ATRECSS — ATR ENGLISH SPEECH CORPUS FOR SPEECH SYNTHESIS

Jinfu Ni[1,2], *Toshio Hirai*[3,a)], *Hisashi Kawai*[2,4], *Tomoki Toda*[1,5], *Keiichi Tokuda*[1,6]
Minoru Tsuzaki[1,7], *Shinsuke Sakai*[1,2], *Ranniery Maia*[1,2], *and Satoshi Nakamura*[1,2]

[1]National Institute of Information and Communications Technology, Japan
[2]ATR Spoken Language Communication Labs, Japan
[3]Arcadia Inc., Japan
[4]KDDI Research and Development Labs, Japan
[5]Nara Institute of Science and Technology, Japan
[6]Nagoya Institute of Technology, Japan
[7]Kyoto City University of Arts, Japan

ABSTRACT

This paper introduces a large-scale phonetically-balanced English speech corpus developed at ATR for corpus-based speech synthesis. This corpus includes a 16-hour American English speech data spoken by a professional male narrator in "reading style." The contents of prompt sentences concern basically news articles, travel conversations, and novels. The prompt sentences were selected from huge collections of texts using a greedy algorithm to maximize the coverage of linguistic units, such as diphones and triphones. A few measures were taken to control undesirable recording variations in voice quality in the short term (daily) and long term (monthly) while recording the prompt sentences. Statistical figures of the corpus developed as well as those of subsets provided for Blizzard Challenge 2006 and 2007 are presented.

1. INTRODUCTION

ATR has been studying corpus-based speech synthesis techniques for more than a decade. These techniques have been used to develop three text-to-speech (TTS) systems, namely, ν-talk [1], CHATR [2] [3], and XIMERA [4]. The latter was utilized to output speech in the multilingual speech-to-speech translation system developed at ATR [5]. The quality of synthetic speech is highly affected by the coverage of the linguistic units in the target domain by the speech corpus used. Therefore, coverage and balance of the linguistic units have been very important aspects of the design of the speech corpora utilized for speech synthesizer development. The CMU ARCTIC database [6], for example, was developed considering these aspects. It consists of several speakers, each uttering approximately one hour of speech. Recently, a 16-hour phonetically-balanced American English speech corpus from a male speaker was built at ATR. A subset of this speech corpus has been released to the Blizzard Challenge 2006 and 2007 various as a common dataset for participants to build their synthetic voices [7]. This paper introduces this corpus, named ATRECSS — ATR English speech Corpus for Speech Synthesis. Because it is time-consuming and costly to construct a large-scale speech corpus, the design of appropriate prompt sentences is necessary for reducing the corpus size and maximizing the linguistic unit coverage of the target language. Also, recording such a speech corpus may still last from several weeks to months. This poses an important problem: how to avoid recording variations in voice quality [8] in the short-term (daily) and long-term (monthly). We presented a practical approach in [9] to dealing with the issues. This consists of (1) selecting a source text corpus that well represents the target domains; (2) analyzing the source text corpus to obtain the unit statistics; (3) automatically extracting prompt sentences from the source text corpus to maximize the intended unit coverage with a given amount of text; and (4) recording prompt sentences while controlling undesirable voice variability.

The rest of this paper is organized as follows. Section 2 outlines the selection of prompt sentences. Section 3 describes the prompt recording and the statistic results. Section 4 presents the contents of a subset that was provided for Blizzard Challenge 2006 and 2007, and Section 5 is a summary.

2. SELECTING PROMPT SENTENCES

Prompt sentences for recording a speech corpus for speech synthesis need to reflect target domains, in particular, by their phonetic characteristics. Achieving a good coverage of some phonetic units (e.g., diphone and triphone) is straightforward for limited domains, although perfect-quality open-domain synthesis is still not yet possible [6]. Therefore, designing prompt sentences basically involves the following stages:

- select a source text corpus to represent the target domains;
- analyze the source text corpus to obtain the unit statistics;
- select appropriate prompt sentences from the source text;
- inspect and remove unsuitable sentences.

[a)] He contributed to this work when he was affiliated with ATR-SLC.

2.1. Source text corpus

Two domains were taken into account to construct our English speech corpus for speech synthesis: (1) conversational communication, and (2) news-reading. Consequently, we extracted prompt sentences from huge collections of texts (hereafter referred to as source text corpora) described below.

