Designing Speech Database with Prosodic Variety for Expressive TTS system

Hiromichi Kawanami*, Tsuyoshi Masuda†, Tomoki Toda* and Kiyohiro Shikano*

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

†Currently working at Asahi Kasei Cooperation

Abstract

For the purpose of building speech synthesis system that can generate high-quality speech with wide range in prosody and realize fine prosody control, we propose new speech database constructing method. As a speech synthesis method, we select a hybrid system which consists of two part : speech unit selection and prosody modification part by STRAIGHT (vocoder type high quality analysis-synthesis method). Our viewpoint for designing database is to reduce amount of prosody modification. which causes quality deterioration. Hence, to make it possible to generate arbitrary prosody within permissible range of prosody modification, we designed 9 sub-databases those consist of same phonetic balanced text set with different prosody. In this paper, we report the designing method and general features of obtained databases. Listening tests focused on durational fearure were also conducted. The results show effectiveness of the method and the necessity to change unit selection cost according to speech rate.

1. Introduction

In recent years, speech synthesis technology has progressed remarkably because of large scale speech corpus. But most of temporal synthesis systems are designed to generate high quality speech which reproducing original speaker's normal reading style speech. That is to say, monotonous, normal pitch range and normal speech rate. Next generation synthesizer will be expected to output expressive speech like human. Here, expressive speech means that it carries, in addition to verbal information, discourse or para- and extra-linguistic informations, for example speaker's attitude, intent, speaking style and individuality. As information variation expands, the acoustic features of the speech also spread. Particularly, difference in prosodic features are observed explicitly. Hence, we mention to build synthesis system that can output flexible prosody and that can control prosody to a given target value strictly.

To realize such system without losing perceptual quality, two synthesis approaches are conceivable. One is a waveform concatenation synthesizer with very large speech database that includes various informations. Another is to conduct acoustic feature modification after waveform unit selecting. However, the both method have problems caused by lack of a proper unit from speech database. On the first method, it cannot output objective prosody when database does not include proper units. Therefore, a huge size database should be required to solve. On the latter case, large prosody modification from a natural speech unit causes quality degradation. To avoid this problems, we focused on database designing.

The research by Kawai et al. (2000), which take account perceptual capability into measurement for degradation by prosody modification. Referring the result, we designed and recorded 9 phonetic-balance text sets with different prosody by two female narrators.

In the following sections, we describes the database designing and recording method and the general features of databases in sections 2 and 3, respectively. The result of evaluation tests that investigate effectiveness of the databases as Text-to-Speech (TTS) database and proper cost function for them are described in section 4. There, the tests are operated about the variaion of speech rate. It is followed by Discussions, then we conclude this paper.

2. Databases with Prosody Variation

2.1. Designing

As we mentioned above, the problem with prosody modification is speech quality degradation, that correlate with modification rate.

Kawai et al. (2000) investigate relationship between prosody modification using PSOLA and quality degradation using word utterance In which they deal with F_0 and duration of word utterance. It describes that modification is acceptable within –0.2 [octave] to +0.2 [octave] and from –0.5[octave] to +0.1[octave], for F_0 and for duration modification, respectively. Here, acceptable range is defined that obtained 4 point or greater by MOS score (5 is maximum) by listening test. Although we refer these values, we use vocoder-type analysis-synthesis method STRAIGHT (Kawahara et al., 1999a) for modification. Because, we our preliminary experiment (Masuda et al., 2001) shows resynthesized speech by STRAIGHT is perceived better than that by PSOLA.

In the previous analysis for emotional speech in Japanese (Hirose et al., 2000), F_0 is observed that F_0 changes within 1 [octave] in male normal reading utterance and it expands to about 2 [octave] in emotional(angry) speech. In respect of duration, human speech can be presented in arbitrary speech rate unless it lose its intelligibility. By considering human mechanism of speech production and practical use of synthetic speech, as to this study we aim to generate speech rate that have more than 1 [octave] range.

