CONSTRUCTING A PHONETIC-RICH SPEECH CORPUS WHILE CONTROLLING TIME-DEPENDENT VOICE QUALITY VARIABILITY FOR ENGLISH SPEECH SYNTHESIS

Jinfu Ni[†], *Toshio Hirai*[†], *and Hisashi Kawai*[†‡]

[†]ATR Spoken Language Communication Research Laboratories, Japan
[‡]KDDI R&D Laboratories Inc., Japan

ABSTRACT

This paper presents a practical approach to constructing a large-scale speech corpus for corpus-based speech synthesis. This consists of (1) selecting a source text corpus that fits limited target domains; (2) analyzing the source text corpus to obtain the unit statistics; (3) automatically extracting prompt subjects (sentences) from the source text corpus to maximize the intended unit coverage with the given amount of text; and (4) recording prompt subjects while controlling such critical factors that cause undesirable voice variability. This paper describes related computational methods, such as a greedy algorithm for prompt selection, the proximity effects found in a real recording system, and a technique for detecting the time-dependent voice variations. While the approach is demonstrated in English, it is also promising for other languages.

1. INTRODUCTION

One of the main trends in constructing text-to-speech (TTS) systems is to apply corpus-based unit selection technology, as exemplified in [1]. As anyone who has built a unit selection synthesizer knows, the quality of synthetic speech is highly dependent on the unit coverage of a speech corpus [2]. Because it is time- and cost-consuming to construct a large-scale speech corpus, designing appropriate *prompt subjects* (sentences) is necessary for reducing the corpus size and maximizing the unit coverage of a spoken language, as described in [2], [3], and other works. On the other hand, recording such a speech corpus may still last from several weeks to months. This poses an important problem: how to avoid in the recording short-term (daily) and long-term (monthly) variations in voice quality [4]. This is important because concatenating speech segments with different voice qualities produces audible discontinuities that degrade the naturalness of synthetic speech. While a few techniques have been proposed to correct voice variability in a large speech corpus, as described in [5] and [6], our experiments in [7] indicated that active correction of the channel variability in a speech corpus, for example, would to some extent cause degradation of voice quality.

The rest of this paper is organized as follows. Section 2 describes a practical approach, focusing three main stages: designing prompt subjects to fit a few target domains; suppressing the critical factors related to time-dependent voice variability while recording the prompt subjects; and detecting voice quality variations in the recording. Section 3 presents experimental results, and Section 4 concludes this paper.

2. DESCRIPTION OF THE APPROACH

A recorded speech corpus needs to reflect the target domains, in particular, by being phonetically balanced. Achieving coverage is straightforward for limited domains but very difficult for others, since perfect quality open-domain synthesis is not yet possible [2]. Thus, designing prompt subjects involves the following stages:

- Select a source text corpus to fit the target domains.
- Analyze the source text corpus to obtain the unit statistics.
- Select appropriate prompt subjects from the source text.
- Inspect and remove unsuitable sentences.

2.1. Source text selection and analysis

• Determine target domains. There exist two common domains related to TTS applications: news-reading and conversation applied to spoken language communication systems. In the experiment described here, two text corpora are adopted: a Basic Travel Expression Corpus, henceforth referred to as BTEC, which was collected at ATR, and an English newspaper corpus, hereafter, NEWS. Both are assumed to meet the intended domains (simple conversation and news-reading).
• Use *Festival* [8] to analyze a source text corpus to obtain the statistics of basic units (monophones, diphones, triphones), POS (part of speech), and existing diphone and triphone types in the text corpus. This consists of a few steps: (1) Decomposing every paragraph in the text corpus into *utterances* (a predicted unit in [8] whose size extends from a phrase to a clause). (2) Grouping *utterances* into *sentences* determined simply by specific punctuation marks, such as ". ! ?"; thus a *sentence* may comprise one or more *utterances*. Note that there might exist wrongly phonetized words due to OOV (out-of-vocabulary) such as proper names and potential typos in

This research was supported in part by the National Institute of Information and Communications Technology.

1-4244-0469-X/06/$20.00 ©2006 IEEE

text. Also, we have considered including words with only a single pronunciation by excluding homographs, but the effect is very limited. Accordingly, there exist 34 POS tags and 40 phonemes plus an extra one /pau/ (pause/silence). A /pau/ is always assumed at the beginning and end of a *sentence*, and it is also used to separate any two adjacent *utterances*.

