DESIGNING JAPANESE SPEECH DATABASE COVERING WIDE RANGE IN PROSODY FOR HYBRID SPEECH SYNTHESIZER

Hiromichi KAWANAMI, Tsuyoshi MASUDA, Tomoki TODA and Kiyohiro SHIKANO*

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

ABSTRACT

For the purpose of building Text-to-Speech (TTS) system that can generate high-quality and wide range speech in prosody, we conducted speech database construction. As a speech synthesizer, we use a hybrid system which consists of a unit selection module and prosody modification by STRAIGHT (vocoder type high quality analysis-synthesis method). Our viewpoint is to reduce an amount of prosody modification which causes quality deterioration. In other words, it is to generate any prosody at will within permissible prosody modification. Based on the aspect, we designed 9 sub-databases those consist of same phonetic balanced texts with different prosody. In this paper, we describe the designing policy and general features of the obtained database and the results of listening tests focused on the effectiveness about durational feature. They shows the advantage of the proposed database. and but it is also observed the necessity to change unit selection cost function according to output speech rate.

1. INTRODUCTION

By realization of practical use of large scale speech corpus, speech synthesis technology has progressed remarkably in this decade. But most of temporal synthesis systems are designed to generate high quality speech in standard normal reading style. That is to say, monotonous, normal pitch range and normal speech rate. One subject for next generation TTS system is to output expressive speech which carries, in addition to verbal informations, para- and/or extra-linguistic informations – for example speaker's attitude, intent, emotion, speaking style, individuality and so on. However, as variations of information expand, acoustic features of the speech also come to have various values. Particularly, difference in prosodic features are observed explicitly.

To realize such synthesis system without losing speech quality, we apply hybrid synthesizer which consists of a non-uniformal waveform unit selection module and an acoustic feature modification module. But large amount of prosody modification from a natural speech unit causes quality degradation. To avoid this problem, we focus on database designing. The research by Kawai et al.[1] take account perceptual capability into measurement for degradation by prosody modification. Referring the result, we designed and recorded 9 phonetic-balanced utterance sets with different prosody by two female narrators[2].

In this paper, we report results of two listening tests which indicate effectiveness of proposed database from the viewpoint of speech rate. And based on the results, we discuss the further application of our databases.

In the following sections, database designing and recording method and the general features of databases are described in sections 2 and 3, respectively. The result of evaluation tests which investigate effectiveness of the databases as Text-to-Speech (TTS) database are described in section 4. After that we discuss about more proper unit selection method in the following section, then we conclude this paper.

2. DATABASES WITH PROSODY VARIATION

2.1. Detabase Designing

As we mentioned above, the problem with prosody modification is speech quality degradation, that correlate with modification rate. Kawai et al.[1] investigate relationship between prosody modification using PSOLA and quality degradation using word utterances. 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, respectively. (Acceptable range is defined that it gets 4 point or greater by MOS score in average (5 is maximum) by listening test.)

Although we refer these values, we use vocoder-type analysis-synthesis method called STRAIGHT [3] for prosody modification. Because, we our preliminary experiment shows re-synthesized speech by STRAIGHT is perceived better

*Currently working at Asahi Kasei Corporation than that by PSOLA [4] .

And the previous analysis for emotional speech in Japanese reports [5], it 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 the respect of duration, human speech can be generated in speech rate at will unless it lose its intelligibility. So we considered human mechanism constraint for speech production and practical use of synthetic speech, in this study we decided to operate about 1 [octave] range in speech rate.

Based on the studies above, we designed database that consists of 9 phonetic balanced sub-databases. Each set has same texts and different prosody. 3 variations of F_0 are defined. They 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 are defined. 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 names of each sub-database and their target prosodic values are illustrated in Table 1.

For each sub-database, we use 525 sentences set, which include ATR phonetic balanced sentence set consists of 503 Japanese sentences [6] and additional 22 sentences to compensate foreign phonemes.

2.2. Detabase Recording

Two female professional narrators (speaker FME, speaker FOR) were asked to utter the 525 sentence text set 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 styles. We call this utterance set, the **reference** database.

2. 9 re-synthesized speech sets in target prosody were generated from the **reference** database using STRAIGHT method.

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 the **normal** database were also recorded to avoid voice quality differences with other sub-databases. At this time re-synthesized speech from the **reference** database without prosody modification was also presented before each utterance.

high-fast F0: +0.425 (+0.4) dur.: –0.449 (–0.5)	**high** F0: +0.405 (+0.4) dur.: +0.063 (0.0)	**high-slow** F0: +0.413 (+0.4) dur.: +0.354 (+0.5)
fast F0: +0.013 (+0.0) dur.: –0.432 (–0.5)	**normal**	**slow** F0: +0.042 (+0.0) dur.: +0.427 (+0.5)
low-fast F0: –0.264 (–0.4) dur.: –0.458 (–0.5)	**low** F0: –0.294 (–0.4) dur.: –0.050 (0.0)	**low-slow** F0: –0.293 (–0.4) dur.: +0.370 (+0.5)

Table 1. Prosodic features of obtained sub-databases from speaker FME. (Numbers in parentheses are the target values.)

