Communicative Speech Synthesis with XIMERA: a First Step

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

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

{shinsuke.sakai,jinfu.ni,ranniery.maia,satoshi.nakamura}@atr.jp
tokuda@ics.nitech.ac.jp,minoru.tsuzaki@kcua.ac.jp
tomoki@is.naist.jp,hisashi.kawai@kddilabs.jp

Abstract

This paper presents a corpus-based approach to communicative speech synthesis. We chose "good news" style and "bad news" style for our initial attempt to synthesize speech that has appropriate expressiveness desired in human-human or human-machine dialog. We utilized 10-hour "neutral" style speech corpus as well as smaller corpora with good news and bad news styles, each consisting of two to three hours of speech from the same speaker. We trained target HMM models with each style and synthesized speech with unit databases containing speech with the relevant style as well as neutral speech. From the listening tests, we found out that intended communicative styles were comprehended by listeners and that considerably high mean opinion score on naturalness was achieved with rather small, style-specific corpora.

1. Introduction

Corpus-based approaches to speech synthesis have been very popular in the past decade and concatenative synthesizers have been especially successful due to its high naturalness [1, 2, 3, 4]. After achieving highly natural-sounding synthetic speech, however, the research and user communities of speech synthesis have become more aware about the issues with using speech synthesizers that speaks in an articulate but uniform manner in all the situations in human-machine dialogs or machine-mediated human-human dialogs. Research efforts on expressive and emotional speech synthesis, therefore, have become more and more active these days [5, 6]. We are aiming at developing speech synthesis technology that is useful for human-machine dialogs such as those in speech-enabled automatic conversational services as well as machine-mediated human-human dialogs such as conversations through the speech-to-speech translation system [7]. To investigate the possibilities of achieving synthetic speech appropriate for the communicative purposes in those systems, we looked at different styles of spoken communication such as conveying *good news*, *bad news*, and *focus* (or emphasis) [5], and selected good news and bad news for the styles to handle in our first attempt at synthesizing communicative speech. Part of the reason that we did not choose the focus was that it seems the objective of making some part of the utterance more salient than the others is often achieved by some other linguistic means such as using a different syntactic structure or adding small function words, rather than prosodic means in Japanese, which was the first target language for our communicative speech synthesis efforts.

In this paper, we report our initial attempt at communicative speech synthesis in the framework of XIMERA, a concatenative speech synthesis system [4]. We developed two-hour additional speech corpora in good news and bad news styles and trained HMM target models using these corpora. We also used these corpora together with the 10-hour corpus of neutral speech to generate speech with communicative styles. We tested how much desired styles were achieved and how much of naturalness was maintained by subjective listening tests. In the rest of the paper, we introduce the XIMERA concatenative speech synthesis system, followed by a description of the present approach to communicative speech synthesis. We then report on the experiments followed by the conclusion.

2. XIMERA

The block diagram of Figure 1 shows the main procedures conducted by XIMERA. Like most concatenative TTS systems, XIMERA is composed of four major modules, namely text processing, prosodic parameter generation, segment selection, and waveform generation modules. The target languages of XIMERA are Japanese, Chinese and English. Although the framework of corpus-based synthesis is language independent, most modules, in reality, must be developed or tuned for a target language. The language dependent modules comprise text processing, acoustic models for prosodic parameter generation, speech corpora, and the cost function for segment selection. The search algorithm for segment selection is also related to the target language via the cost function.

