Perceptual Evaluation of Quality Deterioration Owing to Prosody Modification

Kazuki Adachi*, Tomoki Toda†, Hiromichi Kawanami*, Hiroshi Saruwatari* and Kiyohiro Shikano*

*Nara Institute of Science and Technology
8916-5, Takayama-cho, Ikoma-shi, Nara-ken, Japan
{kazuki-a, kawanami, sawatari, shikano}@is.aist-nara.ac.jp

†Nagoya Institute of Technology, JAPAN / Carnegie Mellon University, U.S.A.
tomoki@ics.nitech.ac.jp

Abstract

Our reasearch goal is to construct a Japanese TTS (Text-to-Speech) system that can output various kinds of prosody. Since such synthetic speech is useful for a practical use, many TTS systems have implemented global prosodic control processing. But fundamentally they're designed to output speech with standard pitch and speech rate. We discuss synthesis method for high quality speech with extreme prosody (very high, low, fast and slow) from a viewpoint of a speech database. As a speech synthesis method, we employ a unit selection-concatenation method. We also introduce an analysis-synthesis process to give precise target prosody to output speech. Many research has reported that speech quality get worse in proportion to an amount of prosody modification by analysis-synthesis or PSOLA. Following the reports, we take an approach to reduce prosody modification of a speech segment. Nine Japanese speech databases with different characteristics in prosody are prepared. First we confirm relationship between speech quality deterioration and prosody modification, using synthetic speech with through objective and subjective tests. We also investigate relationship between a speech deterioration tendency and each speech database. The result indicates that the tendencies depend on prosodic features of original speech.

1. Introduction

From a practical TTS (text-to-speech) system or a expressive spoken dialogue system, a speech synthesizer that can generate speech with various kinds of prosody is requested. Many already developed TTS systems have implemented global prosodic control, fundamentally they're designed to output speech with standard pitch and speech rate.

In this decade, a unit selection synthesis method using large speech corpus (corpus-based speech synthesis) has been attracted for naturalness of output speech (Black et al.,1995). Because its output speech quality depends on its corpus size and corpus character, some unit selection approaches introduce prosody or/and spectrum modification process. We also employ a unit selection-concatenation method and analysis-synthesis process to give precise target prosody a user want, to output speech.

However, many research has reported that speech quality get worse in proportion to an amount of prosody modification by analysis-synthesis or PSOLA (Pitch Synchronous OverLap Add). Following the reports, we take an approach that aims to reduce prosody modification of a speech segment. We discuss synthesis method for high quality speech with extreme prosody (very high, low, fast and slow) by constructing Japanese speech databases with various kinds of prosody.

In section 2, database designing and recording process and their prosodic characteristics are described. After we confirm relationship between speech quality and prosody modification, using synthetic speech with various kinds of prosody in section 3, we also investigate relationship between a speech deterioration tendency and a speech database. And we conclude this paper.

2. Speech Database

2.1. Database designing

Speech quality deterioration is occurred in proportion to prosody modification by analysis-synthesis or PSOLA method. Our approach is not to modify spectral parameters according to prosody modification but to prepare speech database to reduce prosody modification.

Based on the idea, we record nine phonetically-balanced Japanese speech databases. Each set has same texts and different prosody (Masuda et al.,2004, Kawanami et al.,2002). Target values have three variations of F_0: F_0 in natural reading speech for a speaker (normal), 0.4 [octave] higher (high) and 0.4 [octave] lower (low). In the same way, three target values for duration: normal duration (normal), 0.5 [octave] shorter (fast), 0.5 [octave] longer (slow). Combination of these 2-dimensional feature, we records nine databases. Each database consists of 525 Japanese sentences, which are based on the phonetically-balanced 503 sentence set designed by ATR.

2.2. Database recording

Two female professional narrators (speaker FME, FOR) are asked to speak the sentences in nine prosodic variations. Recording procedure is as follows.

1. 525 sentences are recorded. Speakers are asked to speak in their natural reading style.
2. Using STRAIGHT method (Kawahara, 1997), analysis-synthesized speech is generated in nine kinds of target prosody, from the original speech database.
3. Nine databases are recorded. Synthesized speech with target prosody are presented before each utterance. Speakers ware asked to refer general features of the prosody.