3. DATABASE ANALYSIS

3.1. Parameter Extraction

All speech samples were first down-sampled to 16k [Hz]. STRAIGHT-TEMPO method [7] was used to extract F_0 by 1 [msec] frame shift. For each sentence utterance, mean log F_0 was calculated. Then that was compared with that of corresponding speech in the **normal** database.

About duration, forced phoneme alignment was done by 5 [msec] frame shift using HMM (3 states monophone, gender-dependent model) for each sentence. Automatic pause detection was also executed. Total duration was calculated first as a sum of phoneme durations except pauses. Then each sentence duration was also compared to correponding one from the **normal** database.

Hereafter, we describe about the speech databases from the speaker, FME.

3.2. General Features

The average values for each sub-database are shown in Table 1. The three sub-databases with low F_0 are observed that they do not have enough distance from the **normal** database (about 0.1 [octave] higher than expected). They are considered constraint of the speaker's utterance ability for lowest F_0. About duration, it is shown that all values do not reach to target values. These can be assumed because speech rate control are done also by deletion and insertion of pauses. Although the tendencies above are observed, we considered that each database has prosodic features practically.

4. EVALUATION

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

On the first experiment, we verify the advantage of the **fast** and **slow** databases by comparison with the **normal** database at generating fast and slow speech. On the latter test, the effectiveness of database integration is examined.

4.1. Speech Synthesis

In the following, synthesis speeches for evaluation tests are generated from TTS system, which consists of a non-uniform unit selection module 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 and F_0 labels are calculated automatically. To avoid quality deterioration owing to mis-estimation of prosodic features by TTS system, natural prosody extracted from evaluation set is given as a target prosody. The target F_0 and phoneme labels are corrected manually.

4.2. Comparison with Normal Database

On first experiment, we confirmed the advantage of the **fast** and the **slow** databases by comparison with the **normal** database. AB preference tests were conducted by 10 adult listeners. The 20 pairs of sentences were presented as 16k [Hz] PCM files from 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 1 shows the results. It is observed the effectiveness of the **fast** and **slow** databases. Especially in the case of the **slow** for slow speech rate, it is clearly observed.

4.3. Effectiveness of Database Integration

Based on the result of the previous evaluation, another AB listening test that compares integration of the three sub-databases and pre-selection of sub-database is performed. The experimental conditions are almost the same but 15 sets of utterance are presented. One speech is generated from database that corresponds with target speech rate, the other one is made from the integrated database of the three. As well as the preceding test, each target prosody was extracted from the natural utterance. In this test, the case of normal speech rate is also examined.

The results are illustrated in Figure 2. In spite of decreasing of data amount for searching, pre-selecting is effective for slow prosody. This fact suggests that a proper unit selection method is different according to output speech rate.

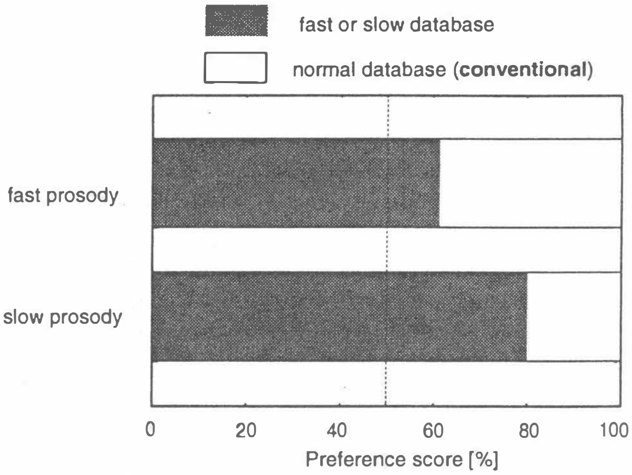

Fig. 1. Comparison with the fast/slow database and the normal database.

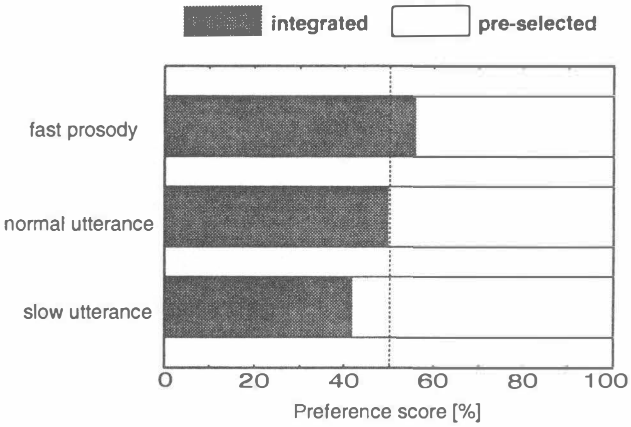

Fig. 2. Comparison with database integration and pre-selection.

