SPONTANEOUS SPEECH SYNTHESIS WITH LINGUISTIC-SPEECH CONSISTENCY TRAINING USING PSEUDO-FILLED PAUSES

Yuta Matsunaga, Takaaki Saeki, Shinnosuke Takamichi, and Hiroshi Saruwatari

Graduate School of Information Science and Technology, The University of Tokyo, Japan.

ABSTRACT

We propose a training method for spontaneous speech synthesis models that guarantees the consistency of linguistic parts of synthesized speech. Personalized spontaneous speech synthesis aims to reproduce the individuality of disfluency, such as filled pauses. Our prior model includes a filled-pause prediction model and synthesizes filled-pause-included speech from text without filled pauses. However, inserting the filled pauses degrades the quality of the linguistic parts of the synthesized speech. This might be because filled-pause insertion tendencies differ between training and inference, and the synthesis model cannot represent connections between filled pauses and surrounding phonemes in inference. We, therefore, developed a linguistic-speech consistency training that guarantees the consistency of linguistic parts of synthetic speech with and without filled pauses. The proposed consistency training utilizes not only ground-truth-filled pauses but also pseudo ones. Our experiments demonstrate that this method improves the naturalness of the synthetic linguistic speech and the entire predicted-filled-pause-included synthetic speech.

Index Terms—spontaneous speech synthesis, filled pause, individuality, linguistic-speech consistency

1. INTRODUCTION

Voice cloning techniques that reproduce the individuality of a speaker’s speech [1–3] have been developed for the past few years, but they only handle fluent reading-style speech. For a speech synthesis system that can substitute human speech in situations where people speak spuriously (e.g., lectures), we need voice cloning that can deal with more human-like spontaneous speech including disfluency [4–7]. We, therefore, address personalized spontaneous speech synthesis, which reproduces the individuality of disfluency. Our particular focus is filled pauses (FPs), which are a kind of disfluency and play a significant role in human speech [8–13].

Some studies have addressed FP prediction on text [14–17], while others have examined spontaneous speech synthesis including FPs [18–20]. There are also a few studies on the effect of FPs on synthetic speech [18,22]: however, the evaluation of the individuality has been limited to the presence of FPs [20]. In our previous work, we constructed a spontaneous speech synthesis model with FP insertion and subjectively evaluated the effects of FPs in synthetic speech in terms of individuality [23]. The synthesis model consists of an FP prediction model and a text-to-speech synthesis model and synthesizes speech with FPs from text without FPs. The FP prediction model inserts FPs into text without FPs, and the text-to-speech model synthesizes FP-included speech from an FP-inserted text. While this model can synthesize FP-included speech from text without FPs, the FP insertion degrades the quality of the FP-included synthetic speech. Since the FP position is not unique, the phoneme features around the FP differ between the training and evaluation data, which means the model sometimes cannot represent the connections between FPs and surrounding phonemes and the quality therefore degrades. We could circumvent this by increasing the amount of training data, but this is difficult due to the high cost of FP annotation. Therefore, in the current study, we propose a method to synthesize FP-included speech without changing the linguistic-speech parts from the case without FPs.

Specifically, as shown in Fig. 1, we developed a training method that guarantees the consistency of the linguistic parts of synthetic speech with and without FPs. We train a synthesis model so that the linguistic parts of synthetic speech with FPs approximately match those of speech without FPs synthesized by a pre-trained model. This method is performed using not only ground-truth FPs but also pseudo ones sampled from the pre-trained FP prediction model so as to be robust against various FPs inserted when synthesizing FP-included speech. We subjectively evaluate our method in several settings and investigate how it improves the naturalness of synthesized speech with FPs in detail. Audio samples are available from our project page[1]. The key contributions of this work are as follows:

• To improve the quality of FP-included synthetic speech, we propose a training method that guarantees the linguistic-speech consistency of synthesized speech with and without FPs.

• Our experimental results demonstrate that the proposed method improves the naturalness of the entire synthetic speech with predicted FPs and the linguistic parts of all the FP-included synthetic speech.

