INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Investigating Effective Additional Contextual Factors in DNN-based Spontaneous Speech Synthesis

Yuki Yamashita[1], *Tomoki Koriyama*[2], *Yuki Saito*[2], *Shinnosuke Takamichi*[2], *Yusuke Ijima*[3], *Ryo Masumura*[3], *Hiroshi Saruwatari*[2]

[1]Faculty of Engineering, The University of Tokyo, Japan
[2]Graduate School of Information Science and Technology, The University of Tokyo, Japan
[3]NTT Media Intelligence Laboratories, NTT Corporation, Japan

yukiyama913@g.ecc.u-tokyo.ac.jp, tomoki_koriyama@ipc.i.u-tokyo.ac.jp

Abstract

In this paper, we investigate the effectiveness of using rich annotations in deep neural network (DNN)-based statistical speech synthesis. General text-to-speech synthesis frameworks for reading-style speech use text-dependent information referred to as context. However, to achieve more human-like speech synthesis, we should take paralinguistic and nonlinguistic features into account. We focus on adding contextual features to the input features of DNN-based speech synthesis using spontaneous speech corpus with rich tags including paralinguistic and nonlinguistic features such as prosody, disfluency, and morphological features. Through experimental evaluations, we investigate the effectiveness of additional contextual factors and show which factors enhance the naturalness as spontaneous speech. This paper contributes as a guide to data collection for speech synthesis.

Index Terms: speech synthesis, context, spontaneous speech, annotation, deep neural network

1. Introduction

Speech synthesis has been applied not only to speech interaction systems for smart phones and interactive robots, but also to various other applications such as announcement in public transportation and voice-overs for video works. With the expansion of these applications, there are increasing expectations for speech synthesis systems that can not only generate speech for a given text, but also express various emotions and intentions of speech.

In speech synthesis based on statistical models, since it is generally difficult to predict speech waveforms directly from input text, a pipeline model is used to statistically model the relationship between the context obtained from the text and speech features for waveform generation. Recently, end-to-end text-to-speech synthesis, which uses only text strings and phonemes as direct input has been widely studied [1, 2]. However, the performance of the end-to-end systems depends on the target language. For example, it has been reported that the performance of Japanese end-to-end speech synthesis, in which accent is important, is improved by adding accent information as context [3], and this indicates that context is still important.

In this paper, we focus on the construction of context in speech synthesis using deep neural networks (DNN) [4] among the statistical models of speech synthesis. Typical contexts are the triphones that indicate the previous and next phoneme information, the kind of prosody associated with a word or a phrase, and the length of a phrase. Such basic contexts can be used in DNN-based reading-style speech synthesis to produce high-quality speech. On the other hand, the context is flexible and it is easy to feed various information into the input of DNN. For example, adding a speaker vector as a context is successful in multi-speaker speech synthesis [5, 6]. The emotional expression in synthetic speech can be controlled by using the degree of expressiveness as a contextual factor [7, 8]. It is also reported that local features such as emphasis can be used as the input of DNN [9].

From the above, in order to synthesize speech with a wide variety of paralinguistic and nonlinguistic features, adding such features as context is promising. However, the use of too many input features often makes the training of DNN difficult and causes overfitting problem. Therefore, in this paper, we investigate the effectiveness of contextual factors in spontaneous speech synthesis using the Corpus of Spontaneous Japanese (CSJ) [10], which contains highly spontaneous speech samples such as lectures and dialogues. Specifically, the CSJ core data has a large amount of tags to describe information unique to spontaneous speech such as tone, phone prolongation, speaking style, and disfluency. We use these tags as additional contextual factors for DNN-based speech synthesis. We investigate the effective contextual factors for spontaneous speech synthesis comparing the reproducibility of synthesized speech to original speech in subjective evaluation experiments. Experimental results show that adding one context and combining two contexts tend to enhance the naturalness of spontaneous synthetic speech.