2.1. Text processing

The text processing module consists of three sub-modules for morphological analysis, rough syntactic analysis, and pronunciation and accent generation. The morphological analysis is conducted based on a bigram language model and a morpheme dictionary consisting of 239,591 Japanese or 195,959 Chinese entries. The rough syntactic analysis determines (1) a depen-

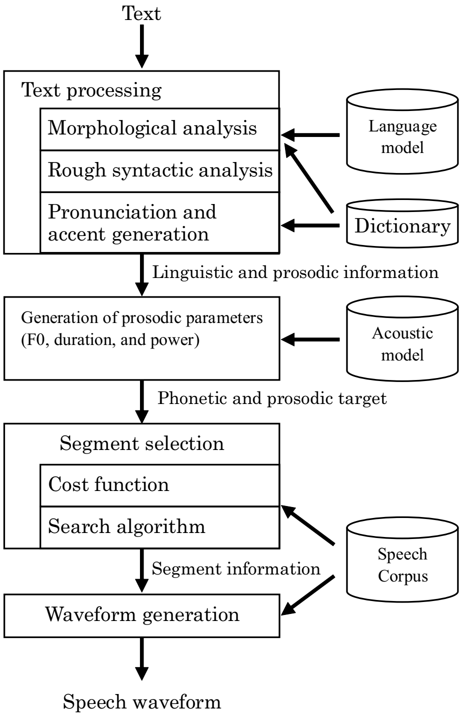

Figure 1: Main procedures performed by the XIMERA TTS system.

dency between adjacent words, which is mainly used for F0 generation, and (2) clause boundaries, which is mainly used for pause insertion. The pronunciation generation determines the readings of homographs and euphonic changes of unvoiced to voiced sounds. The accent generation determines the accent type of an accentual phrase based on accent types and the accent concatenation features of the constituent morphemes.

2.2. Generation of prosodic parameters

Prosodic parameters, namely $F0$, phone duration, and power, are generated according to the HMM-based speech synthesis technique [8, 9]. In other words, XIMERA includes an HMM-based synthesizer whose purpose is to produce the duration and power of the final concatenated waveform. Therefore, each HMM observation vector is composed of: (1) power; (2) mel-cepstral coefficients (without the 0-th coefficient); and $F0$. The generated parameters are also included in the concatenation cost for target selection.

2.3. Segment selection

2.3.1. Processing unit

The minimal processing unit is a half-phone [10, 11]. For Japanese synthesis, concatenation at a C-V boundary is inhibited by definition of the cost function. Moreover, a half phone unit in the resultant unit sequence should be at least either followed or preceded by a unit that was adjacent to it in the original speech corpus.

2.3.2. Cost function

The cost function of a sentence for segment selection is given by

$$C_g = \frac{1}{N}\sum_{i=1}^{N} C_l(u_i, t_i)^p, \quad (1)$$

where N denotes the number of targets in the sentence, C_l denotes a local cost at the target t_i, and u_i and t_i respectively denote the i-th target and segment candidate. The power p was determined to be 2 as a result of perceptual experiments [12]. The local cost is given by

$$\begin{aligned} C_l(u_i, t_i) = {} & w_{F0}C_{F0}(u_i, t_i) + w_{dur}C_{dur}(u_i, t_i) + \\ & w_{cen}C_{cen}(u_i, t_i) + w_{F0c}C_{F0c}(u_i, t_i) + \\ & w_{env}C_{env}(u_i, t_i) + w_{spg}C_{spg}(u_i, t_i), \quad (2) \end{aligned}$$

where $C_{F0}(u_i, t_i)$, $C_{dur}(u_i, t_i)$, and $C_{cen}(u_i, t_i)$ respectively denote errors in $F0$, segment duration, and spectral centroid between the target and a segment candidate; representing therefore the target costs. On the other hand, $C_{F0c}(u_i, t_i)$, $C_{env}(u_i, t_i)$, and $C_{spg}(u_i, t_i)$ respectively denote discontinuities of $F0$, phonetic environment, and spectrum between adjacent segments; representing the concatenation costs. w_{F0}, w_{dur}, w_{cen}, w_{F0c}, w_{env}, and w_{spg} are corresponding weights for the local costs. Mappings from acoustic measures into the above local costs and weights were optimized based on perceptual experiments [13].

2.3.3. Search

The optimal sequence of waveform segments is searched for by using the Viterbi algorithm [14]. A problem due to large corpora is the heavy computation load required for evaluating candidate segments. To reduce the amount of computation, pre-selection based on target sub-costs is adopted.

