Entropy-Based Sentence Selection for Speech Synthesis Using Phonetic and Prosodic Contexts

Takashi Nose[1], *Yusuke Arao*[2], *Takao Kobayashi*[2], *Komei Sugiura*[3], *Yoshinori Shiga*[3], *Akinori Ito*[1]

[1]Graduate School of Engineering, Tohoku University
[2]Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology
[3]National Institute of Information and Communications Technology

tnose@m.tohoku.ac.jp, komei.sugiura@nict.go.jp, yoshinori.shiga@nict.go.jp

10.21437/Interspeech.2015-692

Abstract

This paper proposes a sentence selection method using a maximum entropy criterion to construct recording scripts for speech synthesis. In the conventional corpus design of speech synthesis, a greedy algorithm that maximizes phonetic coverage is often used. However, for statistical parametric speech synthesis, phonetic and prosodic contextual balance is important as well as the coverage. To take account of both of the phonetic and prosodic contextual balance in the sentence selection, we introduce and maximize the entropy of the phonetic and prosodic contexts, such as biphone, triphone, accent, and sentence length. The objective experimental results show that the proposed method achieves better coverage and balance of contexts and reduces spectral and F0 distortions compared to the random and coverage-based sentence selection methods.

Index Terms: speech synthesis, sentence selection, entropy, corpus design

1. Introduction

In corpus-based speech synthesis, design of the corpus is very important when we construct a new speech corpus. Specifically, constructing phonetically and prosodically balanced text corpus with high contextual coverage as a recording script is essential to synthesize natural-sounding speech close to the original one. In addition, when we attempt to synthesize speech with various speakers, emotional expressions, and speaking styles, the costs for speech recording and context labeling increase and thus the optimization of sentence selection becomes more important.

There have been many studies for corpus design and sentence selection mainly in the area of concatenative speech synthesis. In [1] and [2], a constraint of the minimum number of samples was introduced for biphones/triphones and a sentence set was constructed that maximizes the phonetic coverage under the constraint. This type of selection problem is known as an NP problem that cannot be solved in polynomial time, and a greedy approach was taken instead of achieving a strict optimization. In [1], the minimum number of samples is given in an ad-hoc manner and the influence of the number of samples on the synthesis performance and quality is not well discussed.

It has been shown that sentence selection methods using phonetic information are important for the area of speech recognition and speech translation [3, 4]. In addition to the phonetic information, prosodic contextual coverage influences naturalness and reproducibility of synthetic speech as well as phonetic coverage, and several methods have been proposed for maximizing phonetic and prosodic coverage [5–7]. In [5], a measure of coverage was introduced that incorporates the distribution of fundamental frequency (F0) and phoneme duration predicted by the prosody generation module of a TTS. However this approach is designed for concatenative speech synthesis where the perceptual damage to naturalness due to prosody modification is taken into account, and is not designed for statistical parametric speech synthesis that we focus on. There is another approach where diphone is used as a synthesis unit and prosodic coverage and/or balance is taken into account by classifying each diphone into a lexical stress or a prosodic unit called prosodeme [6, 7].

In the above studies, coverage and balance of constructed sentence sets were examined with concatenative speech synthesis systems while no experimental evaluations using the resultant synthetic speech were shown. Recently, statistical parametric speech synthesis such as HMM-based one is widely studied. HMM-based speech synthesis can generates smooth and stable speech parameter sequences using less amount of speech data than concatenative synthesis and has a small footprint (typically a few MBytes). In addition, the emotional expressions and speaking styles are well modeled and controlled by using style modeling, style interpolation, and style control techniques [8–10]. These techniques are indispensable to realizing a humanoid robot in the future.

In this paper, we propose a novel sentence selection method for constructing a more compact text corpus for speech synthesis where phonetic and prosodic contextual balances is simultaneously taken into account using entropy criterion. In this method, phonetic and prosodic contexts that are prominently important in the HMM-based speech synthesis are selected in advance and the total entropy of these contexts are maximized in a greedy manner for the given number of sentences. We show the superiority of the proposed method to the random and conventional phonetic-coverage-based methods.