2. Related Work

This paper follows the previous studies of spontaneous speech synthesis: hidden Markov model (HMM)-based one [11] and DNN-based single-speaker one [12]. In the HMM-based spontaneous speech synthesis [11], the tags in CSJ were used as context of HMM-based speech synthesis [13], and the effect of CSJ tags on decision tree division was investigated. Although the experimental results show that tone information and phone prolongation were effective to enhance naturalness of synthetic speech, others did not affect the performance. A possible reason is that since the prediction in HMM-based speech synthesis is based on decision trees, it was difficult to deal with the complicated combination of contextual factors.

In our previous work [12], we showed that the use of additional contexts enhances the naturalness of spontaneous speech. However, the effectiveness of respective contexts was not clear because the trained data was limited to the data of a single speaker. Since DNN-based speech synthesis can easily model multiple speakers at the same time, in this paper we perform the

http://dx.doi.org/10.21437/Interspeech.2020-2469

evaluation on the synthetic speech in detail using a large amount of speech data.

3. Corpus of Spontaneous Japanese(CSJ)

Corpus of spontaneous Japanese (CSJ)[10] is designed for various purposes such as the analysis and modeling of spontaneous speech. It is a database containing a large amount of spontaneous speech of modern Japanese with rich annotations. Especially, the *core data* in CSJ has a huge amount of manual annotation. It contains 201 talks by 137 speakers.

The XML file in CSJ includes detail tags [14] which has hierarchical structure in the core data as follows:

```
<Talk>
  <IPU>
    <LUW>
      <SUW>
        <TransSUW>
          <Mora> or <NonLinguisticSound>
            <Phoneme>
              <Phone>
                <XJToBILabelTone>
                <XJToBILabelWord>
                <XJToBILabelBreak>
                <XJToBILabelPrm>
                <XJToBILabelMisc>
```

Here, we describe how to use these XML elements to form contexts.

<Talk>: In CSJ, speech data is stored as a set of utterance sequences referred to as a "talk." Talks include academic presentation speech, simulated public speaking, reading aloud, and dialogues (i.e., interviews, free dialogues, and task-oriented conversations). The types of talk and speaker information can be obtained from this tag.

<IPU>: For reading-style speech, in general, sentences are used as a unit of utterance. However, this is not appropriate for spontaneous speech because the end of the sentence is not always uttered. In CSJ, IPU (inter-pause unit) is used as an utterance unit of transcriptions, in which 200 ms pauses are regarded as the boundaries of utterances.

<LUW>, <SUW>: Since Japanese is an agglutinative language, there is a high degree of freedom in the definition of "word" [15]. Therefore, CSJ takes two types of unit in words: LUW (long-unit word) and SUW (short-unit word). An SUW is the shortest unit defined by the dictionary UniDic [16]. An LUW is a longer unit representing compound words. Information about the part of speech, conjugation type, and conjugation form can be obtained from these tags.

<TransSUW>: This tag is used to indicate disfluent utterances such as fillers, word fragments, and restatements.

<Mora>: This tag has kana information, and it can be also used to count the number of moras in phrases or clauses and its position in them.

<NonLinguisticSound>: Nonlinguistic information such as breaths and laughing is labeled in this tag. This tag also includes a vowel-nasal filler denoted by "VN," which appears in the response utterances in dialogs.

<Phoneme>, <Phone>: <Phone> tags have phone-related annotation in detail, e.g., the beginning and end times and devoiced vowels. In this study 59 types of phone entities are used. The <Phoneme> tag is a group of <Phone> tags, which we ignore in this study.