2.4. Signal processing

XIMERA does not utilize prosodic modification of the final concatenated waveform. Toda et al. reported in [15] that the unnaturalness caused by prosodic modification algorithms, such as STRAIGHT [16], degrades the synthesized speech when the corpus size is greater than two hours. Therefore the waveform generation module is based on simple waveform concatenation. The concatenation point is searched for within a 5-ms range around the segment boundaries so that a short-term cross-correlation coefficient is maximized.

3. Communicative speech synthesis with XIMERA

We developed corpora of good news and bad news styles with the same speaker that we recorded 60 hours of neutral speech. Due to the limited time available for developing the prompt text, we reused part of the existing prompt text. The set of prompt sentences were equivalent to those used for a 2.6-hour subset of the existing neutral speech. Roughly 50% of the prompts were newspaper sentences, 25% were phonetically balanced sentence set, and the rest were travel conversation and novel sentences. The sentences in written form were modified to have conversational sentence ends.

The procedures of database collection, correction, labeling and phonetic segmentation, were conducted as described in [4], following the same directions employed for the construction of

Table 1: *Speech corpora used in this experiment.*

Speech corpus	ID	# utterances	# phones	Size
Neutral	N10	12,169	515,845	10 h
Neutral	N2	1,962	135,142	2.6 h
Good news	G2	1,962	139,551	3.0 h
Bad news	B2	1,962	138,558	3.2 h

Table 2: *The corpus sizes for training the target HMMs.*

Database style	Size (h)	# utterances	# phones	# feature labels
Neutral	2.3	1,807	118,701	117,638
Good news	2.8	1,852	128,468	126,962
Bad news	2.7	1,726	118,640	117,070

the original XIMERA database. The sizes of the corpora developed, with good-news and bad-news styles, were 3.0 and 3.2 hours, respectively.

4. Experiments and results

4.1. Goal

The effectiveness of an approach can be evaluated by some designed experiments. For this purpose, we built several TTS systems from the corpora listed in Table 1, under XIMERA framework, and conducted two perceptual experiments to investigate the ability of conveying communicative speech synthesis with a certain degree of naturalness. Therefore, Experiment I is intended to evaluate the naturalness of synthetic speech in each target speaking style, namely "good news", "bad news" and neutral styles[1], whereas Experiment II is focused on rating the appropriateness of conveying "good news" and "bad news" by the synthetic speech.

4.2. Prosody generation modules and unit databases

In order to synthesize "good news", "bad news" and "neutral" speech, we trained contextual HMMs for three style-specific prosody generation modules. Table 2 shows the amount of database used to train each style and the resulting number of feature labels.

The unit database were generated from the database sets shown in Table 1. Note that two versions of "neutral" database were generated, one with two hours and other with ten hours.

4.3. Different system versions used in the experiments

Through a few combination of the corpora shown in Table 1, the six databases listed in Table 3 were developed. The database N2 is a sub-set of N10. Further, by combining the HMM-based prosody generation modules of Table 2 with these databases, eight TTS systems are used in this experiment. Table 4 illustrates which acoustic model is combined with each individual database. In the following, each system is described.

Table 3: *Description of the six databases for use in this test.*

ID	Size	Content
G2	3.0	3 h "good news" speech
G2+N10	13	3 h "good news"+10h neutral speech
N2	2.6	2.6 h neutral speech
N10	10	10 h neutral speech
B2	3.2	3.2 h "bad news" speech
B2+N10	13.2	3.2h "bad news"+10 h neutral speech

Table 4: *Combination of the three HMM-based targets with the six unit databases to form eight systems.*

Style	G2	G2+N10	N2	N10	B2	B2+N10
G	●	●		●		
N			●	●		
B				●	●	●

1. **System G–G2**
   - Target: "good news";
   - Unit database: G2.
2. **System G–G2+N10:**
   - Target: "good news";
   - Unit database: G2+N10.
3. **System G–N10:**
   - Target: "good news";
   - Unit database: N10.
4. **System N-N2:**
   - Target: "neutral";
   - Unit database: N2.
5. **System N-N10:**
   - Target: "neutral";
   - Unit database: N10.
6. **System B–B2:**
   - Target: "bad news";
   - Unit database: B2.
7. **System B–B2+N10:**
   - Target: "bad news";
   - Unit database: B2+N10.
8. **System B–N10:**
   - Target: "bad news";
   - Unit database: N10.