1. An English Basic Travel Expression Corpus (BTEC),
2. a large-scale English newspaper corpus (NEWS).

The two text corpora were first pre-processed by

1. Using a Festival [10] tool to decompose every paragraph in the text corpora into *utterances* (a predicted unit in [10] 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*.
3. Filtering out long *sentences* in order to enable natural uttering. More specifically, all the *sentences* that have more than 25 words were filtered out. Also, we filtered out short *sentences* in NEWS corpus so as to keep the average sentence length relatively long.

A Festival [10] tool was then used to analyze the source text corpora for obtaining the statistics of basic units (monophones, diphones, triphones), POS (part of speech), and finding out potential distinct diphones and triphones (hereafter referred to as diphone/triphone types) existing in the source text corpora. There were 34 POS tags and 40 phonemes (monophones) plus an extra /pau/ (pause/silence). A /pau/ is always assumed at the beginning and ending of a *sentence*, and it is also used to separate any two adjacent *utterances*.

Figure 1 shows the monophone distributions for the two source text corpora. The monophone with lowest occurrence is /zh/ with 0.034% in BTEC and 0.069% in NEWS. Table 1 shows counts of the units existing in the text corpora. While there are, for example, 1,680 distinct diphones in theory [1,680 = 41× 41 − 1 (i.e., /pau/–/pau/)], there actually exist only 1,472 diphone types in BTEC and 1,597 in NEWS. Table 2 shows Kullback-Leibler divergences between BTEC and NEWS for POS, monophones, diphones, and triphones. These results indicate that the distributions of units such as POS, diphone and triphone are different to some extent between news-writing and conversational texts. For this reason, we extracted two *sentence* sets, each from BTEC and NEWS separately, using a greedy algorithm that is roughly described in the next section.

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

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

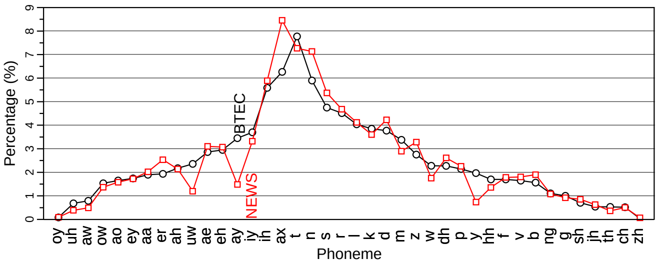

Fig. 1. Distribution of the 40 monophones existing in BTEC and NEWS text corpora.

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

2.2. Prompt sentence extraction

A greedy algorithm in [9] was used to extract a sentence set from a text corpus while maximizing the unit coverage. A unit type may be diphone, triphone, or POS in this paper, and the unit coverage is calculated in the following way.

Let U^i be a predefined set of linguistic units like POS, monophone, etc., where $U^i = \{u_1^i, ..., u_{n^i}^i\}$ with size n^i. Let S be a *sentence* set extracted from a text corpus. The coverage of S for U^i, denoted by $C_S^{U^i}$, is defined as $C_S^{U^i} = \sum_{j=1}^{n^i} p(u_j^i) \times \delta(u_j^i)$, where $p(u_j^i)$ indicates the occurrence frequency of unit u_j^i in the source text corpus, and $\delta(u_j^i) = 1$, if $u_j^i \in S$. Otherwise, $\delta(u_j^i) = 0$. By definition, $\sum_{j=1}^{n^i} p(u_j^i) = 1$. After this, C_S^{mono}, C_S^{di}, C_S^{tri}, and C_S^{POS} stand for the coverage of sentence set S for monophone, diphone, triphone, and POS, respectively.

We intended to maximize C_S^{di}, C_S^{tri}, and C_S^{POS} while designing a sentence set. This was achieved by using a multi-level evaluation method [9] while deciding which sentences would be extracted out from the source text corpus. More specifically, we chose the sentence each time that has the highest score among the source text corpus (or N sentences for balancing computational cost), after comparing the contributions of all the sentences to current sentence set S in the following priority.

1. The sentence maximizes C_S^{di}.
2. If there are more than one sentence that achieve 1., pick the sentence that maximizes C_S^{tri}.
3. If there still are more than one sentence that achieve 2., pick the sentence that maximizes C_S^{POS}.
4. If there still are more than one sentence that achieve 3., pick the sentence that maximizes the number of triphone variants at specific positions: the beginning, ending, and a few middle positions of *utterances*.