5. DISCUSSION

In this section, we disscuss the method that make use of the proposed database properly.

The problem that brought up in the previous section is the necessity to change method of unit selecting in proportion to output speech rate. And we must also consider that the output speech is required not only in this three speech rate but any speech rate practically. In this meaning, pre-selecting of database is not the best solution for our goal. To construct proper cost function for the integrated database is needed.

As a preliminary experiment for that, we performed an additional AB hearing test that comapared deterioration by shortening of duration and that by extending. For each pair of synthetic speech, one was generated by extending or shortening from a utterance in the **normal** database, another was generated from the corresponding utterance in the **fast** or **slow** databases. They were given various speech rates from –0.5 [octave] to +0.5 [octave] by 0.1 [octave] step. Speech shortening and extending is executed as follows; first, for each speech rate, sentence speech durations and each phoneme durations are caluculated by linear complement from the phoneme labels extracted by the natural utterances in two speech rates (normal and fast, or slow). Then, pairs of synthetic speeches were generated by durational modification by STRAIGHT method. Here calculated phoneme durations are strictly refered.

AB preference tests were conducted to 10 adult listeners using four different sentences. The experimental conditions were almost the same with those of the previous section, but sample speech pairs were presented from DAT player randomely.

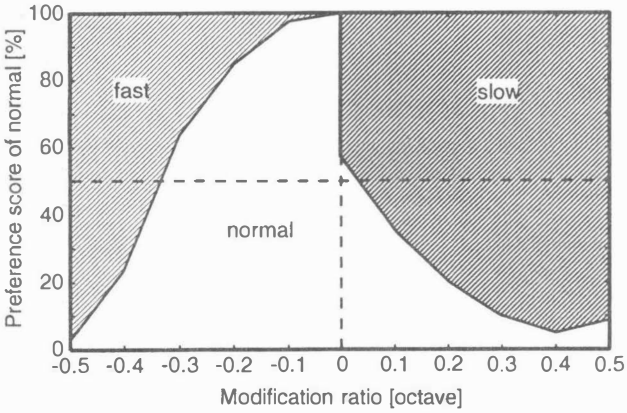

Fig. 3. Selection rate of each database.

The result is illustrated in Figure 3. Here we should notice that the natural utterances are existing with three speech rates, they're about –0.43 [octave] (as fast), 0 [octave] (as normal) and +0.43 [octave] (as slow) as indicated in Table 1.

From the result, the tendency is observed that shortening natural speech keeps high quality than extending. And the tendency is emphasized at slow speech rate.

As our future work, we plan to construct cost function for unit selecting process for any speech rate by perceptual tests. The cost function is supposed to be transformed corresponding to the rate. And we're also going to conduct database evaluations from the viewpoints of F_0 features.

6. CONCLUSION

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

Listening tests using outputs from hybrid TTS system focused on speech rate were also reported and the results show the effectiveness of the **fast** and the **slow** databases to synthesize corresponding speech outputs. As the furter discussion, another experiment was conducted for more proper use of the database. The result shows the necessity of cost function transformation for output speech rate.

Acknowledgments

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

7. REFERENCES

[1] H. Kawai, S. Yamamoto, N. Higuchi and T. Shimizu, "A Design Method of Speech Corpus for Text-to-Speech Synthesis Taking Account of Prosody,"*Proc. ICSLP, Vol. 3,* pp.420-425 (2000)

[2] H. Kawanami, T. Masuda, T. Toda and K. Shikano, "Designing Speech Database with Prosodic Variety for Expressive TTS system," *Proc. LREC*(2002)

[3] Kawanara, et al., "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,* pp. 187-207 (1999)

[4] T. Masuda, T. Toda, H. Kawanami, H. Saruwatari and K. Shikano, "STRAIGHT-based Prosody Modification of CHATR Output," *Proc. ASJ meeting, Autumn, Vol. 2,* pp.245–246 (2001-10) (in Japanese)

[5] K. Hirose, N. Minematsu and H. Kawanami, "Analytical And Perceptual Study on the Role of Asoustic Features in Realizing Emotional Speech," *Proc. ICSLP, Vol. 2,* pp. 369-372 (2000)

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

[7] Kawahara, et al., "Fixed Point Analysis of Frequency to Instantaneous Frequency Mapping for Accurate Estimation of F0 and Periodicity," *Proc EUROSPEECH, Vol. 6,* pp. 2781-2784 (1999)