2. SPONTANEOUS SPEECH SYNTHESIS WITH FILLED PAUSE INSERTION AND ITS CHALLENGES

2.1. Spontaneous speech synthesis with filled pause insertion

Figure 2 shows the spontaneous speech synthesis model consisting of an FP prediction model and a text-to-speech synthesis model [23].

---

1. https://sites.google.com/g.ecc.u-tokyo.ac.jp/yuta-matsunaga/publications/fp_consistency

[1] This work was supported by JST, Moonshot R&D Grant Number JP-MIPS2011.
sentences with and without FPs, using the model described in Sec. 2. Specifically, we compare intermediate representations of each of the five modules in FastSpeech 2. Changes in the representations are accumulated from the first (i.e., encoder) to the last (i.e., decoder) module, but the investigation of each module requires quantification of the effect of each one separately rather than cumulatively. Therefore, we replace intermediate representations of the module just before the targeted one inferred from FP-included sentences with those inferred from FP-removed ones. With this procedure, only FP parts of inputs of the module are different between the conditions with and without FPs, and by comparing these outputs, we can investigate the effect of only the targeted module. Figure 3 shows an example of this investigation process for pitch predictor. We can also investigate other modules (e.g., energy predictor) in the same manner.

4. LINGUISTIC-SPEECH CONSISTENCY TRAINING

We propose a training method for spontaneous speech synthesis models that guarantees consistency of linguistic parts between synthesized speech with and without FPs. First, we introduce a consistency guarantee for synthesized speech with ground-truth FPs (i.e., actually used by speakers) to ensure that the linguistic parts of synthesized speech with ground-truth FPs do not change from those of synthesized speech without FPs. We then extend this method for not only ground-truth FPs but also pseudo FPs that are predicted to be inserted with high probability. The loss function used in the proposed linguistic-speech consistency training is written as

\[ L = L_{\text{TTS}} + \alpha (L_{\text{GT-FP}} + \beta L_{\text{Pseudo- FP}}) \]  

where \( L_{\text{TTS}} \) is the loss used in the training of the text-to-speech synthesis model (i.e., FastSpeech 2 in this paper), and \( L_{\text{GT-FP}} \) and \( L_{\text{Pseudo- FP}} \) are the loss regarding the linguistic-speech consistency of synthesized speech with ground-truth FPs and with pseudo FPs, respectively. \( \alpha \) and \( \beta \) are hyperparameters to control the weight of the consistency loss and the ratio of the consistency loss for pseudo FPs to that for ground-truth ones, respectively.

4.1. Basic linguistic-speech consistency training

First, we introduce a consistency guarantee for ground-truth FPs. We pre-train a spontaneous speech synthesis model using FP-included text with ground-truth FPs. Then, we train a speech synthesis model (i.e., FastSpeech 2) on the pre-trained teacher model. This training procedure is expected to mitigate the degradation of the linguistic parts of synthesized speech with ground-truth FPs due to FP insertion.
spontaneous speech data as a teacher model. Then, we train a student model using the teacher model with fixed model parameters. In the training, the teacher and student models synthesize speech from the FP-removed and ground-truth-FP-included phoneme inputs, respectively. Next, we calculate $L_{\text{GT-FP}}$, the distance between intermediate representations of linguistic parts of speech without FPs synthesized by the teacher model and those of speech with ground-truth FPs by the student one. We arbitrarily decide which intermediate representation is used for consistency loss (i.e., the output of the energy predictor, the pitch predictor, etc.).

4.2. Extended linguistic-speech consistency training with pseudo-filled pauses

We next extend the consistency training for ground-truth FPs to that for pseudo FPs. We obtain a text with pseudo FPs by inserting sampled FPs (including no FP insertion) into FP-removed text on the basis of the predicted probabilities using the FP prediction model described in Sec. 4.1. Note that the prediction model inserts a most probable FP in inference, but in this training process, it inserts probabilistically sampled FPs. The student model then synthesizes speech with pseudo FPs. We calculate $L_{\text{Pseudo-FP}}$, the distance between intermediate representations of linguistic parts of speech without FPs synthesized by the teacher model and those of speech with pseudo FPs by the student one. The intermediate representations subject to consistency loss are the same as in the basic training. We also utilize an alternative method that randomly inserts FPs without using the pre-trained FP prediction model, which is described above. In this method, we randomly select a position of an FP in the sentence and an inserted FP word and insert one FP per sentence.

