



Speech Database Reduction Method for Corpus-Based TTS System

Mitsuaki ISOGAI and Hideyuki MIZUNO

NTT Cyber Space Laboratories, NTT Corporation, Japan

{isogai.mitsuaki, mizuno.hideyuki}@lab.ntt.co.jp

Abstract

We propose a new speech database reduction method that can create efficient speech databases for concatenation-type corpus-based TTS systems. Our aim is to create small speech databases that can yield the highest quality speech output possible. The main points of proposed method are as follows; (1) It has a 2-stage algorithm to reduce speech database size. (2) Consideration of the real speech elements needed allows us to select the most suitable subset of a full-size database; this yields scalable downsized speech databases. A listening test shows that proposed method can reduced a database from 13 hours to 10 hours with no degradation in output quality. Furthermore, synthesized speech using database sizes of 8 and 6 hours keeps relatively high MOS of more than 3.5; 95% of MOS using full size database.

Index Terms: text to speech, speech database design, speech database reduction

1. Introduction

In recent years, a lot of text-to-speech (TTS) systems based on the corpus-based concatenation approach have been developed [1-5]. The speech database is a key component of all corpus-based TTS systems. The large speech database allows a corpus-based TTS system to yield highly natural sounding speech. On the other hand, because large databases are expensive to implement, the speech database should be as small as possible. That is to say, reducing the size of speech database, while keeping the synthesized speech quality, is the key issue for practical applications.

To synthesize high-quality speech, the database must contain a wide variety of speech parts: words, syllables, and phonemes. Several methods to generate the scripts used by the narrator have been reported [6-10]. For example, one approach of exchanging sentence pairs based on the entropy of diphones and triphones[6] was proposed, another approach that maximizes the synthesis unit coverage by taking account of prosody[8] was studied. These methods extract the sentences with estimated phoneme chains and prosody from a large text corpus. However, the estimation error inherent in text analysis or prosody estimation disrupts the sentence extraction process. This degradation directly leads to poor synthesis quality.

In this paper, we propose a new speech database reduction method that can generate small speech databases that yield the highest possible speech quality. We generate a small speech database by selectively choosing and using samples from a large speech database including hand labeled phonemes and F0 data. To raise the precision of sentence selection, we take account of correct phoneme labels and F0 data. The large database considered here was generated using our recording script generation method [11]. Therefore, it inherently contains a wide variety of speech parts. In [11], we indicated that taking account of the balance of acoustic speech

parts and linguistic ones would be effective in raising the coverage of speech elements. So we use these speech parts in the reduction method proposed here. In this paper, F0 variations are taken as the acoustic parts.

Section 2 defines the phonetic elements. Section 3 introduces our speech database reduction method. Section 4 details an analysis of the phonetic element coverage in speech database reduction experiments. Section 5 explains listening tests on the results of speech database reduction. Section 6 provides results of a task-specific reduction experiment. Our conclusion is given in Section 7.

2. Definitions

2.1. Phonetic element definitions

In this paper, we define four phonetic elements as shown in Table 1. Syllable(represented by SYL) is a basic element to ensure that any text can be synthesized. The three expansion elements provide high-quality synthesis. Of the three expansion elements, two are acoustic elements and the other is a linguistic one. The acoustic elements cover variations in short-time speech features, while the linguistic element provides long-period speech features such as words. The acoustic elements are “Syllable with several F0 variations” (CV) and “Syllable with several F0 variations and phoneme environment” (S). The linguistic element is “Morpheme with phoneme environment” (M).

Table 1. *Phonetic element definitions. In the form column, P means any phoneme, S means syllable, and M means morpheme. ‘[]’ means the enclosed unit has several F0 variations. ‘()’ means the enclosed phoneme is a phoneme environment. In the example column, /KURUMA/ (it means ‘car’) is a Japanese morpheme.*

| | phonetic element | symbol | form | example |
|--------------------|---|--------|-----------|---------------------|
| basic element | syllable | SYL | S | /NO/ |
| | syllable with F0 variations | CV | [S] | [/NO/] |
| expansion elements | syllable with F0 variations and phoneme environment | S | (P)[S](P) | (/O/)[/NO/](/T/) |
| | morpheme with phoneme environment | M | (P)M(P) | (/A/)/KURU-MA/(/G/) |

According to a preliminary examination we define F0 variation as follows. F0 variation consists of two dimensions. One is the quantized F0 average of an element. The F0 average is generated for a vowel section in the element quantized every 20 mels, from 60 mels to 600 mels. The other is the quantized F0 tilt of the element. We quantized the tilt to three types by coefficient $a[\text{mel}/\text{msec}]$ which is defined by the F0 regression line of the vowel section in the element. The first type is *raise* (if $a > 0$), the second one is *flat* (if $-0.3 < a \leq 0$) and the last one is *fall* (if $a \leq -0.3$).