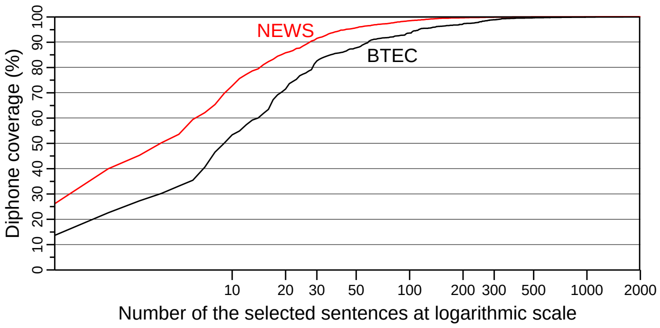

Fig. 2. Diphone coverage for the first 2,000 sentences extracted from BTEC (black curves) and NEWS (red curves).

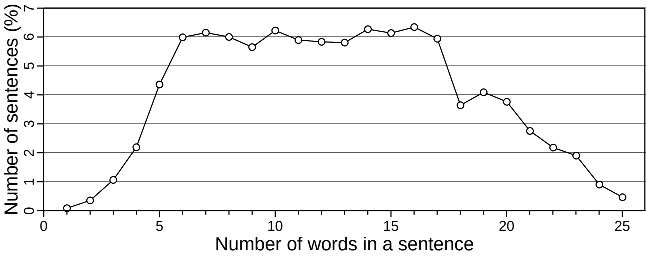

Fig. 3. Length distribution in words in the two sentence sets.

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 this step-by-step procedure. This is, however, extremely time-consuming. In practice, N was fixed at 2,000.

What we actually did was to extract independently 7,633 sentences from BTEC and 4,904 sentences from NEWS. Table 3 lists the coverage of the sentence sets for monophone (C_S^{mono}), diphone (C_S^{di}), triphone (C_S^{tri}), and POS (C_S^{POS}). In addition, 21.6% of sentences end with a question mark selected from BTEC and 3% from NEWS. Figure 2 shows in part the diphone coverage as a function of set sizes. Figure 3 shows the sentence length distribution in words for the two sentence sets; the average length is 12.7 words.

Table 3. Unit coverage of the two extracted sentence sets.

	#Sent.	C_S^{mono}	C_S^{di}	C_S^{tri}	C_S^{POS}	#Ques.
BTEC	7,633	100%	99.99%	99.87%	100%	21.6%
NEWS	4,904	100%	99.99%	99.86%	100%	3.0%

3. RECORDING PROMPT SENTENCES

The extracted prompt sentences together with ARCTIC [6] and some other sentences were recorded in a sound-proof room at ATR by an American male native speaker, who won against other three natives in a well-designed audition. The talent speaker was 52 years old at the recording time. His birthplace was the east coast of the USA and had been brought up there at least twelve years. The recording period lasted more than one month, including 18 recording days.

A major factor causing undesirable voice variability may be the physical and mental conditions of a speaker besides the influence of an audio equipment and the speaker's recording experiences. In order to suppress the time-dependent effects on voice variations as described in [8] during the recording process, we took the following measures to minimize the impact of the factors on the speech corpus in recording time [9].

1. Using identical audio equipment throughout the recording. The layout of the recording studio was kept the same, while the volume of the microphone amplifier could be adjusted, when it was really necessary.

2. Keeping the mouth-microphone distances as close to 30 cm as possible. 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. The experimental results as described in [9] suggested the proximity effect could be limited to 3 dB when keeping the mouth-microphone distances as close to 30 cm as possible during the recording period.

3. Limiting the amount of speech data to be collected in each recording day (less than one hour pure speech).

4. Dividing a recording day into several sessions, 20-minute work, 20-minute break, alternately.

5. Having the speaker to listen to a few selected samples for anchoring a "normal" voice for each recording day.

6. Having the speaker to insert a two-second (or longer) pause between any two sentences while uttering.

To obtain high SNR, a large diaphragm condenser microphone with a cardioid directional pattern was used [4]. The speech data was digitized at a sampling frequency of 48 kHz with 24-bit precision, and harddisk-recorded. After reading errors were removed by human inspection, speech data were separated into utterances, high-pass filtered at 70Hz, and finally precision-converted down to 16 bits after amplitude adjustment. After that, phone segmentation and F0 extraction were then conducted. The results were not manually corrected.

The contents of this English corpus include ARCTIC [6], news, and travel conversations as shown in Table 4. All of these sentences were uttered in "reading style." The sizes in hours do not include silences at utterance initials and finals but include utterance medial pauses. Figure 4 shows the distribution of the 40 monophones in this speech corpus. The phoneme /zh/ has the smallest number of instances: 544.