• Measure divergence in unit statistics between two sources by the symmetrizing Kullback-Leibler divergence:

$$KL(P(p_i), Q(q_i)) = \sum_{i=1}^{n} \left(\frac{p_i}{2} \times \ln \frac{p_i}{q_i} + \frac{q_i}{2} \times \ln \frac{q_i}{p_i} \right). \quad (1)$$

$P(p_i)$ and $Q(q_i)$, $i = 1, 2...n$, each represents the unit probability distribution of one of the two sources. A unit set may be POS (n = 34), monophones (n = 40), diphones (n = 1,680 (= 41× 41 -1 (/pau/-/pau/)), and triphones. The two sources may be either BTEC and NEWS or a source text corpus and prompt subjects extracted from the source text corpus.

2.2. Prompt subject generation

2.2.1. Coverage definition

To select a *sentence* set from a large text corpus, it is necessary to define the metric of coverage of the *sentence* set [3]. Let X have elements $\{\mu_1^x, \mu_2^x, ..., \mu_{n_x}^x\}$, where n_x is the number of elements. X indicates a unit type, such as diphone, triphone, or POS. That is, $X \in$ {diphone, triphone, POS}. Assume $p(\mu_i^x)$ to be the occurrence frequency of unit μ_i^x in a source text corpus. By definition, $\sum_{i=1}^{n_x} p(\mu_i^x) = 1$. Additionally, let S denote a *sentence* set selected from the text corpus. Accordingly, the coverage of S for X, denoted by C_S^X, is defined as $C_S^X = \sum_{i=1}^{n_x} p(\mu_i^x) \times \delta(\mu_i^x)$, where $\delta(\mu_i^x)$ = 1, if $\mu_i^x \in S$. Otherwise, $\delta(\mu_i^x)$ = 0.

2.2.2. A greedy algorithm

An algorithm for extracting a *sentence* set from a text corpus to maximize the unit coverage is described as follows.

Step 1: A score is calculated for each of the currently focused N^* *sentences*, where this focus runs through the entire text corpus in succession during the loop process. The score is defined as the increase in coverage that would occur if the *sentence* were added to *sentence* set S. S is none at initiation.

Step 2: The *sentence* that has the highest score among the N^* *sentences* is extracted according to the following priority, and it is added to S; the size of S increases by 1 at each loop.

(1) Maximizing C_S^{diphone} (or simply denoted by C_S^{di}).

(2) Maximizing C_S^{triphone} (or C_S^{tri}), if (1) is satisfied.

(3) Maximizing C_S^{POS}, if (2) is satisfied.

(4) Maximizing the number of triphone variants at specific positions: the beginning, end, and a few middle positions of *utterances*, if (3) is satisfied. This tactic is intended to consider necessary prosodic effects on the basic units

Step 3: Halt and output S if its size reaches a predefined value. Otherwise, return to Step 1.

If N^* *sentences* cover the entire source text corpus, S would reach a global optimum solution according to the defined criterion. This is simply because the *sentence* with the highest score among the rest of the text corpus could always be selected through the step-by-step procedure.

2.3. Prompt recording

2.3.1. Controlling time-dependent voice variations

We intend to suppress the time-dependent effects on voice variations described in [4] during the recording process rather than correcting them after recording. First, identical audio equipment is used throughout the recording, and only the volume of the microphone amplifier is adjusted. Of course, the layout of the recording studio is also kept the same.

Second, a critical factor is microphone setting, more specifically the distance between the microphone and the speaker's mouth. It is known that the low frequency responses of a microphone with a directive response pattern are boosted due to the proximity effect when a sound source is set close to the microphone. Figure 1 shows the frequency responses of a target microphone (Neumann Microphone, Type TLM 103, which has a cardioid directional pattern) set in a recording system with respect to those of a nominative microphone fixed at an impulse mouth; this measurement was reported elsewhere [9]. It is clear from this figure that, when the distances between the target microphone and the artificial mouth change from 5 cm to 60 cm in several steps, the boosting is found in the frequency range of 50-300 Hz. In order to suppress the proximity effect, the results suggest keeping the mouth-microphone distances as close to 30 cm as possible during the recording period. Consequently, the proximity effect can be limited to 3 dB as shown in Fig. 1.

Finally, a main factor causing voice variability is the physical and mental conditions of the speaker. While it is difficult to suppress this kind of effect, many efforts have been made to minimize the impact of this factor to the recording in voice quality. For example, choosing a speaker with a good ability to control his/her voice through a careful audition process; limiting the speech data collected in each recording day; dividing the recording day into several sessions (20-minute work and 20-minute break, alternately); and having the speaker listen to several nominative samples so that he/she can anchor a normal voice for each recording day.