2. Entropy-based sentence selection using phonetic and prosodic contexts

2.1. Definition of phonetic and prosodic contexts

In this study, we set the target language to Japanese and chose phonetic and prosodic contexts, i.e., phoneme, accent, and sentence length (number of moras), that were shown to be important in our previous experimental evaluations for HMM-based speech synthesis [11]. Each context is defined as follows:

2.1.1. Phoneme

The design of the phonetic context is based on the research on constructing ATR Japanese speech database [12]. To efficiently take account of the phonetic balance, all possible combinations

of two phonemes (biphones) appearing in the source sentences are used as symbols. Since some of phone segments are susceptible to the preceding and succeeding phonemes, those sequences of three phonemes (triphones) are also taken into account. There are four groups for the triphone symbols.

- Unvoiced vowel
  In Japanese phones, vowels /i/ and /u/ are often devoiced when the preceding and succeeding phonemes are unvoiced consonants. We extract the unvoiced vowels from phoneme labels of the speech corpus used in the experiment and use the triphones of unvoiced vowel with preceding and succeeding unvoiced consonants as symbols.
- Nasal
  Similarly to the unvoiced vowel, the vowels /i/ and /u/ are sometimes nasalized when the preceding and succeeding phonemes are nasal. We use the triphones of nasal with the preceding and succeeding vowels as symbols.
- Semivowel
  Semivowel has strong connection to the preceding and succeeding vowels. We use the triphones of semivowel with preceding and succeeding vowels as symbols.
- Contracted consonant
  Phoneme symbol /yy/ used as a part of contracted consonant [1] has strong connection to the preceding and succeeding phonemes and we use the triphones of contracted consonant with the preceding and succeeding phonemes as symbols.

The resultant unique numbers of biphones and triphones used in the experiments are 354 and 105, respectively.

2.1.2. Accent

Japanese is a pitch-accent language and the dominant factors of accentual contexts are (i) length (number of moras) of an accent phrase and (ii) accent type. To take account of the balance of pitch patterns, we combine the accent-phrase length and the accent type and use as a single symbol. There are 103 unique symbols related to the accentual context included in the sentences of the experiments.

2.1.3. Sentence length

The sentence length r is defined as the number of moras in the sentence. Since the prosodic characteristics of speech is thought to be not so sensitive to small differences of sentence length, we classify the sentence length for every q moras as follows:

$$n = \mathrm{floor}(r/q), \tag{1}$$

where we set $q = 5$ in this study. Resultantly, fourteen unique symbols ($n = 1, 2, ..., 14$) are used for sentence length.

2.2. Sentence selection based on contextual entropy

In this section, we describe the proposed entropy-based sentence selection algorithm where the balance of phonetic and prosodic contexts of Sect. 2.1 are simultaneously taken into account. The objective function S for stepwise optimization is defined as follows:

$$S = \sum_{m=1}^{M} w_m S_m, \quad S_m = -\sum_{n=1}^{N_m} p_{mn} \log_2 p_{mn}, \tag{2}$$

where M is the number of contexts that is set to $M = 3$ in this study. S_m and N_m are the entropy and the total number of symbols for the mth context, respectively. $w_m (0 \leq w_m \leq 1)$ is the weight for each context and is set to 1.0 in this study. p_{mn} is probability of symbol n appearing in the current sentence set and is approximated by relative frequency as used in [12]. Since maximizing the objective function for the population of the sentence set (source text corpus) is an NP problem, we maximize S in a greedy manner that was proposed in [13]. We show the process of the maximization when the target number of sentences D is given.

1. Calculate entropy for all sentences of the source text corpus and construct the initial sentence set L by selecting the sentence having the largest entropy.
2. Select a new sentence that maximizes the current total entropy S from the source text corpus and add the sentence to L.
3. Repeat the process 2 until the number of sentences in L is equal to D.