Based on the studies above, we designed database that consists of 9 phonetic balanced sub-databases. Each set

high-fast	**high**	**high-slow**
F_0: +0.4	F_0: +0.4	F_0: +0.4
dur.: –0.5	dur.: 0.0	dur.: +0.5
fast	**normal**	**slow**
F_0: 0,0		F_0: 0.0
dur.: –0.5		dur.: +0.5
low-fast	**low**	**low-slow**
F_0: –0.4	F_0: –0.4	F_0: –0.4
dur.: –0.5	dur.: 0.0	dur.: +0.5

Table 1: Target relative prosodic features from normal database (octave).

high-fast	**high**	**high-slow**
F_0: +0.425	F_0: +0.405	F_0: +0.413
dur.: –0.449	dur.: +0.063	dur.: +0.354
fast	**normal**	**slow**
F_0: +0.013		F_0: +0.042
dur.: –0.432		dur.: +0.427
low-fast	**low**	**low-slow**
F_0: –0.264	F_0: –0.294	F_0: –0.293
dur.: –0.458	dur.: –0.050	dur.: +0.370

Table 2: Relative prosodic features from normal database (octave).)

has same texts and different prosody. 3 variations of F_0 are, normal F_0 (F_0 in natural reading speech for a speaker), 0.4 [octave] higher F_0 than that of corresponding normal speech, 0.4 [octave] lower than normal speech. In the same way, 3 variations for duration. Namely, normal duration, 0.5 [octave] shorter than normal, 0.5 [octave] longer than normal. By integrating these sub-databases, prosodic range of output speech with acceptable quality expands ideally, 0.8 [octave] at F_0, and 1.0 [octave] at duration. The name of each sub-database and their target value of prosodic features are shown in Table 1.

For each sub-database, we use 525 sentences set, which include ATR phonetic balanced sentence set consists of 503 Japanese sentences (Abe et al., 2000) and additional 22 sentences to compensate foreign phonemes.

2.2. Recording

Two female professional narrators (speaker FME, speaker FOR) were asked to utter the sub-database in 9 prosodic variations. They were recorded in a soundproof room in digital format 16 bit, 48 kHz.

Recording procedure is as follows.

1. 525 sentences were recorded without special instruction for prosody. Speakers were asked to speak in their natural reading style. We call this utterance set, reference database.

2. 9 re-synthesized speech sets in different prosody were generated from reference database. using STRAIGHT method. To realize target prosody described in Table 1,

3. 9 sets of sub-database were recorded. Re-synthesized speech from speaker's own voice was presented before each utterance from a loud speaker. Speakers ware asked to refer general features of the prosody.

   The utterances for normal database was also recorded to avoid difference of voice quality from other databases. At this time re-synthesis speech without prosody modification was also demonstrated before each utterance.

3. General Features of Sub-databases

In this and the following sections, we focused on the utterances of speaker FME. General features of obtained sub-databases are described in this section.

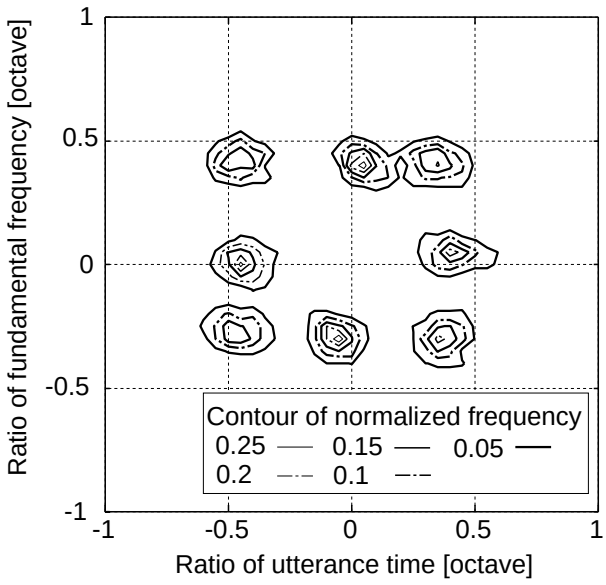

Figure 1: The contour map of normalized histogram.

3.1. Parameter Extraction

All speech samples were first down-sampled to 16k [Hz]. STRAIGHT-TEMPO (Kawahara et al., 1999b) was used to extract F_0 by 1 [msec] frame shift. For each sentence utterance, mean log F_0 was calculated and it was compared with that of corresponding speech in normal database. About duration, forced phoneme alignment was done by 5 [msec] frame shift using HMM(monophone, gender-dependent model(female)). Automatic pause detection and labeling were also processed. For each sentence, total duration was calculated first as sum of phoneme duration except pauses. Then each sentence duration was also compared with that of normal database.