2.3.2. Technique for detecting voice variations

The proposed technique for evaluating the recording is based on a measure of the minus log-likelihood of long-term power spectral densities (PSDs) in terms of one mixture Gaussian $N(\mu, \Sigma)$ modeling the acoustic space of time-dependent voice variations; μ and Σ indicate a q-dimensional mean and covariance matrix, respectively. Long-term PSDs have been found to be effective for detecting voice quality variability in large-scale speech corpora [5] [6]. The PSD of long-term voiced segments y_i, denoted by $P_{y_i}(k)$, can be expressed as follows.

$$P_{y_i}(k) = \frac{1}{\|w\|^2 M} \sum_{m=1}^{M} P_{y_i}^{(m)}(k), k = 0, ..., K-1, \quad (2)$$

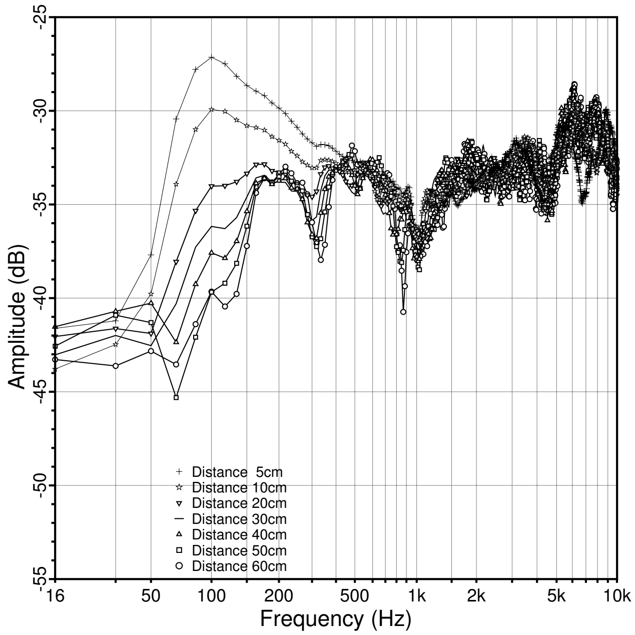

Fig. 1. Frequency responses of a target microphone with respect to those of a nominative microphone at different distances between the microphone and an impulse mouth.

$$P_{y_i}^{(m)}(k) = \frac{1}{K}\sum_{n=0}^{K-1}\left|w(n)y_i(n)e^{-\frac{j2\pi kn}{K}}\right|^2, \qquad (3)$$

where w is a hamming window, and M is the number of speech frames extracted from y_i. In practice, the judgment of whether the *m*th frame of y_i is voiced is simply done by comparing its energy $E_{y_i}^{(m)}$ with the maximum $E_{y_i}^{(max)}$:

$$E_{y_i}^{(m)} = \sum_{n=0}^{K-1}|w(n)y_i(n)|^2 \left\{\begin{array}{c}\text{voiced if } E_{y_i}^{(m)} \geq \delta E_{y_i}^{(max)}\\ \text{unvoiced, otherwise}\end{array}\right\}, \qquad (4)$$

where δ indicates a factor and K is FFT size.

A long-term PSD is then converted to a q-order MCEP (mel-frequency cepstrum) vector, denoted by x_i. Thus the minus log-likelihood of x_i to $N(\mu, \Sigma)$ can be expressed as

$$\ell(x_i) = \frac{\ln((2\pi)^q\text{det}(\Sigma)) + (x_i-\mu)'\Sigma^{-1}(x_i-\mu)}{2}. \qquad (5)$$

We use one mixture model, rather than multi-mixtures, in order to capture the divergence of the acoustic source of voice variability, thus revealing relative voice variations.

3. EXPERIMENTAL RESULTS

3.1. Prompt subjects

BTEC and NEWS are the text corpora for analyzing unit statistics and extracting prompt subjects. Table 1 lists the count of words, *sentences*, diphone and triphone types for each corpus. Note that a *sentence* was filtered out if its word number was more than 25 in BTEC and not between 10 and 25 in NEWS.

Table 1. Count of basic units in source text corpora.

Corpus	#Words (million)	#Sentences	#Diphone types	#Triphone types
BTEC	3.77	749.5 k	1,472	26,657
NEWS	22.02	4,985.2 k	1,597	40,499

Table 2. Divergence between BTEC and NEWS text corpora.