The use of style-specific target HMMs combined with neutral database is intended to test how good performance can be achieved by use of a neutral speech corpus only.

4.4. The test sentences

The eight TTS systems above were used to supply synthetic speech for use in a listening test. We chose ten *ambiguous* Japanese sentences. These sentence can be literally interpreted as "good news", "bad news", or "neutral." The use of *ambiguous* sentences is expected to be suited for testing synthetic speech in delivering an intended speaking style. Since there are eight versions of synthetic speech for each sentence, 80 distinct stimuli in total were yielded . The 80 stimuli are divided into two groups, 40 stimuli for each, using a different randomized order across groups. One group was used for Experiment I and the other for Experiment II.

[1] Hereafter referred to as G, B, and N, respectively.

4.5. Subjects

Twelve listeners participated this listening test, six male and six female speakers, all of whom are Japanese natives with normal hearing. These stimuli were presented to listeners with headphones in a silent office. The listeners were allowed to listen to a few samples before starting this test so as to get some idea of the quality of synthetic speech in the three styles. During this listening test, they could listen to each stimulus as many times as they liked, but could not go back and forth anyway.

4.6. Results and discussion

In Experiment I, the listeners were asked to rate the naturalness of synthetic speech on a 5-point scale from 1, the worst naturalness, to 5, very natural. In Experiment II, the same listeners were then instructed in the listening task of evaluating the appropriateness of synthetic speech in conveying "good news", neutral news, and "bad news" on a 7-point scale from -3 (very good "bad news"), 0 (neutral), and 3 (very good "good news").

Figure 2 shows Mean Opinion Scores (MOS) for each of the TTS systems enumerate above, and Figure 3 shows the MOS obtained in Experiment II for each system. Table 5 lists the number of stimuli in percentage evaluated as "bad news", "neutral," or "good news" on the 7-point scale for each system.

Several observations may be made from the experimental results. Firstly, synthetic speech in "good news" style has high naturalness, even we use the "neutral" unit database. The MOS obtained by System G-N10 in Figure 2 illustrates this fact.

Secondly, when both the targets and databases were built from style-relevant speech corpus, the resulting systems achieved better performance than the others. For instance, System G-G2 outperforms systems G-G2+N10 and G-N10, and System B-B2 outperforms systems B-B2+N10 and B-N10. The degradation in naturalness from systems G-G2 to G-G2+N10, and from B-B2 to B-B2+N10, perhaps might be partly caused by the unit selection algorithm, since the former was included in the latter. On the other hand, the MOS values obtained in Experiment II for rating the appropriateness of intended styles were quite similar in both systems G-G2 and G-G2+N10 as well as systems B-B2 and B-B2+N10, as shown in Figure 3.

Thirdly, when focusing on the naturalness of systems G-G2, N-N2, and B-B2, which are similar in speech corpus size, systems with "good news" and "bad news" styles achieved considerable better performance than the neutral system N-N2. This result might indicate that appropriate styles could possibly improve the naturalness of the synthesized speech. In other words, an effective way to improving naturalness in small corpus speech synthesis is to generate synthesized speech in varied styles.

Finally, while "good news" speech presented better naturalness than "bad news," "bad news" speech could give clearer impression than "good news" speech, according to Figure 3.

Basically, the results of Experiment II showed that listeners could correctly identify synthesized speech in a particular style ranging from 98.4% for "bad news" to 66.7% for "good news".