<XJToBILABEL*>: These tags are the prosodic information provided by the X-JToBI [17], a prosodic labeling scheme for spontaneous Japanese. The tag <XJToBILabelTone> includes tone labels of accent (A), initial boundary tone (%L, %H), boundary pitch movement (L%, HL%, LH%, HLH%), other tags (LTBPM, PT, pointer, extender, filler). <XJToBILabelWord> presents intonation information associated with words. Specifically, the perceived position of accent nucleus is annotated in this tag. <XJToBILabelBreak> is used to indicate break index (BI) labels. The labels "1", "2", and "3" correspond to the boundaries of words, accent phrases, and intonation phrases, respectively. The hierarchical structure of XML is different from that of the context set described in Sect. 4.2. Hence, we reconstruct the structure using the break index labels of <XJToBILabelBreak>. Specifically, we use labels 2 and 3 as the boundaries of the accent phrase and breath group, respectively. We can also flexibly modify these boundaries in spontaneous speech by using the information about pauses and disfluency assigned to this tag. <XJToBILabelPrm> is used as an auxiliary label to represent lexically irregular prominences which cannot be interpreted by other tags. <XJToBILabelMisc> tag is used to annotate extremely diverse spontaneous intonation that are not supported by XJToBI.

4. Contexts

4.1. Contexts for Japanese reading-style speech synthesis

Since the contextual factors depend on the languages and we use Japanese speech data for experiments, we explain the Japanese context set used in the demo script of HMM/DNN-based Speech Synthesis System (HTS) [18] as an example of the context of reading-style speech synthesis. The Japanese context set uses five hierarchical speech units: utterance, breath group, accent phrase, mora, and phone. To take the effect of adjacent units into account, we use the contextual factors of the previous and next speech units as well as those of current units. We refer to this context set as the *baseline context*. The baseline context also includes meta-information about the speaker (e.g., speaker ID, gender, age, and birth place) and the category of the talk.

4.2. Additional contexts for Japanese spontaneous speech synthesis

As additional context for spontaneous speech synthesis, tone labels, word units, clauses, phone prolongation, speaking style, disfluency, phoneme addition information has been examined in earlier work on HMM-based speech synthesis [11]. In this paper, we propose to use the following information contained in the XML file of CSJ as the additional contexts. Since most of the tags are categorical information, we encode such tags into one-hot features.

4.2.1. Tone label

It is difficult to model pitch movements in spontaneous speech using only accent-type information because they are much more complicated than those in reading-style speech. For example, a rise-fall type of boundary pitch movement is observed in utterances including dialogue acts such as turn-keeping and requesting an agreement. We utilize the labels of low, high, high-low, low-high, and high-low-high boundary pitch movements for the additional contextual factors based on the tag <XJToBILabelTone>. Moreover, We use other tone labels including "A", "pH", and "L", and add contextual factors about these labels to not only accent phrases but also phones, because the position of tone labels is critical information for pitch contours. Irregular pitch movements annotated in <XJToBILabelPrm> are also used. In addition, we use detail boundary information of <XJToBILabelBreak>, such as "2+p", which represents the existence of pause after the phrase.

4.2.2. Word

In the reading-style speech synthesis described in Sect. 4.1, word-unit features are omitted from contextual factors because they are not important in practice [19]. In this study, we incorporate word-unit features into the extended context to examine the effectiveness of such features for spontaneous speech. As the contextual factors of the word unit, we use the information of the part of speech, the conjugate type and form, and euphonic sounds included in the tags <LUW> and <SUW>.

4.2.3. Clause

Although IPU is a useful unit for spontaneous speech in which the end of a sentence does not often appear explicitly, it is often too short to model sequential information. As a grammatical unit related to a sentence, we can use clauses automatically determined by the transcription texts. Attributes ClauseBoundaryLabel in the tag <SUW> is related to the *strength* of the boundary classified into weak, strong, or absolute. The absolute boundary is equivalent to the sentence boundary of reading-style speech and it frequently becomes the utterance boundary. On the other hand, the weak boundary rarely becomes the utterance boundary. The types of clause boundary are determined according to the final word of the phrase. Also, manual correction of clause information is assigned to the CU_OperationSign in the tag <SUW>.