5. EXPERIMENTAL EVALUATION

5.1. Setting

We used JLecSponSpeech [23], a lecture speech corpus by two Japanese speakers. We downsampled speech to 22.05 kHz and utilized 97 sentences in subjective evaluations. Other preprocessing, model hyperparameters, and training conditions followed our previous work [23].

We investigated the causes of the quality degradation of the spontaneous speech synthesis model trained under the above conditions. To investigate the duration predictor, encoder/decoder, and pitch/energy predictor, we calculated relative errors of predicted values of linguistic parts with FP-included inputs compared to those with FP-removed inputs, cosine similarities of intermediate representations of linguistic parts between with and without FPs, and those of each predictor’s outputs, added with values of the residual path, respectively.

In the consistency training, we used the teacher model trained under the above conditions and trained the student model using the loss function described in Sec. 4. We used L1 loss for the consistency loss. In Eq. 1 we set $\alpha = 1.0$ in all the consistency training and $\beta = 0.0$ in the basic one. The hyperparameters and other training conditions of the student model are the same as the teacher model. For sampling pseudo FPs, we used the “non-personalized” FP prediction model from our previous work [24] and created 128 sentences with pseudo FPs inserted, and in training, we randomly selected one of them and calculated the consistency loss.

5.2. Preliminary investigation

First, we describe the results of the investigation of quality degradation in a spontaneous speech synthesis model. Since we obtained similar results for both speakers in JLecSponSpeech, we present the results of only one speaker here. Violin plots of distributions of relative errors and cosine similarities are shown in Fig. 5. The relative errors of the duration predictor outputs and the similarities of the encoder/decoder outputs are distributed around 0.0 and 1.0, respectively, indicating that these modules are not significantly affected by FP insertion. However, while most of the similarities of the pitch and energy predictor outputs are distributed around 1.0, we can also see peaks around 0.3 and 0.6, respectively, indicating that the pitch/energy predictor is affected by FP insertion to some extent, especially among several modules. We also investigated in which phoneme positions around FPs the output values are affected by FP insertion in the pitch/energy predictor, and the right side of Fig. 5 shows the results. We can see that the similarities are mostly distributed around 0.3 and 0.6 in the next phonemes by 1 or 2 from FPs, indicating that the features after adding the pitch and energy predictor outputs change by FP insertion in phonemes around FPs.

5.3. Evaluation of linguistic-speech consistency training

5.3.1. Basic consistency training

We next investigated the effect of consistency training for ground-truth FPs (Sec. 4.1). We used pitch/energy predictor outputs as the targets of consistency loss, on the basis of the results in Sec. 5.2. We conducted a preference AB test to compare the naturalness of synthesized speech samples from a conventional model trained without consistency loss and the proposed one trained with consistency loss for energy or for pitch and energy. We evaluated FP-included speech synthesized from the ground-truth-FP-included text (TrueFP). A total of 30 listeners participated, and each listener evaluated ten pairs of speech samples. In the following preference tests, the number of listeners and speech samples were the same as in this test.

Table 1. Preference scores of synthesized speech by a conventional model trained without consistency loss and the proposed one with basic consistency loss only for ground-truth FPs. Pitch and energy predictor outputs are targets of consistency loss.

| Target          | Score       | p-value  |
|-----------------|-------------|----------|
| -               | 0.450 vs. 0.550 | 1.43 × 10^{-2} | energy     |
| energy          | 0.547 vs. 0.453 | 2.22 × 10^{-2} | pitch and energy|
| energy          | 0.523 vs. 0.477 | 2.54 × 10^{-1} | pitch and energy|

Fig. 5. Results of preliminary investigation. Distributions of cosine similarities or relative errors of linguistic parts of outputs from each module in a speech synthesis model (left) and cosine similarities of pitch/energy predictor outputs in adjacent phonemes to FPs (right).
Table 2. Preference scores of consistency training with pseudo FPs sampled randomly or probabilistically with an FP prediction model.

