the other's headphone through the mixer. The recording equipments are shown in ***Table 1***.

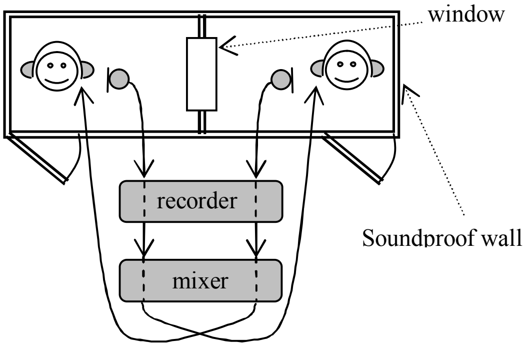

Fig.6: *Soundproof chamber for conversation recording*

Table 1: *Recording equipments and conditions*

Microphones	SONY ECM-77B
Recorder	marantz PMD670
Mixer	MACKIE 1202-VLZ Pro
Sampling frequency	48 kHz
Quantizing bit	16 bit

We recorded for three days and collected the spontaneous speech wave of about 140 minutes. From these data set, in this time, one speaker's voice was picked up and made into the corpus of about 100 minutes (including pause time) describing in the following subsection.

2.2 Corpus construction

The procedure that the recorded data is made into the corpus is the following.

2.2.1 Speech wave division

The recorded data was divided by sentences (it is difficult to define the sentence. One segment placed between silences is considered to be the sentence in this paper.).

As the result of this operation, 3004 speech files were obtained.

2.2.2 Utterance text dictation

The pronunciation of utterance of each speech wave file is dictated to text file with "hiragana (Japanese syllabary)". Conversation speech contains not only accurate pronunciation but also imprecise articulation, laugh, filler, disfluency and so on. Although these phenomena have an important role for composing natural communication, it is difficult to control them by a computer. A method having the possibility to solve this problem is to introduce some label set (e.g. speech act labels) [7] and develop a new unit selection technique which uses them. However in this paper, these label set were not used and all of utterances were described with "hiragana" phonologically as much as possible to examine the performance of the conventional speech unit selection.

2.2.3 Phoneme segmentation

Each speech wave is furthermore divided into phoneme segments. This operation was done by the Julian-segmentation-kit included in the Julius: an open-source continuous speech recognition software based on Hidden Markov Model [8]. Speech wave file, phoneme utterance text and acoustic model are necessary for this kit performing. Speech file must be down-sampled to 16 kHz. Phoneme utterance text is made from above-mentioned "hiragana" utterance text by the Perl script. Finally, the acoustic model for female speakers is used.

2.2.4 Feature extraction

Unit selection of concatenative speech synthesis needs some features (F_0, Power and MFCC) to estimate the distortion degree between speech wave units. F_0 and power are extracted by the free software snack sound toolkit [9]. This toolkit has two methods of F_0 extraction. We selected the one that is equal to the get_f0 command included in the toolkit ESPS/waves+ by Entropic Inc. Speech waves were pre-emphasized ($\alpha = 0.97$) for power extraction. MFCC is extracted by the tool (called wav2mfcc) provided by the Julius.

3 EVALUATION

To examine the quality of spontaneous synthetic speech, author has a subjective evaluation experiment. Ten sentences were prepared at random. Four subjects listened synthetic speech with a headphone in an office (not in a soundproof chamber), and told what they said. Two types of answers were requested to subjects: (a) the answer when they listened to a synthetic speech one times and (b) the answer when they listened to it again several times (the number of listening times are unrestricted). Sentence intelligibility is calculated as follows:

$$SI = \frac{100}{N_s}\sum_{i=1}^{N_s}\frac{CP_i}{P_i}\ [\%], \qquad (4)$$

where N_s, P_i and CP_i are the number of subjects ($N_s = 4$ in this paper), the number of phonemes of synthetic sentences for i-th subject, and the number of correct phonemes included in the answer sentences of i-th subject respectively.

The result is shown in ***Table 2***. The result of synthetic speech by only the unit selection (post processing has not used) and the result of natural speech are also shown for comparison. Although the result of synthetic speech with using TD-PSOLA is lower than the result of natural speech, it is higher than the conventional. That is, TD-PSOLA improved the intelligibility of the conventional synthetic speech (+4% for type (a) and +5.8% for type (b)).

Table 2: *Sentence intelligibility (number of listening times are (a) once and (b) unrestricted)*

Kind of speech	Sentence intelligibility [%]	
	(a)	(b)
Unit selection	62.9	77.4
Unit selection + TD-PSOLA	66.9	83.2
Natural speech	92.6	94.2

On the other hand, deteriorated synthetic speech by the proposal method was observed. It is thought that big modification to the prosody is a one of causes of the quality degradation. There are future works as follows: (a) employment of the sub-cost function corresponding to degree of prosodic modification, (b) comparison with other speech waveform modification method.

4 CONCLUSION

In this paper, TD-PSOLA is introduced as post processing of spontaneous concatenative speech synthesis to obtain the synthetic speech decreasing spectral and prosodic discontinuity which causes degradation of intelligibility. From the subjective evaluation experiment, intelligibility of synthetic speech has been improved +4% when once listening and +5.8% when listening time is unrestricted. As future work, (a) employment of the sub-cost function considering about degradation caused by prosodic modification, (b) comparison with other speech waveform modification method.

Acknowledgements

The author is grateful to the Applied Linguistics Laboratory in Nara Institute of Science and Technology for the conversation recording cooperation.

References

[1] The CREST/ESP Project, http://feast.his.atr.jp/

[2] K. Maekawa, "Corpus of Spontaneous Japanese: Its Design and Evaluation," Proceeding of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003), 2003.

[3] T. Akagawa, K. Iwano and S. Furui, "Toward realization of HMM-based spontaneous speech synthesis," (in Japanese) Technical Report of the Institute of Electronics, Information and Communication Engineers (SP2005-16), pp.25-30, 2005.

[4] K. Fujii, R. Ueda, H. Kashioka and N. Campbell, "A trial to apply concatenative speech synthesis to spontaneous speech," Proceeding of International Technical Conference on Circuits/Systems, Computers and Communications, Vol.2, pp.653-656, 2006.

[5] 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.

[6] E. Moulines and F. Charpentier, "Pitch-synchronous Waveform Processing Techniques for Text-to-speech Synthesis Using Diphones," Speech Communication, No.9, pp.453-467, 1990.

[7] N. Campbell, "What Type of Inputs Will We Need for Expressive Speech Synthesis?," Proceeding of IEEE 2002 Workshop on Speech Synthesis, 2002.

[8] http://julius.sourceforge.jp/

[9] http://www.speech.kth.se/snack/