4.2.4. Important sentence

In the spontaneous speech, there is a difference in utterance between the important and non-important parts. In the CSJ core data, 177 talks are independently labeled by three workers for clause units in the top 10% and 50% of important parts of the speech, respectively, according to the tag SE_{Subject1—Subject2—Subject3}_{10p—50p} in the tag <SUW>.

4.2.5. Position in dialogue

In a dialogue, there may be a difference in utterance between when we respond and when we ask questions. In other words, the current position in the exchange of dialogue can affect speaking style.

4.2.6. Phone prolongation

When a speaker is thinking, surprised, or emphasizing, phones are often pronounced for longer than usual. Since this prolongation is not lexical, additional annotation is required. The labels about phone prolongation can be obtained from the attributes TagVLong and TagCLong in the tag <Mora>, which denote the prolongation of vowels and consonants, respectively.

4.2.7. Speaking style

When a speaker utters with expressiveness such as laughing and whispering, the speech waveform changes depending on the speaking style. This information is in the attributes of Tag{Whisper—Laughing} in the tags <Mora> and <NonLinguisticSound>.

4.2.8. Disfluency

Spontaneous speech includes many disfluent utterances. CSJ includes filled pauses, word fragments, and restatements as the labels of disfluency in the tag <TransSUW>. The use of these labels is expected to distinguish such disfluent utterances from normal utterances.

5. Experiments

5.1. Experimental condition

We used 195 talks corresponding to the whole academic presentation speech, simulated public speaking, and dialogues in the CSJ core data. The amounts of IPU were 57973, and the amounts of speech data were about 34.1 hour. The training data was segmented into IPUs. The context labels were created for each IPU using XML files in CSJ. We individually trained phone duration and acoustic feature models. The phone duration model used phone-level context as an input feature vector and predicted phone durations, whereas the acoustic feature model predicted frame-level acoustic features from corresponding input features.

We extracted the spectrum envelope, aperiodicity, and f_0 using WORLD [20] from 16 kHz waveforms. We converted the WORLD features into 187-dimensional acoustic features, which consisted of the 0–59th mel-cepstrum, log f_0, one-dimensional code aperiodicity, their delta and delta-delta features, and voiced/unvoiced flags. For the baseline context, the dimensions of input feature vectors were 360 and 364 for the duration and acoustic feature models, respectively.

The architecture of the DNN was a basic feedforward neural network. The number of hidden layers was five and each layer had 1024 hidden nodes. We used the ReLU activation functions and Adam optimizer [21] with a learning rate setting to 0.001. To avoid the overfitting problem, we used weight decay with a coefficient of 10^{-6} and a dropout with a rate of 0.5. The minibatch size was 1024 and we ran 20 epochs.

For a subjective evaluation test, we merged multiple synthetic IPUs into the speech samples whose durations were approximately 5 seconds because some of the IPUs were too short to evaluate. In evaluating tone label, word, clause contexts, 100 randomly selected speech samples were used as test speech samples. In evaluating the important sentence contexts, 100 randomly selected speech samples from 177 talks which had annotation of the context, were used as test speech samples. In evaluating the position-in-dialogue contexts, 100 randomly selected speech samples from dialogue talks were used as test speech samples. In evaluating phone prolongation, speaking style, dis-

Table 1: *XAB test comparing additional context with baseline one using DNN.*

Context	Selected rate	p-value
baseline vs. +tone label	44.3 % vs. **55.7** %	< 0.01
baseline vs. +word	43.0 % vs. **57.0** %	$< 10^{-3}$
baseline vs. +clause	43.7 % vs. **56.3** %	< 0.01
baseline vs. +important sentence	48.0 % vs. 52.0 %	0.33
baseline vs. +position in dialogue	49.3 % vs. 50.7 %	0.74
baseline vs. +phone prolongation	46.7 % vs. 53.3 %	0.10
baseline vs. +speaking style	45.0 % vs. **55.0** %	0.014
baseline vs. +disfluency	44.7 % vs. **55.3** %	< 0.01