3.2. General Features

The results for the average values are shown in Table 2. Figure 3. illustrate the contour map of normalized histogram. Three sub-databases with low F_0 are observed that they do not have enough difference with normal database (about 0.1 [octave] higher than expected). They are considered due to constraint of utterance ability for the lowest F_0. The results of duration show that all values does not reach to target values. These can be considered that speech rate control are done also by deletion and insertion of pauses.

However, it can be considered that each database has expected prosodic features generally.

4. Evaluation

To evaluate the database, two kinds of listening tests were conducted focused on speech rate using three sub-databases: **normal**, **fast** and **slow**. The tests are held as comparison of two synthesized speech generated from TTS system with different database or cost function.

On first experiment, we verify the advantage of fast and slow databases by comparison with normal database on generating fast and slow prosody speech. Next, integration of database is investigated from the viewpoints of pre-selection and cost function of database and cost function about duration.

4.1. Speech Synthesis

In the following, synthesis speeches for evaluation tests are generated from TTS system, that is non-uniform unit selection and subsequent prosody modification by STRAIGHT.

For each sub-database, 53 sentences (J-set from ATR database) are used for evaluation. The rest 472 sentences are used for TTS database, which phoneme labels and F_0 are calculated automatically. To avoid quality deterioration owing to mis-estimation of prosody by TTS, natural prosody that is extracted from evaluation set is given as a target prosody. The target F_0 are corrected manually.

4.2. Comparison with Normal Database

On first experiment, we confirmed advantage of fast and slow databases by comparison with normal database. To investigate effect of speech rate, two tests are done separately. That is to say, a comparison test between synthetic speech from fast database and that from normal database in fast and a test between slow database and normal database in slow speech rate.

AB tests were conducted to 10 adult listeners. A set of 20 sentence pairs was presented in 16k [Hz] PCM files in a Personal Computer IBM ThinkPad A21 with a headphone. Listers were allowed to playback speech files any number of times by operating icons in a display.

Figure 2 shows the results. It is observed the effectiveness of slow database for slow speech rate, clearly.

4.3. Effectiveness of Pre-Selection

AB listening test that shows effect of pre-selection of database is performed. The experimental conditions are almost same but 15 sets of utterance are used. One speech is generated from database that corresponds with target speech rate. The other is made from integration database of the aforementioned three. As well as the preceding test, target prosody was extracted from natural utterance, but the normal database is also used.

The results are illustrated in Figure 3. Selection rate from integrated database is also illustrated in Figure 4. Pre-selecting make effort for prosody from slow database, in spite that data amount for searching becomes narrow to one thirds. But it becomes bad influences It is considered that cost function about duration can be optimized in proportion to output speech rate.

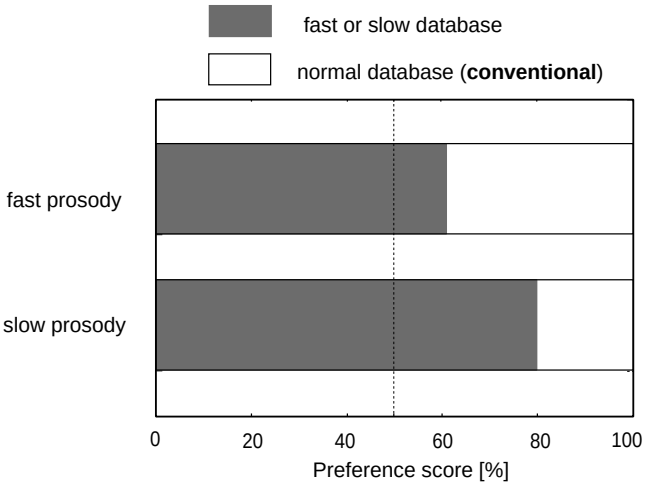

Figure 2: Comparison by speech rate of databases.