2.2. Measurement definitions

We define the metric of coverage to select sentences that will constitute the reduced speech database from the source speech database. We define $Ec(X)$ as ‘Element Cover Rate(ECR)’ based on the total number of variations of the phonetic elements in the full-size (source) database. We define $Sc(X)$ as ‘Sentence Cover Rate(SCR)’ based on the frequencies of the phonetic elements in the source speech database. The definition of $Sc(X)$ is as follows:

$$Ec(X) = \frac{M(X)}{N(X)} \quad (1)$$

$$Sc(X) = \frac{\sum_{i=1}^{N(X)} n_i(X) d_i(R)}{\sum_{i=1}^{N(X)} n_i(X)} \quad (2)$$

where X is the phonetic element type such as SYL. $M(X)$ is the number of variations of X in the reduced database R . $N(X)$ is the number of variations of X in the source speech database D . $n_i(X)$ is the frequency of $u_i(X)$ in D . $\{u_1(X), \dots, u_i(X), \dots, u_{N(X)}(X)\}$ are phonetic elements included in D . Function $d_i(R)$ outputs 1 if $u_i(X) \in R$, otherwise it outputs 0.

3. Proposed Algorithm

In this section, we detail our speech database reduction method based on a greedy algorithm with 2-stage selection. It is based on our script generation method in [11]. The 1st stage handles the basic element. The 2nd stage handles the expansion elements.

The algorithm is described as follows:

1st stage. The scores of all sentences in the source speech database are calculated. This score, denoted by $s(\text{SYL})$, is defined as the increase in $Ec(\text{SYL})$ that would occur if the sentence were added to the reduced speech database. If only one sentence has the highest score, add this sentence to the reduced speech database, and iterate the 1st stage.

2nd stage. If there are multiple sentences with equally highest score in terms of $s(\text{SYL})$, calculate new score S using expansion elements to decide the most suitable of these sentences. This new sentence score T is calculated as follows:

$$T = w(\text{CV})s(\text{CV}) + w(\text{S})s(\text{S}) + w(\text{M})s(\text{M}) \quad (3)$$

where, $s(\text{CV})$ is defined as the increase in $Sc(\text{CV})$, that would occur if the sentence were added to the reduced speech database. $s(\text{S})$ and $s(\text{M})$ are defined in the same way. $w(\text{CV})$, $w(\text{S})$ and $w(\text{M})$ are weight coefficients of CV, S, and

M, respectively. If there is only one sentence with highest score, add this sentence to the reduced speech database. If there are multiple sentences with equally highest score, select the sentence with the shortest length. Return to the 1st stage.

This loop is iterated until the size of the reduced speech database reaches some application-specific limit.

4. Coverage Analysis

4.1. Conditions

This section examines our speech database reduction algorithm from the viewpoint of the difference in phonetic element coverage. Speech database reduction was carried out using four sets of weight coefficients. Moreover, random selection was employed as the baseline. The details of the four sets, sets(a)~(d), are as follows.

In weight coefficient sets(a)~(c), we set one of the weights to 1.0, the others to 0. For example, $w(\text{CV})=1.0$, $w(\text{S})=w(\text{M})=0$. In weight coefficient set(d), we set all weights to 1.0. That is $w(\text{CV})=w(\text{S})=w(\text{M})=1.0$. These sets were intended to examine which types of phonetic elements were added to the speech database due to the weight coefficients of phonetic elements.

The source speech database had about 13 hours of material. The reduced database size variations were 10, 8, 6, 4, 2, 1 and 0.5 hour(s). Table 2 shows the contents of the source speech database in this experiment.

Table 2. Contents of the source speech database.

| genre | | newspaper, newscast, novel, etc. |
|---|-----|----------------------------------|
| number of sentences | | 30811 |
| number of mora | | 308928 |
| number of elements in the source database | SYL | 324 |
| | CV | 5900 |
| | S | 90469 |
| | M | 95466 |

4.2. Results and discussion

Figures 1~3 show the relations between speech database size and SCR. Figure 1 shows that $Sc(\text{CV})$ offers high coverage even as the speech database size is reduced. $Sc(\text{CV})$ drops rapidly if the database size is less than 2 hours. However, it offers more than 90% coverage at any condition. Figure 2 shows that $Sc(\text{S})$ decrease more rapidly than $Sc(\text{CV})$. Furthermore, figure 3 shows that $Sc(\text{M})$ drop off more rapidly than $Sc(\text{S})$.