Table 2: *XAB test comparing additional context with baseline one using LSTM.*

Context	Selected rate	p-value
baseline vs. +tone label	44.7 % vs. **55.3** %	< 0.01
baseline vs. +word	42.0 % vs. **58.0** %	$< 10^{-4}$
baseline vs. +clause	41.7 % vs. **58.3** %	$< 10^{-4}$
baseline vs. +important sentence	47.7 % vs. 52.3 %	0.25
baseline vs. +position in dialogue	46.7 % vs. 53.3 %	0.10
baseline vs. +phone prolongation	45.7 % vs. **54.3** %	0.033
baseline vs. +speaking style	43.7 % vs. **56.3** %	< 0.01
baseline vs. +disfluency	43.3 % vs. **56.7** %	< 0.01

fluency contexts, 50 randomly selected speech samples from labeled samples and 50 randomly selected speech samples from unlabeled samples were used as test speech samples since the labels appeared only rarely. The test speech samples were not included in the training data.

5.2. Subjective evaluation results

To evaluate the perceptual quality of synthetic speech, we performed subjective evaluation based on XAB tests, which are generally used to evaluate two samples and find the subtle difference of their quality. The participants on crowd-sourcing service first listened to the original reference X, and chose which of synthetic speech samples A and B was similar to the reference. For each test, the number of participants was 30, and each participant evaluated randomly chosen 10 speech segments. For each context, we examined the situations using both architectures of DNN and LSTM-RNN.

5.2.1. Comparison of additional context with baseline one

Table 1 shows the result where baseline and additional contexts are compared in the experiments using DNN. In tone label, word, clause, speaking style, disfluency, it is seen that the additional context gave significantly higher scores than the baseline context using DNN.

Table 2 shows the result where baseline and additional contexts are compared in the experiments using LSTM. In tone label, word, clause, phone prolongation, speaking style, disfluency, it is seen that the additional context gave significantly higher scores than the baseline context using LSTM.

5.2.2. Comparison of one or two additional contexts

To investigate the combination of the multiple contexts, we compared the cases that used one or two additional contexts. We picked up tone label, word, and clause, which are the context whose test speech samples are selected from all, and which gave significantly higher scores than the baseline context in Sect. 5.2.1. Table 3 and Table 4 show the results where DNN and LSTM are used as statistical models, respectively. It is seen from the results that the combination of additional contexts tended to increase the score. This indicates that the use of multiple contexts is effective in the DNN-based spontaneous speech synthesis.

Table 3: *XAB test comparing adding one context and two contexts using DNN.*

Context	Selected rate	p-value
+tone label vs. +tone label+word	42.0 % vs. **58.0** %	$< 10^{-4}$
+tone label vs. +tone label+clause	41.0 % vs. **59.0** %	$< 10^{-5}$
+word vs. +tone label+word	46.3 % vs. 53.7 %	0.072
+word vs. +word +clause	43.3 % vs. **56.7** %	< 0.01
+clause vs. +tone label+clause	47.0 % vs. 53.0 %	0.14
+clause vs. +word +clause	43.7 % vs. **56.3** %	< 0.01

Table 4: *XAB test comparing adding one context and two contexts using LSTM.*

Context	Selected rate	p-value
+tone label vs. +tone label+word	43.3 % vs. **56.7** %	< 0.01
+tone label vs. +tone label+clause	43.7 % vs. **56.3** %	< 0.01
+word vs. +tone label+word	46.7 % vs. 53.3 %	0.10
+word vs. +word +clause	44.7 % vs. **55.3** %	< 0.01
+clause vs. +tone label+clause	47.7 % vs. 52.3 %	0.25
+clause vs. +word +clause	43.7 % vs. **56.3** %	< 0.01