As a better sub-optimum solution of sentence selection, the replacement of two sentences for the current sentence set were employed in [12] and [14]. However, those methods takes huge calculation time when the size of the source text corpus is large. On the other hand, the calculation time increase linearly depending on the number of the sentences of the corpus in the proposed greedy approach, which is a strong advantage for a practical use.

3. Corpus for experiments

In this study, we use a dialogue speech database used in [15], This database was designed based on Kyoto tour guide dialogue corpus [16]. The corpus includes speech dialogue of 160 hours uttered by 328 tourist-guide sets. In [15], the recording scripts was constructed by transcribing the twenty-one dialogues having active conversation. The recording was conducted in a soundproof room and in the recording two professional voice talents sat across a table and read the scripted dialogues naturally without overlapping each other. The size of the dialogue speech data was 466 min per person. In this study, we use 8439 utterances of the guide person.

4. Experiments

To examine the effectiveness of the proposed method, we compared the following three methods in the experiments.

Random Sentences are randomly selected from the source text corpus until the total number of moras of the current sentence set is equal to or larger than the target number of moras D.

Coverage Sentences are selected sentence by sentence from the source text corpus so as to maximize the triphone coverage until the total number of moras of the current sentence set is equal to or larger than D.

Entropy Sentences are selected sentence by sentence from the source text corpus using the proposed entropy-based sentence selection until the total number of moras of the current sentence set is equal to or larger than D.

It is noted that we used the number of moras instead of the number of sentences because the amount of training data for speech synthesis is not determined by the number of sentences but the

[1] Contracted consonant called "youon" in Japanese is specific to Japanese language. The English "can" is pronounced in Japanese as /k yy a N/ and /k yy/ is the contracted consonant.

Table 1: *Entropy of respective contexts for constructed sentence sets.*

	Random	Coverage	Entropy
Phoneme	7.58	7.54	**7.69**
Accent	4.75	4.72	**5.39**
Sentence length	2.76	2.91	**3.10**

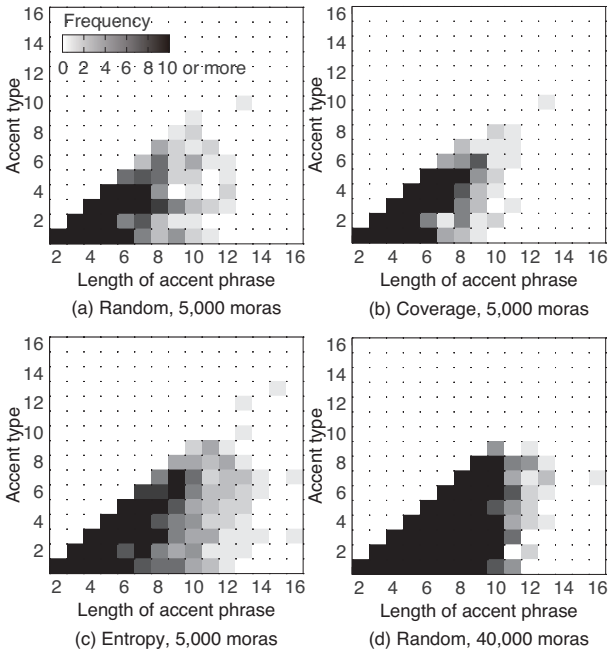

Figure 1: Two-dimensional histogram of length and type of accent phrase.

number of moras (or phonemes)[2]. We conducted objective experiments and compared the performance using the common database described in Sect. 3.

4.1. Experimental conditions

From the 8439 sentences, first we selected 93 sentences as test data using the proposed sentence selection algorithm described in Sect. 2.2. The remaining 8346 sentences was used as the source text corpus from which sentences were selected based on each method. In the random sentence selection, we constructed ten sets for the given number of sentences to examine the variation of the experimental results depending on the randomness. In the conventional coverage-based sentence selection, triphone coverage is maximized in a greedy manner using a similar manner to [1, 2]. In the corpus, accent labeling and phone segmentation were conducted automatically using Japanese text analysis and phoneme alignment, respectively. For the test data, the phone segmentation result was manually modified by a professional labeler.