For set(a), Figure 1 shows that $Sc(\text{CV})$ offers higher coverage than the other sets and random selection. Figures 2 and 3 show that $Sc(\text{S})$ and $Sc(\text{M})$ of set(a) are inferior to the others. This indicates the following: If we set the weight of a certain phonetic element to 1.0, and the others to 0, the coverage of that certain element can be high. However, the other elements provide performance inferior to random selection. In the cases of set(b) and (c), $Sc(\text{S})$ and $Sc(\text{M})$ have similar tendencies, respectively. Meanwhile, SCR with set(d) shows that taking account of several expansion elements can cover a wider variety of phonetic elements on average.

Moreover, it has higher coverage than random selection regardless of database size. From the viewpoint of phonetic element coverage, this experiment shows that weight coefficient set(d) is better than the other coefficient sets and indeed random selection.

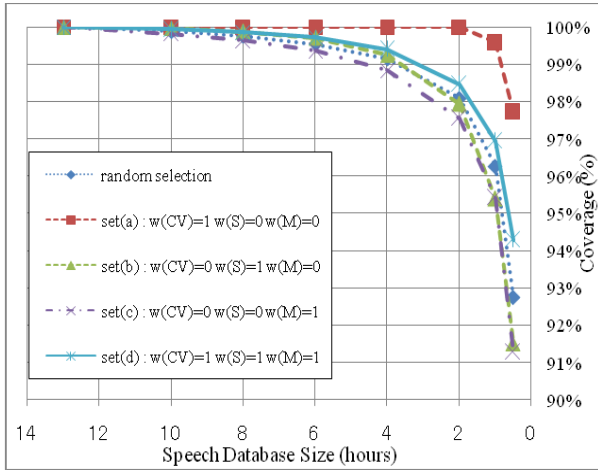


Figure 1: Coverage of speech element CV: $Sc(CV)$

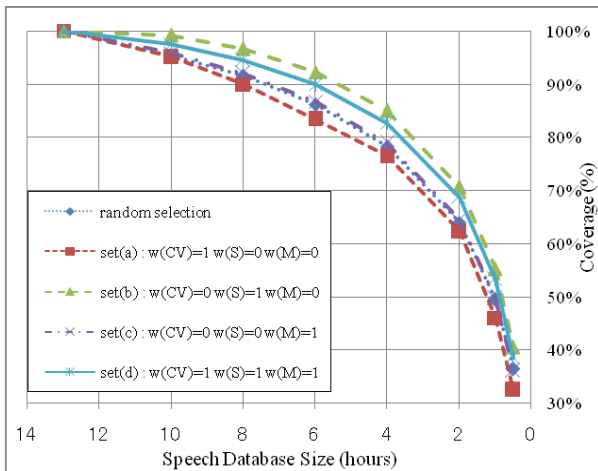


Figure 2: Coverage of speech element S: $Sc(S)$

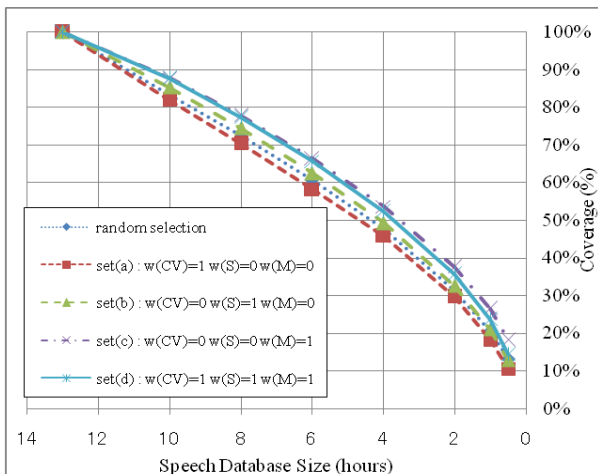


Figure 3: Coverage of speech element M: $Sc(M)$

5. Listening Tests

5.1. Conditions

Listening tests were carried out to examine the quality of the synthesized speech created from the reduced speech databases. 16 sentences were synthesized by our corpus-based TTS system[5] using the speech databases generated in 4.1: 13(source), 10, 8, 6, 4, 2, 1, and 0.5 hour(s). 32 subjects rated the naturalness of each synthesized speech using a five level (1~5) Mean Opinion Score (MOS).