6. Conclusions

In this paper, we have investigated the effectiveness of additional contexts for DNN-based speech synthesis using the spontaneous speech corpus. The contexts are extracted from the rich tags of the corpus and used as the input of DNN- and LSTM-RNN-based speech synthesis frameworks. The subjective evaluation results using a large amount of training data showed that adding one context and combining two contexts tend to enhance the naturalness of spontaneous synthetic speech. We also showed the effectiveness of the several information such as word-unit features, clause, speaking style and disfluency, which were not shown to be effective in the conventional HMM-based study with a limited amount of training data [11]. For future work, we should examine the trade-off between the performance of synthetic speech and the cost of collecting additional contexts. Furthermore, we will investigate the automatic extraction of contexts from more variety of speech data. Moreover, we will compare the effectiveness of manually constructed contexts and that of automatically extracted contexts such as BERT [22] semantic vector embedding.

7. References

[1] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," 2017.

[2] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. INTERSPEECH*, 2017, pp. 4006–4010.

[3] Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, "Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language," in *Proc. ICASSP*, 2019, pp. 6905–6909.

[4] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, 2013, pp. 7962–7966.

[5] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for DNN-based speech synthesis," in *Proc. INTERSPEECH*, 2015, pp. 879–883.

[6] N. Hojo, Y. Ijima, and H. Mizuno, "DNN-based speech synthesis using speaker codes," *IEICE Transactions on Information and Systems*, vol. E101.D, no. 2, pp. 462–472, 2018.

[7] S. An, Z. Ling, and L. Dai, "Emotional statistical parametric speech synthesis using LSTM-RNNs," in *Proc. APSIPA ASC*, 2017, pp. 1613–1616.

[8] J. Lorenzo-Trueba, G. E. Henter, S. Takaki, J. Yamagishi, Y. Morino, and Y. Ochiai, "Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis," *Speech Commun.*, vol. 99, pp. 135–143, 2018.

[9] M. Wang, Z. Wu, X. Wu, H. Meng, S. Kang, J. Jia, and L. Cai, "Emphatic speech synthesis and control based on characteristic transferring in end-to-end speech synthesis," in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, 2018, pp. 1–6.

[10] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of japanese." in *Proc. LREC*, 2000, pp. 947–952.

[11] T. Koriyama, T. Nose, and T. Kobayashi, "On the use of extended context for HMM-based spontaneous conversational speech synthesis," in *Proc. INTERSPEECH*, 2011, pp. 2657–2660.

[12] Y. Yamashita, T. Koriyama, Y. Saito, S. Takamichi, Y. Ijima, R. Masumura, and H. Saruwatari, "DNN-based speech synthesis using abundant tags of spontaneous speech corpus," in *Proc. LREC*, 2020.

[13] 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. EUROSPEECH*, 1999, pp. 2347–2350.

[14] K. Maekawa, H. Kikuchi, and W. Tsukahara, "Corpus of spontaneous Japanese : Design, annotation, and XML representation," 2004.

[15] K. Maekawa, M. Yamazaki, T. Ogiso, T. Maruyama, H. Ogura, W. Kashino, H. Koiso, M. Yamaguchi, M. Tanaka, and Y. Den, "Balanced corpus of contemporary written Japanese," *Lang. Resour. Eval.*, vol. 48, no. 2, pp. 345–371, 2014.

[16] Y. Den, J. Nakamura, T. Ogiso, and H. Ogura, "A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation." in *Proc. LREC*, 2008, pp. 1019–1024.

[17] K. Maekawa, H. Kikuchi, Y. Igarashi, and J. Venditti, "X-JToBI: an extended J-ToBI for spontaneous speech," in *Proc. 7th ICSLP*, 2002, pp. 1545–1548.

[18] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. 6th ISCA Workshop on speech synthesis (SSW6)*, 2007, pp. 294–299.

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

[20] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. & Syst.*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.