The results also imply that there is an overlap for distinguishing between "good news" and "neutral" styles. This can be supported by the numbers in Table 6, which shows how many units were selected from N10 in both systems B-B2+N10 and G-G2+N10. As shown in Table 6, 59% of units were selected from N10 when synthesizing the sentences in "good news" style, while only 3% of units were selected from N10 when synthesizing "bad news." Further, a deeper examination showed that the vocal range in uttering "bad news" style (by the same speaker) is sharply narrowed and the mean is considerably lowered. Table 7 lists the voice ranges measured from B2, N2, and G2, and some examples are displayed in Figure 4.

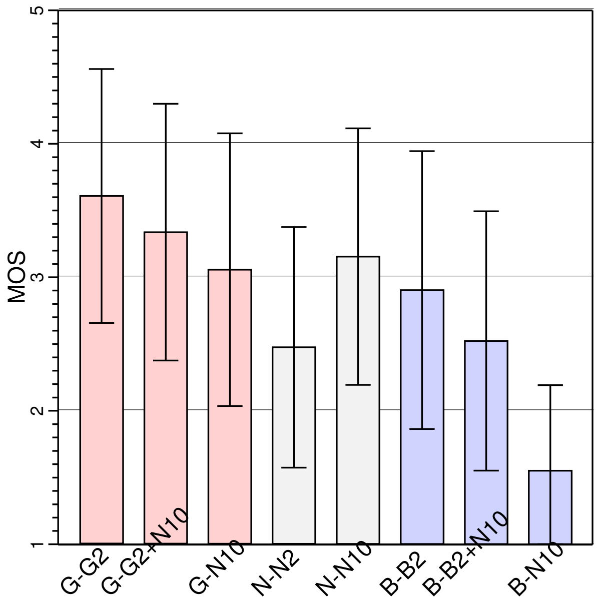

Figure 2: *Mean Opinion Scores and standard deviations for each of the synthesis systems in Experiment I described in the text.*

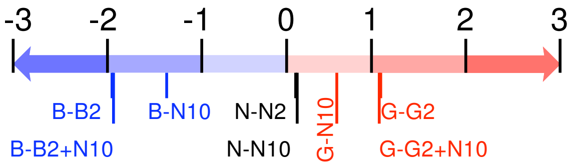

Figure 3: *Mean Opinion Scores for each of the synthesis systems in Experiment II described in the text.*

5. Conclusions

In this paper, we presented a corpus-based approach to communicative speech synthesis. We chose "good news" style and "bad news" style for our initial attempt to synthesize communicative speech and collected speech corpora with those styles. Target HMMs were trained with these style-specific corpora, whereas we also made use of neutral speech corpus for building style-specific unit databases in order to know how this existing resource can be utilized to generate speech with expressions relevant in the communication. From the listening tests, we found out that intended communicative styles were comprehended by listeners and that considerably high mean opinion score on naturalness was achieved with rather small, style-specific corpora. Currently we need to have separate model trees for each of the communicative styles. We plan to examine the possibilities of having a single model tree where styles are handled as one of the features for clustering HMM target models.

Table 5: *Percentage of stimuli evaluated as individual levels in Experiment II.*

System ID	–3	–2	–1	0	1	2	3
B–B2	26.7	41.7	30		1.6		
B–B2+N10	26.7	43.3	23.3	6.7			
B–N10	16.7	25	40	11.7	5	1.7	
N–N2			15	63.3	18.3	3.3	
N–N10			16.7	60	18.3	5	
G–G2			3.3	28.3	30	30	6.7
G–G2+N10			3.3	28.3	35	28.3	5
G–N10			0.86	35	46.7	10	

Table 6: *Percentage of units selected from the neutral speech corpus when synthesizing "bad news" and "good news" styles.*

Style	System	Units selected from subset N10
Bad news	B–B2+N10	3%
Good news	G–G2+N10	59%

6. Acknowledgements

The authors are grateful to J. Kunita for various contributions to the corpus development and implementation of the system described in this paper.

7. References

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," pp. 373–376, 1996.