Speech signals were sampled at a rate of 16kHz, and

[2]The random selection tends to result in a much smaller number of moras than the other two methods when the distribution of the sentence length is not uniform but Gaussian and this was not fair in the evaluation.

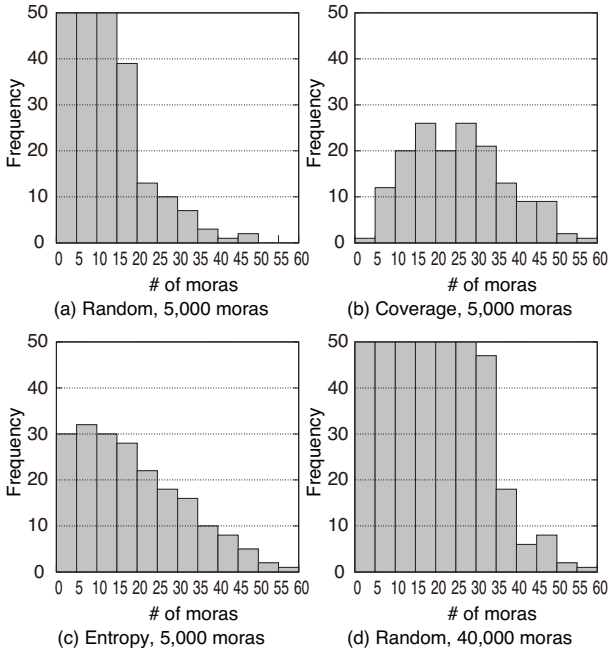

Figure 2: Histogram of sentence length.

STRAIGHT analysis [17] was used to extract spectral envelope, F0, and aperiodicity features with a five msec frame shift. The spectral envelope was converted to mel-cepstral coefficients using a recursion formula. The aperiodicity features were converted to average values for five frequency sub-bands, i.e., 0-1, 1-2, 2-4, 4-6, and 6-8 kHz. The resultant feature vector consisted of 40 mel-cepstral coefficients including the zeroth coefficient, log F0, five average band aperiodicities, and their delta and delta-delta coefficients. The total number of dimension was 138. We used five-state left-to-right hidden semi Markov models [18] with no state skip. Each state had a single Gaussian distribution with a diagonal covariance matrix. In the decision-tree-based context clustering, the minimum description length (MDL) was used as a stopping criterion [19].

4.2. Contextual coverage and balance

To compared the contextual balance of the sentence sets constructed by the three methods, we calculated the total entropy of the phonetic and prosodic contexts described in Sect. 2.1 for the constructed sets where the target number of moras was set to 40,000. Table 1 shows the result. From the table, we found that the proposed entropy-based sentence selection gave the largest entropy values for all contexts. Especially, the improvement was prominent in the accentual context, which is desirable to reducing the pitch distortion. To visually confirm the coverage and balance of accentual context, we plotted the two-dimensional histogram of length (number of moras) and type of accent phrase in Figure 1. In the figure, the frequency over ten is displayed as ten. In the case of Random, the result of the first set in the ten sets is shown. From the figures, we found that the proposed Entropy of (c) gave better contextual coverage and balance than Random and Coverage in the same number of target moras ($D = 5,000$) and almost comparable to the case of $D = 40,000$ in Random of (d). We also examined the coverage and balance of the sentence length. Figure 2 shows

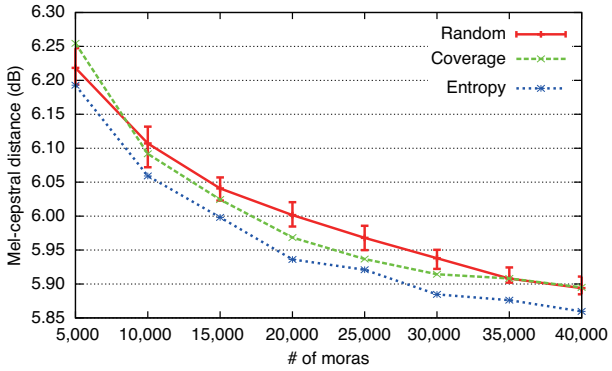

Figure 3: Mel-cepstral distance between original and synthetic speech.