Divergence	POS	Monophone	Diphone	Triphone
KL(BTEC,NEWS)	0.229	0.030	0.153	0.489

Table 3. Unit coverage of prompt subjects.

	#Sent.	C_S^{di}	C_S^{tri}	C_S^{POS}	Type$_S^{\text{di}}$	Type$_S^{\text{tri}}$
BTEC	5,120	99.99%	99.72%	100%	99.93%	73.1%
NEWS	3,100	99.99%	99.33%	100%	92.84%	49.0%

Table 4. Results for criteria of guiding subject selection.

#Sent. chosen by	C_S^{diphone}	C_S^{triphone}	C_S^{POS}	Others
BTEC (5,120)	8.95%	90.6%	0.31%	0.08%
NEWS (3,100)	13.0%	86.5%	0.35%	0.16%

Table 2 shows the divergence calculated by Eq. (1) between BTEC and NEWS, while considering the distributions of POS, monophones, diphones, and triphones. It shows that there exist statistical differences in unit coverage between news-writing and conversational text corpora.

Two subject sets are independently extracted from the two text corpora using the greedy algorithm: 5,120 *sentences* (52 k words, average 10.5 words per *sentence*) from BTEC and 3,100 *sentences* (50 k words, average 16.3 words per *sentence*) from NEWS. Table 3 outlines the unit coverage of the subject sets. For example, the 3,100-*sentence* set involves 92.8% of existing diphone types (Type$_S^{\text{di}}$ in Table 3) and 49.0% of existing triphone types (Type$_S^{\text{tri}}$). In addition, 22.2% of *sentences* end with a question mark from BTEC and 2.2% from NEWS. The least phoneme is /zh/ (139 instances extracted from BTEC and 184 samples from NEWS).

Figure 2 shows the coverage as a function of set sizes in the number of *sentences* in the sets. The first 350 *sentences* from BTEC are selected with N^* taking the total number of *sentences* (i.e., 749.5 k) at each loop. This is, however, extremely time-consuming. Then N^* = 2,000 in the rest of this experiment. It can be seen that there are slightly different increase rates for the triphone coverage around point 350.

Table 4 illustrates how many *sentences* are selected by maximizing C_S^{diphone}, C_S^{triphone}, C_S^{POS}, and the other (triphone variants at specific positions). It is clear that the criteria for maximizing C_S^{diphone} and C_S^{triphone} play a critical role in governing the selection, given a small corpus size.

The prompt subjects are recorded in a sound-proofed room at ATR by an American male native, who won against other three natives in a well-designed audition. The recording period lasted more than one month, including 18 recording days.

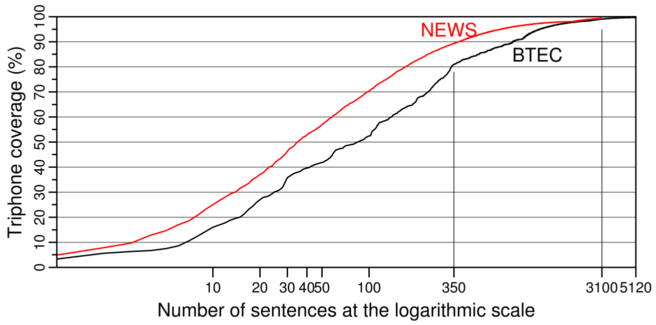

Fig. 2. Coverage increases with increasing subject set sizes.

3.2. Evaluating the recording

Our evaluation focused on time-dependent voice variations in the recording. Part of the speech corpus is provided for those researchers with Blizzard Challenge 2006 [10]. Therefore, an open evaluation on the appropriateness of the speech corpus for constructing TTS systems will be coming soon.

To train a robust model $N(\mu, \Sigma)$, we used the Japanese (male), Chinese (female) and English (male) speech corpora collected at ATR (around 200 recording days in total) to create the acoustic space of the long-term PSDs. More specifically, a long-term PSD was calculated from 3-minute voiced speech segments (48-kHz sampling) with a 10-ms frame rate; K = 1024 for FFT analysis; $\delta = 5 \times 10^{-6}$ for detecting voiced frames; q = 40; and $a = 0.5$ for MCEP conversion. All of the long-term PSDs were then used for estimating both model parameters μ and Σ to model the "complete" acoustic space.