Table 1: Speech Databases

high-fast_db	high_db	high-slow_db
F_0 : +0.425 oct.	F_0 : +0.405 oct.	F_0 : +0.413 oct.
Dur. : −0.449 oct.	Dur. : +0.063 oct.	Dur. : +0.354 oct.
MelCD : 6.829 dB	MelCD : 6.630 dB	MelCD : 6.758 dB
fast_db	**normal_db**	**slow_db**
F_0 : +0.013 oct.	F_0 : 0.000 oct.	F_0 : +0.042 oct.
Dur. : −0.432 oct.	Dur. : 0.000 oct.	Dur. : +0.427 oct.
MelCD : 5.539 dB	MelCD : —— dB	MelCD : 5.371 dB
low-fast_db	**low_db**	**low-slow_db**
F_0 : −0.264 oct.	F_0 : −0.294 oct.	F_0 : −0.293 oct.
Dur. : −0.458 oct.	Dur. : −0.060 oct.	Dur. : +0.370 oct.
MelCD : 6.480 dB	MelCD : 5.985 dB	MelCD : 5.896 dB

(Dur. = Duration , oct. = octave)

A database with normal prosody is also recorded again in the same way to avoid voice quality difference.

2.3. General features

General features of obtained databases (normal_db, high_db, low_db, fast_db, slow_db, high-fast_db, high-slow_db, low-fast_db, low-slow_db) are illustrated in 1. Average distance from normal_db on sentence mean F_0 (F_0), sentence duration (Dur.), mel cepstum distortion (MelCD) are described. MelCD is calculated from 40th order mel cepstral coefficients, extracted from STRAIGHT smoothed spectrum.

3. Modification Ratio and Speech Quality

In this section, we investigate that the recorded speech databases reduce prosody modification for TTS output speech and that reducing prosody modification leads to perceptual speech quality as expected.

In our previous report (Masuda et al.,2004, Kawanami et al.,2002), evaluation on durational modification is already conducted. In this paper, we extent evaluation databases according to F_0, namely, high_db, low_db in addition to fast_db, slow_db. Normal_db is also used for comparison.

3.1. Speech Synthesis

Phoneme sequence and target prosody (F_0 and phoneme duration) are given to the synthesizer. As target prosody, natural prosody is used, speech which prosody is used for target excluded from a synthesis database. 472 sentences is used for synthesis.

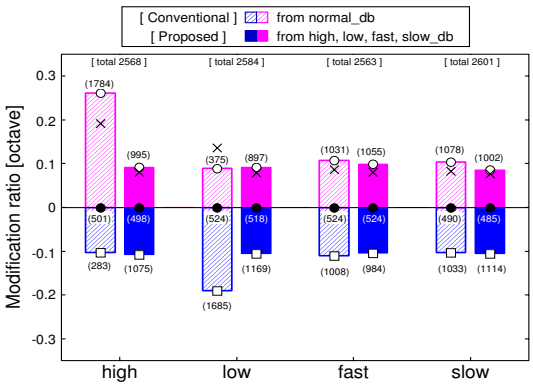
As an unit selection method, we employ unit selection algorithm developed by Toda, et al (2002). Prososodic features (F_0, phoneme duration, power) of selected speech segments are controlled by STRAIGHT method under condition shown in Table 2.

3.2. Objective evaluation

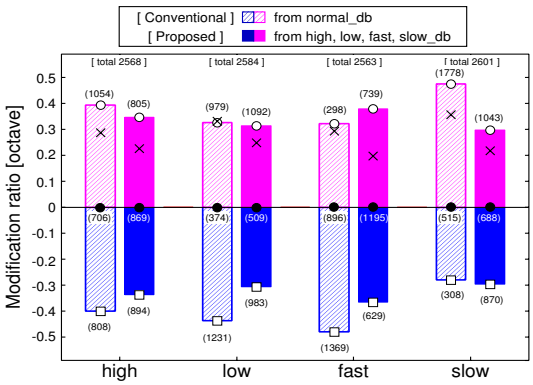
Prosody modification ratio is defined as difference between prosody of a selected speech segment and corresponding target prosody and used for objective evaluation of proposed databases.