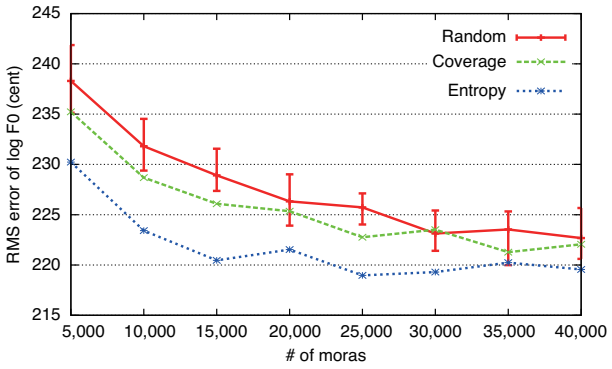

Figure 4: RMS error of log F0 between original and synthetic speech.

the histogram of the numbers of moras in the constructed sentence sets. There is similar tendency to the histograms of the accentual context (Figure 1) and Entropy has better performance when figures (a), (b), and (c) are compared.

4.3. Spectral and prosodic distortions in speech synthesis

Next, we conducted the speech synthesis experiments. HMMs were trained using the training data constructed by the three methods, and the mel-cepstrum, log F0, and phone duration sequences are generated from the HMMs. As the objective similarity, we used mel-cepstral distance, RMS error of log F0, and RMS error of phone duration between original and synthetic speech. The number of training sentences was set from 5,000 to 40,000 with an increment of 5,000 sentences. Figures 3 to 5 show the results. In Random case, the mean value of the ten sets is shown for each number of target sentences. The minimum and maximum values are also shown as error bars.

In terms of the mel-cepstral distance, the distortion decreases by increasing the number of moras of the training data. The proposed entropy-based sentence selection gave the best performance among the three methods, which indicates that the spectral reproducibility is improved by using the proposed method. We also found that the triphone coverage-based method is not effective when the target number of moras is relatively small. In that case, biphone coverage should be used

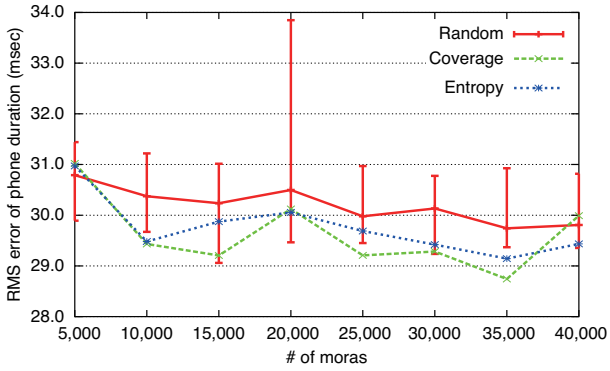

Figure 5: RMS error of phone duration between original and synthetic speech.

instead of triphone to reduce the number of symbols. The proposed method also showed the best performance in terms of the F0 distortion. In the figure, the distortion does not always decrease with the increase of the training data. A possible reason is that the accuracy of the accent labeling is not perfect and the wrong accent labels might affect the F0 reproducibility. In the case of Random, the range of the distortions of the ten sets are wider than those in the mel-cepstral distance, which indicates that randomly selecting the sentences has a larger risk than other two methods in terms of the F0 reproducibility.