[2] M. Chu, H. Peng, Y. Zhao, Z. Niu, and E. Chang, "Microsoft Mulan – a bilingual TTS system," pp. I–264–I–267, 2003.

[3] E. Eide *et al.*, "Recent improvements to the IBM trainable speech synthesis system," pp. I–708–I–711, 2003.

[4] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "XIMERA: a new TTS from ATR based on corpus-based technologies," in *Proc. of ISCA Speech Synthesis Workshop*, 2004.

[5] J. Pitrelli *et al.*, "The ibm expressive text-to-speech synthesis system for american english," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1099–1108, 2006.

[6] C. Wu *et al.*, "Voice conversion using duration-embedded bi-hmms for expressive speech synthesis," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1109–1116, 2006.

[7] S. Nakamura *et al.*, "The atr multi-lingual speech-to-speech translation system," *IEEE Trans. on Speech and Audio Processing*, vol. 14, no. 2, pp. 365–376, 2006.

[8] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. of ICASSP*, 2000.

[9] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. of EUROSPEECH*, 1999.

Table 7: *F0 ranges for the three styles in a speaker.*

Style	Mean	Standard deviation	F0 range
Bad news	161 Hz	33 Hz	(105 Hz, 242 Hz)
Neutral	249 Hz	56 Hz	(131 Hz, 365 Hz)
Good news	283 Hz	61 Hz	(167 Hz, 407 Hz)

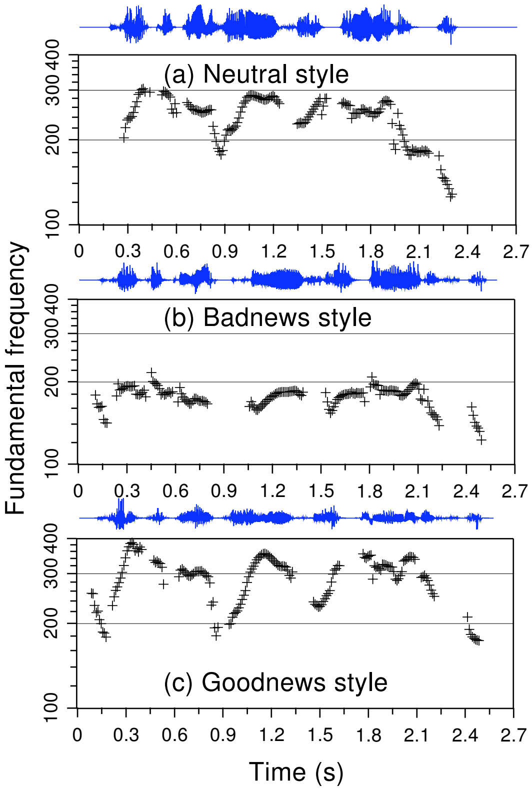

Figure 4: *Examples of F0 contours for a Japanese sentence produced by a native in (a) neutral styles, (b) "bad news" styles, and (c) "good news" styles.*

[10] M. Beutnagel, A. Conkie, and A. Syrdal, "Diphone synthesis using unit concatenation," in *Proc. of International Workshop on Speech Synthesis*, 1998.

[11] T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano, "Unit selection algorithm for Japanese speech synthesis based on both phoneme unit and diphone unit," in *Proc. of ICASSP*, 2002.

[12] T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano, "Segment selection considering local degradation of naturalness in concatenative speech synthesis," in *Proc. of ICASSP*, 2003.

[13] T. Toda, H. Kawai, and M. Tsuzaki, "Optimizing sub-cost functions for segment selection based on perceptual evaluations in concatenative speech synthesis," in *Proc. of ICASSP*, 2004.

[14] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. of ICASSP*, 1996.

[15] T. Toda, H. Kawai, and M. Tsuzaki, "Effectiveness of prosodic modification in concatenative text-to-speech syn-

thesis," in *Proc. of the Fall Meeting of the Acoust. Soc. of Japan*, 2003. In Japanese.

[16] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "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, Apr. 1999.