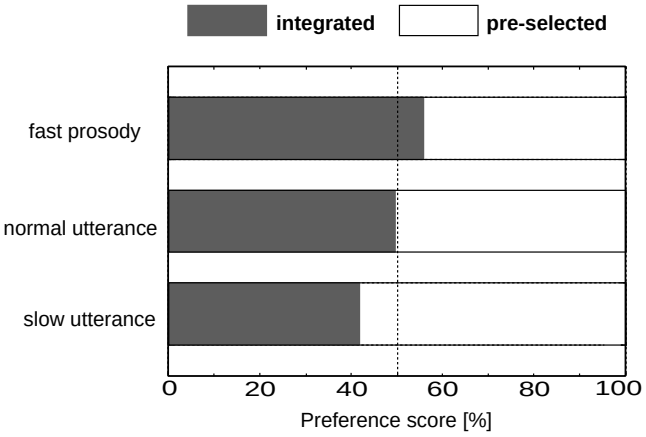

Figure 3: Comparison by speech unit selecting method.

5. Future Work

In the first evaluation test, effectiveness of proposed databases on duration is showed. And the second listening test leaves a subject to be solved in our future work.

That is, consideration for cost function transformation for output speech rate. The work is continuing not only for at speech rate of fast, normal and slow database, but also for arbitrary speech rate. Evaluation for other six databases is also will be conducted. Generally F_0 changes go with changes of segmental features, therefore more detailed analysis will be needed.

6. Conclusion

In this paper, we proposed speech database designing method that enable to cover wide enough range of prosodic features. The database consists of nine sub-databases of phonetic balanced sentences. Analysis of practically recorded databases shows that they generally have objective prosody.

Listening tests using TTS system focused on speech rate were also conducted. The results shows that effectiveness of fast and slow database to synthesize various speech rate

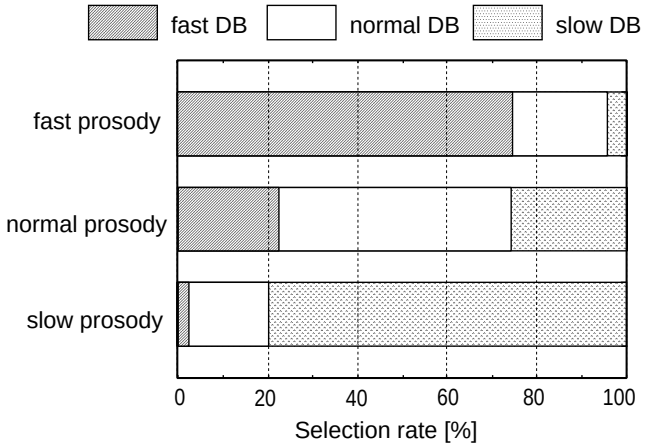

Figure 4: Selection rate of each database.

and that necessity for transforming cost function about duration in unit selection algorithm.

Acknowledgement

This work was partly supported by JST/CREST in Japan.

7. References

M. Abe, Y. Sagisaka, T. Umeda and H. Kuwabara. 1990. Speech Database User's Manual. *ATR Technical Report*, TR-I-0116. (in Japanese)

K. Hirose, N. Minematsu and H. Kawanami. 2000. Analytical And Perceptual Study on the Role of Asoustic Features in Realizing Emotional Speech. *Proc. ICSLP*, Vol. 2:369–372.

H. Kawahara, I. Masuda-Katsuse and A. de Cheveigné. 1999. 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, no. 3–4:187–207.

H. Kawahara, H. Katayose, A. de Cheveigné and R. D. Patterson. 1999. Fixed Point Analysis of Frequency to Instantaneous Frequency Mapping for Accurate Estimation of F0 and Periodicity. *Proc. EUROSPEECH*, Vol. 6:2781–2784.

H. Kawai, S. Yamamoto, N. Higuchi and T. Shimizu. 2000. A Design Method of Speech Corpus for Text-to-Speech Synthesis Taking Account of Prosody. *Proc. ICSLP*, Vol. 3:420–425.

T. Masuda, T. Toda, H. Kawanami, H. Saruwatari and K. Shikano. 2001. STRAIGHT-based Prosody Modification of CHATR Output. *Proc. Acoustic Society of Japan*, Autumn, Vol. 2:245–246. (in Japanese)