Table 2: Conditions of STRAIGHT analysis

Sampling frequency	16 kHz
Analysis window	Gaussian window
Frame shift length	5 ms
Number of FFT analysis	1024



(a) Modification ratio (F_0)



(b) Modification ratio (duration)

Figure 1: Modification ratio of selected speech segments
"○" denotes average of positive modification, "□" denotes average of negative modification,
"●" denotes no modification, "×" denotes average absolute value of modification ratio for all segments. The number of segments are shown in parentheses.

Comparison is done with high F_0 synthetic speech from normal_db (Conventional method) and that from high_db (Proposed method). In the same way, comparisons of normal_db and low_db, normal_db and fast_db, normal_db and slow_db are carried out. 53 sentences are used for evaluation for each combination.

The experimental results are shown in Fig. 3.2.. Results of F_0 and duration modification ratio is illustrated separately. "○" denotes average of positive modification (heightening F_0 or lengthening duration), "□" denotes average of negative modification (lowering F_0 or shortening duration), "●" denotes no modification, "×" denotes average absolute value of modification for all segments. The number of segments are shown in parentheses. It is indicated that reducing prosody modification ratio is realized as expected.

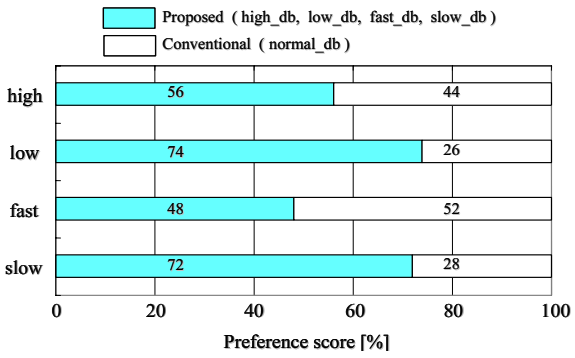


Figure 2: Results of subjective evaluation using TTS output speech

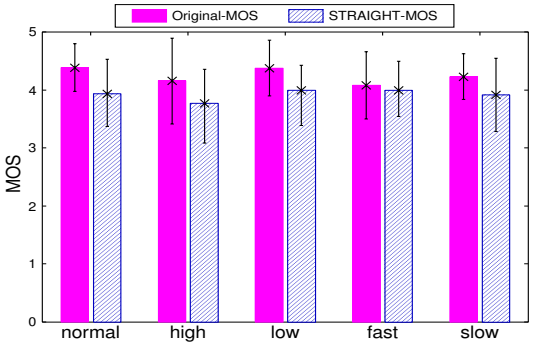


Figure 3: STRAIGHT-MOS

3.3. Subjective evaluation

To investigate relationship between prosody modification ratio and synthetic speech quality, a perceptual test is conducted. 10 adult listeners are asked to choose which of the two synthetic speech has higher quality. 15 sentences are used for evaluation. Examinees are allowed to listen to samples any number of times, The experimental result is shown in Fig. 2.

We can see that no advantage to use fast_db for fast speech synthesis, although reducing absolute value of duration modification is realized as expected (Fig. 1).

4. Detail Analysis

From the result of the perceptual evaluation in 3.3, it is supposed that quality deterioration owing to prosody modification relates to original prosodic features. Kawai et al. (2000) report permissible range of prosodic modification by PSOLA, using word utterances. The define that a "permissible range" is a 2-dimensional space that can obtain more than score four by MOS score for naturalness in average. Their report says the permissible range is $-0.2\sim+0.2$ octave in F_0 and $-0.5\sim+0.1$ octave in duration.

Following the report, we expand the research area using our five databases. Although they reports a permissible range around normal reading speech, we investigate relationship between a permissible range and original prosodic characters. Another difference is a modification method: we use STRAIGHT analysis- synthesis method.

4.1. Speech stimuli

Speech stimuli are not generated from output of a unit selection TTS system but four original speech from each database. Prosody modification is done by STRAIGHT using natural speech by $-0.5\sim+0.5$ octave both in F_0 and duration, in 0.1 octave step each. In this report, F_0 modification and duration modification is not executed simultaneously. To evaluate quality degradation through STRAIGHT analysis-synthesis process, natural speech that is not executed signal processing is add to speech stimuli,

4.2. Perceptual evaluation

Perceptual tests are conducted using MOS(Mean Opinion Score). 8 adult examinees are asked to listened to speech stimuli from headphones and evaluate naturalness from 1(Bad) to 5(Excellent). Speech stimuli are listed randomly and they're allowed to listen to stimuli any number of times. Every stimuli is appeard twice in a test.