5.2. Results and discussion

The results are shown in Figure 4. We verified the difference in MOS between the source speech database and generated speech databases according to t-test (5% significance level). For set(d), the results yielded by 13(source) and 10 hours showed no significant difference. The other comparisons showed significant differences. For set(d), database sizes of 8 and 6 hours yielded relatively high MOS, more than 3.5; 95% of MOS using the source speech database. It indicates that, for the database size of 10~6 hours, the reduced database with set(d) minimizes the reduction in speech quality.

In addition, we compared MOS of the speech synthesized using the reduced speech database, with sets(a)~(d), versus the one with random selection according to t-test (5% significance level). For the database size of 0.5 hour, the comparison of set(b) or (d) with random selection shows a significant difference. For the database size of 1 hour, the comparison of set(b), (c) or (d) with random selection shows a significant difference. It indicates that our method has an advantage over random selection in significantly reducing speech database size.

Table 3 shows the high overall correlation between MOS score and the coverage of the speech elements. Therefore, the any coverage criteria are useful in estimating the MOS of speech synthesized created using reduced speech databases.

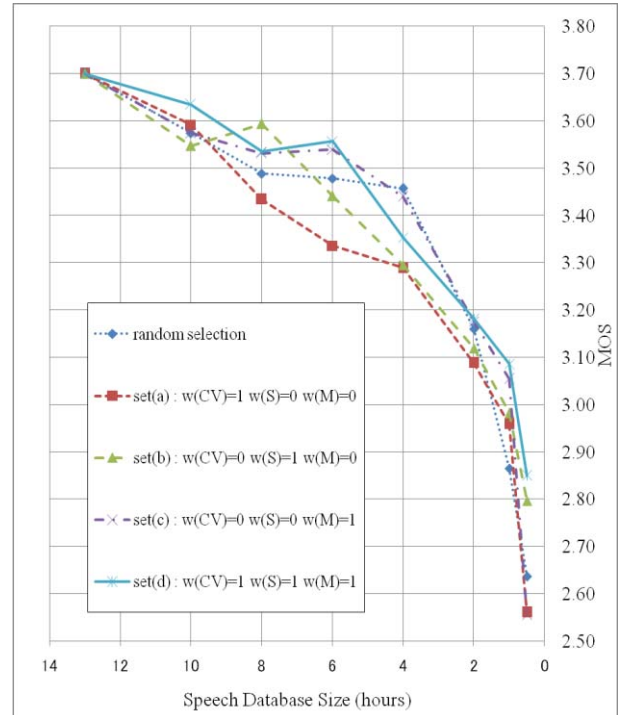


Figure 4: Listening test results: MOS ratings.

Table 3. Correlation coefficients between MOS and coverage.

| weight coefficient conditions | coverage criteria speech elements | | |
|-------------------------------|-----------------------------------|-------|-------|
| | Sc(CV) | Sc(S) | Sc(M) |
| random selection | 0.968 | 0.985 | 0.924 |
| set(a) w(CV)=1 w(S)=0 w(M)=0 | 0.857 | 0.981 | 0.946 |
| set(b) w(CV)=0 w(S)=1 w(M)=0 | 0.926 | 0.987 | 0.978 |
| set(c) w(CV)=0 w(S)=0 w(M)=1 | 0.991 | 0.967 | 0.900 |
| set(d) w(CV)=1 w(S)=1 w(M)=1 | 0.943 | 0.987 | 0.974 |

6. Task-Specific Reduction Experiment

6.1. Conditions

This section examines the effectiveness of task-specific databases. Two small speech databases (A and B) were generated from two large speech databases. Database-A was generated from a task-independent source database that covered various genres. Database-B was generated from a task-specific source database (a subset, 70%, of the task-independent source database) that covered the genre of newscasts and newspapers. Both small databases were created using weight coefficient set(d) from Sections 4 and 5, and both contained about 1 hour of speech.

We carried out two preference tests. One focused on newscast and newspaper tasks. 10 speech pairs of each task were synthesized by the TTS system of Section 5 using the two small speech databases. The other focused on mixed tasks (for example novels, information guidance, conversation, and so on) and again 10 speech pairs were synthesized in the same way. 5 subjects evaluated these speech pairs.