Figure 3 shows the scattered likelihood values of long-term PSDs for the recording (crosses) and those for a Chinese speech corpus as a reference (circles). The lower the likelihood values, the better the performance. First, the measure in Eq. (5) is vital for revealing the potential voice variations. This is demonstrated by the clear evidence of existing time-dependent voice variations in the Chinese speech corpus, such as the channel effects analyzed elsewhere [7]. It is assumed that there are slight time-dependent voice variations in the recording. However, there are also isolated noticeable effects as indicated by such likelihood values that suddenly deviate from the slightly upward trend, for example, the crosses above the line – 45 on the y-axis. This finding may be related to the speaker's temporary physical condition, which was noted in the report on the field recording.

4. CONCLUSION

This paper presented a method for generating prompt subjects used in constructing an English speech corpus. Furthermore, it discussed the recording of prompt subjects while controlling undesirable voice variability. One of the criteria used in the prompt design is to exhibit good diphone and triphone coverage with the given amount of text. Experimental results indicate that the selection algorithm is effective in raising the coverage of all intended units. Also, the time-dependent voice quality variability in the recording is controllable by carefully setting such critical factors as the proximity effect of the microphone, the layout of the recording studio, and the speaker.

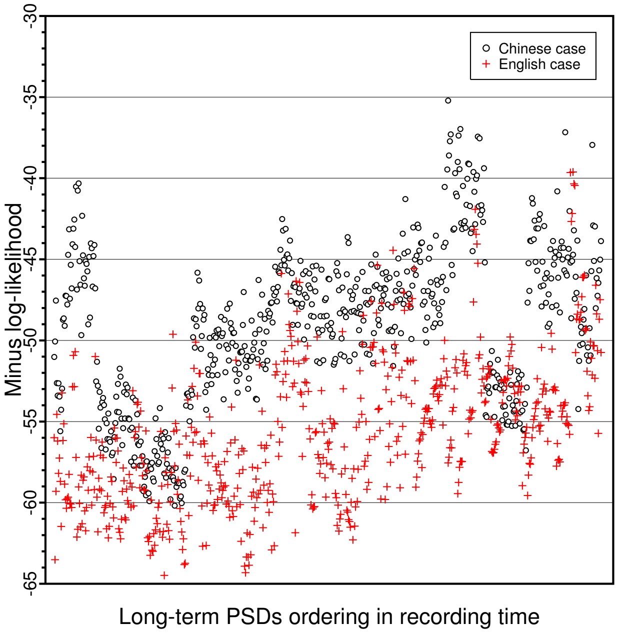

Fig. 3. Scattered likelihood values of long-term PSDs in the English recording (mean: – 55.63) and a Chinese speech corpus (mean: – 49.06) both using the same recording system.

ACKNOWLEDGMENT We would like to thank the anonymous reviewers for their helpful comments.

5. REFERENCES

[1] H. Kawai *et al.*, "XIMERA: a new TTS from ATR based on corpus-based technologies," in *Proc. the 5th ISCA Speech Synthesis Workshop*, pp. 179-184, 2004.

[2] J. Kominek and A.W. Black, "CMU ARCTIC database for speech synthesis," *Technical Report CMU-LTI-03-177*, 2003.

[3] M. Isogai *et al.*, "Recording script design for corpus-based TTS system based on coverage of various phonetic elements," in *Proc. ICASSP 2005*, vol. I, pp. 301-304.

[4] H. Kawai and M. Tsuzaki, "Voice quality variation in a long-term recording of a single speaker speech corpus," in *Text to Speech Synthesis: New Paradigms and Advances*, S. Narayana and A. Alwan (eds.), pp. 19-33, 2004.

[5] Y. Stylianou, "Assessment and correction of voice quality variabilities in large speech databases for concatenative speech synthesis," in *Proc. ICASSP 1999*, pp. 377-380.

[6] Y. Shi *et al.*, "Power spectral density based channel equalization of large speech database for concatenative TTS system," in *Proc. ICSLP 2002*, pp. 2369-2372.

[7] J. Ni, H. Kawai, and M. Tsuzaki, "Detection and correction of the channel variability in a Mandarin speech corpus," *Acoust. Sci. & Tech.* **25**, 4, pp. 303-306, 2004.

[8] http://festvox.org/festival/

[9] J. Ni, H. Kawai, and M. Tsuzaki, "Investigation of power spectral density based channel equalization," in *Technical Report of IEICE*, SP2003-67, pp. 19-24, 2003.

[10] http://www.festvox.org/blizzard/