As for the duration in Figure 5, we could not found a clear tendency of distortion reduction when increasing the target number of moras, which is different from the case of mel-cepstrum and log F0. We found that the automatic phone segmentation caused critical error in some sentences and the segmentation accuracy is not so high. This is because the database used in the experiment is not read speech but dialogue speech and the phonetic (spectral) variations is larger than that in read speech. However, the difference of the average performance of Coverage and Entropy is quite small (less than 0.1 msec), and the proposed method is the best choice of sentence selection in terms of the total performance of spectral and prosodic reproducibility.

5. Conclusions

In this paper, we proposed a novel sentence selection method based on the integration of entropies of phonetic and prosodic contexts. The proposed method was compared to the random and triphone coverage-based methods using a large dialogue corpus. The experimental results showed that the entropy-based sentence selection achieved better coverage and balance of the contexts for phoneme, accent, and sentence length. The speech synthesis experiment was also conducted and there was clear improvement in the reproducibility of spectral and F0 features. In the future work, we will conducted subjective evaluation tests to compare these sentence selection methods. The tuning of the weight for each context is also the remaining task.

6. Acknowledgements

Part of this work was supported by JSPS Grant-in-Aid for Scientific Research 23700195 and 24300071.

7. References

[1] H. François and O. Boëffard, "Design of an optimal continuous speech database for text-to-speech synthesis considered as a set covering problem." in *Proc. EUROSPEECH*, 2001, pp. 829–832.

[2] J. Matoušek, J. Psutka, and J. Krůta, "Design of speech corpus for text-to-speech synthesis," in *Proc. EUROSPEECH*, 2001, pp. 2047–2050.

[3] R. Chitturi, S. H. Mariam, and R. Kumar, "Rapid methods for optimal text selection," in *Recent advances in natural language processing*, 2005.

[4] J.-S. Zhang and S. Nakamura, "An improved greedy search algorithm for the development of a phonetically rich speech corpus," *IEICE Trans. Inf. & Syst.*, vol. 91, no. 3, pp. 615–630, 2008.

[5] H. Kawai, S. Yamamoto, N. Higuchi, and T. Shimizu, "A design method of speech corpus for text-to-speech synthesis taking account of prosody," in *Proc. ICSLP*, vol. 3, 2000, pp. 420–425.

[6] T. Lambert and A. P. Breen, "A database design for a TTS synthesis system using lexical diphones." in *Proc. ICSLP*, 2004, pp. 1381–1384.

[7] J. Matoušek and J. Romportl, "On building phonetically and prosodically rich speech corpus for text-to-speech synthesis," in *Proc. Computational Intelligence*, 2006, pp. 442–447.

[8] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E88-D, no. 3, pp. 503–509, 2005.

[9] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," *IEICE Trans. Inf. & Syst.*, vol. E88-D, no. 11, pp. 2484–2491, 2005.

[10] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 9, pp. 1406–1413, 2007.

[11] S. Yokomizo, T. Nose, and T. Kobayashi, "Evaluation of prosodic contextual factors for HMM-based speech synthesis," in *Proc. INTERSPEECH*, 2010, pp. 430–433.

[12] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.

[13] K. Shikano, "Phonetically balanced word list based on information entropy," Proc. Spring Meet. of the Acoustic Society of Japan, pp.211–212, 1984. (in Japanese).

[14] H. Francois and O. Boeffard, "The greedy algorithm and its application to the construction of a continuous speech database." in *Proc. LREC*, vol. 5, 2002, pp. 1420–1426.

[15] K. Sugiura, Y. Shiga, H. Kawai, T. Misu, and C. Hori, "A cloud robotics approach towards dialogue-oriented robot speech," Advanced Robotics (to appear).

[16] T. Misu, K. Ohtake, C. Hori, H. Kashioka, and S. Nakamura, "Annotating communicative function and semantic content in dialogue act for construction of consulting dialogue systems." in *Proc. INTERSPEECH*, 2009, pp. 1843–1846.

[17] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "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.

[18] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 825–834, 2007.

[19] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Jpn. (E)*, vol. 21, no. 2, pp. 79–86, 2000.