6.2. Results and discussion

The results are shown in Figure 5. We verified the difference of the preference score between database-A and B according to binomial test (5% significance level). For the newscast and newspaper tasks, no significant difference was observed between the synthesized speech using database-A and B according to the preference test. For the other task, speeches generated from database-A were preferred. The difference of the preference score was statistically significant. These results indicate that reduced speech databases should be generated from task-independent databases, regardless of the task of the TTS system.

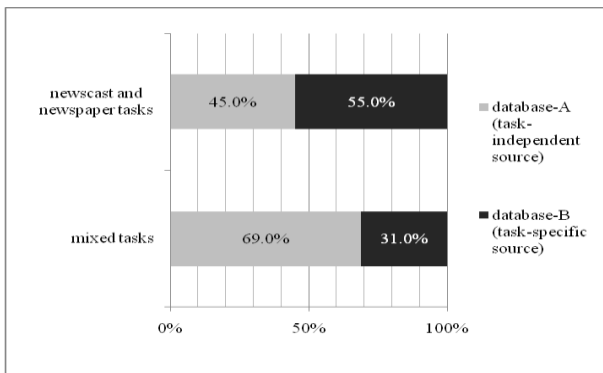


Figure 5: Results of preference tests.

7. Conclusion

This paper introduced a new speech database reduction method that can create efficient speech databases for concatenation-type corpus-based TTS systems. We proposed an algorithm that can create downsized speech databases by selecting subsets of the full-size source database. It identifies the appropriate speech elements by considering correct phoneme labels and F0 values of the source data. A listening test showed that proposed method can reduce a database from 13 hours to 10 hours with no degradation in output quality. A preference test showed that the source database should be task-independent, regardless of the task of the TTS system. In the future, we will examine the difference in opinion score between the database created by using a recording script and an equivalent one created by our reduction method.

8. Acknowledgements

The authors would like to thank the members of the Speech, Acoustics and Language Laboratory for several helpful discussions.

9. References

- [1] T. Hirokawa and K. Hakoda, "Segment Selection and Pitch Modification for High Quality Speech Synthesis using Waveform Segments," *Proc. of ICSLP'90*, pp.337-340, 1990.
- [2] N. Campbell, "CHATR: A High-Definition Speech Resequencing System," *Proc. of 3rd ASA/ASJ Joint Meeting*, pp.1223-1228, 1996.
- [3] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS System," *Proc. of 137th Meeting Acoustic Society America*, 1999.
- [4] H. Kawai, T. Toda J. Ni, M. Tsuzaki, and K. Tokuda, "XIMERA: A New TTS from ATR based on Corpus-based Technologies," *Proc. Of 5th ISCA Speech Synthesis Workshop*, pp.179-184, 2004
- [5] K. Mano, H. Mizuno, H. Nakajima, H. Asano, M. Isogai, M. Hasebe, and A. Yoshida, "Development of a corpus-based concatenative text-to-speech synthesis system 'Cralinet' for contact center services," *Proc. of Autumn Meeting of The Acoustical Society of Japan*, pp.347-348, 2004, (in Japanese).
- [6] K. Iso, T. Watanabe, and H. Kuwabara, "Design of a Japanese Sentence List for a Speech Database," *Proc. of Spring Meeting of The Acoustical Society of Japan*, pp.89-90, 1988, (in Japanese).
- [7] J. P. H. van Santen, "Methods for Optimal Text Selection," *Proc. of Eurospeech97*, pp.553-556, 1997.
- [8] H. Kawai, S. Yamamoto, and T. Shimizu, "A Design Methods of Speech Corpus for Text-to-Speech Synthesis Taking Account of Prosody," *Proc. of ICSLP 2000*, pp.420-425, 2000.
- [9] C. Kuo and J. Huang., "Efficient and Scalable Methods for Text Script Generation in Corpus-based TTS Design," *Proc. of ICSLP2002*, pp.121-124, 2002.
- [10] B. Bozkurt, O. Ozturk, and T. Dutoit, "Text Design for TTS Speech Corpus Building Using a Modified Greedy Selection," *Proc. of Eurospeech 2003*, pp.277-280, 2003.
- [11] M. Isogai, H. Mizuno, and K. Mano, "Recording Script Design for Corpus-Based TTS System Based on Coverage of Various Phonetic Elements," *Proc. of ICASSP2005*, pp.301-304, 2005