Experimental results are shown in Fig.3 and 4. Fig.3 illustrates speech quality degradation owing to only analysis-synthesis process. Comparison of MOS's of natural speech (Original-MOS) and STRAIGHT analysis-synthesized speech (STRAIGHT-MOS) can be seen. The upper number sequence of the horizontal axis in Fig.4 denotes modification rate from original speech and the lower line denotes difference from normal prosody.

4.3. Discussion for perceptual test result

We discuss relationship between speech quality and prosody modification using STRAIGHT-MOS(analysis-synthesized speech by STRAIGHT without prosody modification) as a standard. We define modification permissible range provisionally from the perceptual results.

(1) exclude those are more than 0.5 score inferiour to STRAIGHT-MOS.

(2) exclude those have significant difference between STRAIGHT-MOS, by t-test(significance level 5 %).

Table 3 shows prosody modification permissible range decided by the above difinition.

On stimuli from normal_db, quality deterioration is observed in proportion to F_0 modification generally. On durational feature, a permissible range is more wider for shortening. The same durational tendencies are also observed in high_db and low_db. Comparing these results to those of Kawai et al.(2000), on F_0 is almost same and is slightly sensitive on duration.

On high_db, heightening F_0 remarkably lose quality. Original high_db speech itself is recorded considerably unnatural speech both on F_0 generation and vocal tract control, it is supposed that it is cause not only by miscorrespondence between spectral feature and prosodic feature but by unnaturalness as human pitch range, and so on. Same tendency is observed on the result of low_db.

On fast_db and slow_db, they're characterize deterioration tendencies with durational modification. As a general tendency, it is observed that speech lengthening is sensitive for deterioration than shortening. Based on this tendency, we can see fast_db have wider range for lengthening and slow_db have much wider for shortening. They are also

Table 3: Permissible Range in Each Database

normal	F_0	$-0.2\sim+0.2$	Duration	$-0.3\sim+0.2$
high	F_0	$-0.2\sim\ 0.0$	Duration	$-0.3\sim+0.2$
low	F_0	$-0.1\sim+0.4$	Duration	$-0.3\sim+0.2$
fast	F_0	$-0.2\sim+0.3$	Duration	$-0.2\sim+0.4$
slow	F_0	$-0.1\sim+0.3$	Duration	$-0.5\sim+0.3$

(octave)

supposed to be influenced by unnnaturalness of speech rate as human speech.

If we integrated that of five databases, the permissible range can be expanded within $-0.4\sim+0.4$ octave in F_0, and $-0.6\sim+0.7$ octave in duration but more detailed analysis (e.g. glottal source quality) should be done for database integration.

5. Conclusion

In this paper, we discuss improvement of synthesis speech which have extreme prosodic features. Our approach is to reduce prosody modification in analysis-synthesis process by preparing speech databases with various kinds of prosody. We focused on five speech databases (high, low, fast, slow and normal (conventional)) and confirm average modification ratio is reduced in each database's correponfing prosodic area. Perceptual results also indicate their effect except the fast speech database.

We also investigate relationship between perceptual quality and modification ratio using speech from each database. Using its MOS score results, we define a permissible modification range for each database. The permissible range can be expanded within $-0.4\sim+0.4$ octave in F_0, and $-0.6\sim+0.7$ octave in duration if we integrated that of five databases, but more detailed analysis (e.g. glottal source waveform) is required for database integration. We will also apply the MOS scores to unit selection algorithm.