We evaluated the time-dependent voice variations in the recording based on a measure of the minus log-likelihood of long-term power spectral densities [9]. The experimental results indicated that potential time-dependent voice quality

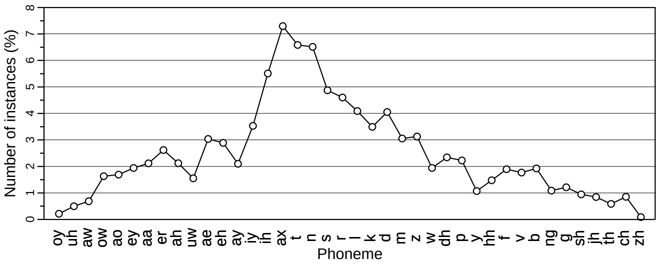

Fig. 4. Distribution of the 40 monophones in the American English speech corpus.

variability in the recording, in comparison with a Chinese corpus developed at ATR, was well controlled through the careful setting of critical factors such as the proximity effect of the microphone, the layout of the studio, and the speaker during the recording process. The details are available in [9].

Table 4. Contents of the English speech corpus.

Genre	Size in hours	Number of sentences	Number of phonemes
ARCTIC	0.868	1,132	35,970
BTEC	7.122	7,633	267,007
NEWS	8.157	4,904	302,325
Miscellaneous	0.563	652	26,836
Total	16.71	14,321	632,165

4. DATASETS FOR THE BLIZZARD CHALLENGE

The Blizzard Challenge was held in January 2005 [11] to evaluate corpus-based speech synthesis techniques. One of the challenge features is to provide a common dataset for all the participants to build their voices for subject evaluation. In 2005, the CMU ARCTIC database [6] was used as the common dataset. Since the size of this database is limited (about one hour by each speaker), it is natural to have interests in the same challenge while using a larger-sized speech corpus. For this reason, ATR-SLC (*http://www.atr.jp*) then provided a subset of this English speech corpus for Blizzard Challenge 2006 and 2007 [7]. Five hours of speech data were released to 18 participants in 2006, and eight hours to 18 participants in 2007, who returned a signed user agreement. Tables 5 and 6 summarize the contents of the two datasets released in 2006 and 2007, respectively.

Table 5. Dataset provided for the Blizzard Challenge 2006.

Genre	Size (h)	Number of sentences	Number of phonemes	C_S^{di} (%)	C_S^{tri} (%)
ARCTIC	0.86	1,132	35,970	n/a	n/a
BTEC	2.00	2,087	77,881	99.98	97.88
NEWS	1.97	1,016	74,938	99.96	96.73
Total	4.84	4,235	188,789	n/a	n/a

Table 6. Dataset provided for the Blizzard Challenge 2007.

Genre	Size (h)	Number of sentences	Number of phonemes	C_S^{di} (%)	C_S^{tri} (%)
ARCTIC	0.86	1,132	35,970	n/a	n/a
BTEC	3.61	3,717	138,196	99.99	99.42
NEWS	3.71	2,030	142,300	99.97	98.06
Total	8.19	6,879	316,466	n/a	n/a

5. SUMMARY

This paper introduced ATRECSS, an American English speech corpus developed at ATR for speech synthesis. Part of the speech corpus was provided as the common corpus for developing speech synthesizers to the participants of Blizzard Challenge 2006 and 2007 under a license agreement.

6. REFERENCES

[1] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura, “ATR ν-talk speech synthesis system,” in *Proc. ICSLP*, 483–486, 1992.

[2] A.W. Black and P. Taylor, “CHATR: a generic speech synthesis system,” in *Proc. COLING*, 983–986, 1994.

[3] W.N. Campbell, “CHATR: a high-definition speech resequencing system,” in *Proc. Joint Meeting of ASA and ASJ*, 1223–1228, 1996.

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

[5] S. Nakamura, *et al.*, “The ATR multi-lingual speech-to-speech translation system,” *IEEE Trans. on Speech and Audio Processing*, Vol. 14, No. 2, 365–376, 2006.

[6] J. Kominek and A.W. Black, “CMU ARCTIC database for speech synthesis,” *Technical Report CMU–LTI–03–177*, 2003.

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

[8] 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.), 19–33, 2004.

[9] J. Ni, T. Hirai, and H. Kawai, “Constructing a phonetic-rich speech corpus while controlling time-dependent voice quality variability for English speech synthesis,” in *Proc. of ICASSP2006*, I-881–I-884, 2006.

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

[11] A.W. Black and K. Tokuda, “The Blizzard Challenge 2005: evaluating corpus-based speech synthesis on common datasets,” in *Proc. of Interspeech*, 77–80, 2005.