Acknowledgement

This work was partly supported by JST/CREST in Japan

6. References

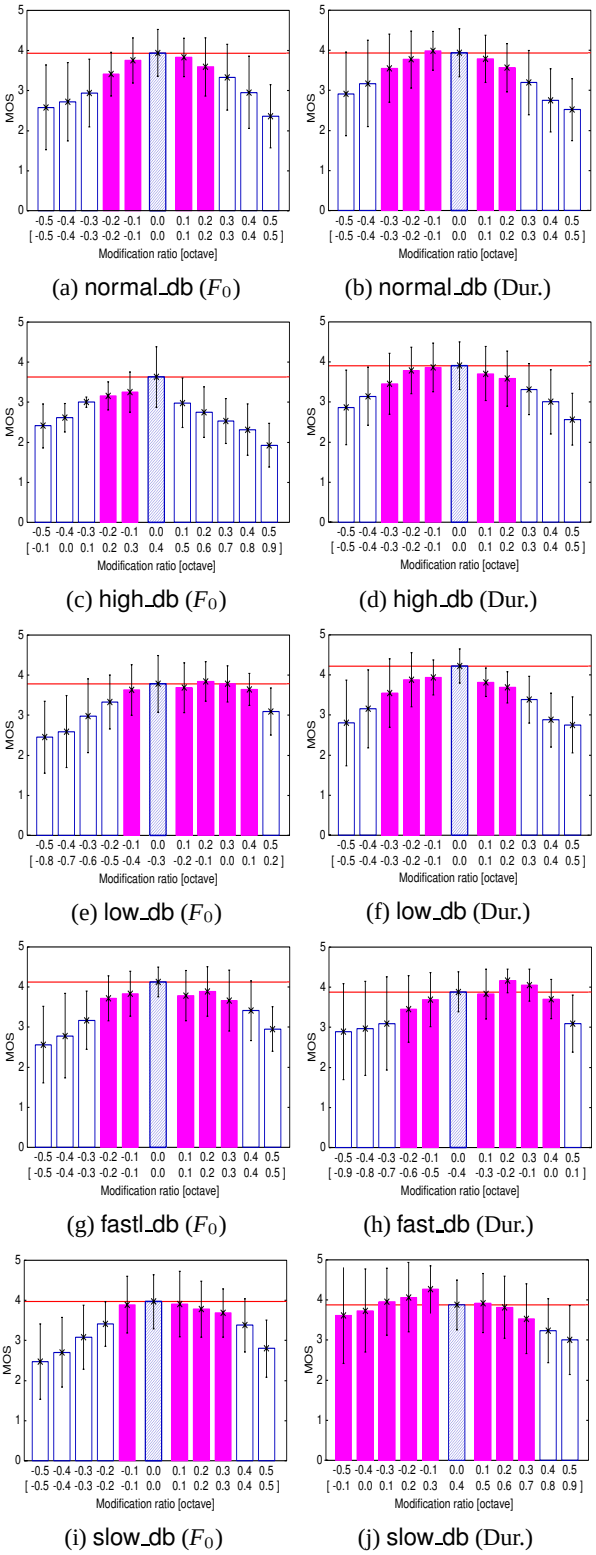
Black, A. and Campbell, N., 1955. Optimising selection of units from speech database for concatenative synthesis, *Proc. EUROSPEECH*, pp. 581–584.

Kawahara, H., 1997. Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited, *Proc. ICASSP, pp. 1303-1306.*

Masuda, T. et al., 2004. Speech Databases with Various Prosody and Its Evaluation on Speech Rate, Journal of IEICE D–II, vol. J87–D–II, no. 2, pp. 447–455. (in Japanese)

Kawanami, H. et al., 2002. Designing Japanese Speech Database Covering Wide Range in Prosody for Hybrid Speech Synthesizer, Proc. ICSLP, pp. 2425–2428.

Kawai, H., Yamamoto, S., Higuchi, N. and Shimizu T., 2000. Design Method of Speech Corpus for Text-to-Speech Synthesis Taking Account of Prosody, Proc. ICSLP, pp. 420–425



(a) normal_db (F_0) (b) normal_db (Dur.)

(c) high_db (F_0) (d) high_db (Dur.)

(e) low_db (F_0) (f) low_db (Dur.)

(g) fastl_db (F_0) (h) fast_db (Dur.)

(i) slow_db (F_0) (j) slow_db (Dur.)

Figure 4: Perceptual result

STRAIGHT-MOS, Permissible range

Upper line : Modification ratio

Lower line : Relative value compared to normal speech

Toda, T. et al., 2003. A Segment Selection Algorithm for Japanese Concatenative Speech Synthesis Based on Both Phonetic Unit and Diphone Unit, Journal of IEICE D–II, vol. J85–D–II, no. 12, pp. 1760–1770. (in Japanese)