| Pseudo FP | Score | p-value |
|-----------|-------|---------|
| Random    | 0.433 vs. 0.567 | $1.06 \times 10^{-3}$ |
| Probabilistic |         |         |

Table 3. Preference scores of speech synthesized by model using consistency loss only for ground-truth FPs and for both ground-truth FPs and pseudo FPs sampled with an FP prediction model.

| $\beta$ | Score | p-value |
|---------|-------|---------|
| 0.0     | 0.423 vs. 0.577 | $1.64 \times 10^{-4}$ |
| 2.0     | 0.547 vs. 0.543 | 3.38 $\times 10^{-2}$ |

Table 4(a). Preference scores by AB tests

| Utterance | Score | p-value |
|-----------|-------|---------|
| NoFP      | 0.497 vs. 0.503 | $8.71 \times 10^{-4}$ |
| TrueFP    | 0.483 vs. 0.517 | $4.15 \times 10^{-1}$ |
| FP part in TrueFP | 0.489 vs. 0.511 | $3.46 \times 10^{-1}$ |
| Linguistic part in TrueFP | 0.407 vs. 0.593 | $4.16 \times 10^{-6}$ |
| PredFP    | 0.410 vs. 0.590 | $9.16 \times 10^{-6}$ |
| FP part in PredFP | 0.547 vs. 0.453 | $7.32 \times 10^{-5}$ |
| Linguistic part in PredFP | 0.457 vs. 0.543 | 3.38 $\times 10^{-2}$ |

Table 4(b). Mean opinion scores

| Method | Utterance | Mean ± 95% conf interval |
|--------|-----------|--------------------------|
| Baseline | TrueFP  | 3.350 ± 0.125 |
|         | PredFP   | 3.280 ± 0.125 |
| Proposed| TrueFP   | 3.590 ± 0.121 |
|         | PredFP   | 3.537 ± 0.117 |

Next, we conducted an absolute evaluation by five-scale mean opinion score (MOS) tests for naturalness. Since this evaluation aims to investigate the naturalness of FP-containing synthetic speech, we only evaluated speech with ground-truth and predicted FPs. A total of 100 listeners participated in each test, and each listener evaluated 12 speech samples. The results are shown in Table 4(b). Compared to the baseline model, the proposed model improves the naturalness of synthetic speech with both ground-truth and predicted FPs.

5.3.2. Extended consistency training

Evaluation of sampling methods. We investigated the effect of using FP prediction models for sampling pseudo FPs in the extended method. We conducted a preference AB test to compare ground-truth-FP-included synthetic speech from the model trained with FPs sampled on the basis of the probabilities predicted by the FP prediction model and those sampled randomly without the prediction model, as described in Sec. 4.2. We set $\beta = 4.0$, which was subjectively evaluated to be the best in the preliminary experiment. The results are shown in Table 3. As we can see, the case using predicted probabilities is significantly preferred, thus demonstrating the effectiveness of using FP prediction models in consistency training with pseudo FP insertions.

Evaluation of consistency training for pseudo FPs. We conducted an AB test to compare the naturalness of the synthesized speech with FPs by the best model trained with consistency loss for both ground-truth FPs and pseudo FPs (i.e., trained with $\beta = 4.0$) and the model trained with consistency loss only for ground-truth FPs (i.e., $\beta = 0.0$). Table 3 shows the results. The case with $\beta = 4.0$ is evaluated significantly higher, indicating that the naturalness of the model with FPs is higher when trained with the consistency loss for both ground-truth and pseudo FPs at a ratio of 4.0 than when trained with the consistency loss only for ground-truth FPs.

Evaluation of best-proposed method. Finally, we evaluated the model trained with the conventional method and the best model trained with the proposed one. The conventional model was trained without consistency loss, as in the previous study. The proposed model with $\beta = 4.0$ was evaluated the best in preliminary experiments; thus, we used the model trained with consistency loss with $\alpha = 1.0$ and $\beta = 4.0$.

We conducted a comparative evaluation. To investigate the tendency of naturalness improvement of synthesized speech with FPs in detail, we evaluated seven kinds of speech samples: speech without FPs synthesized from text without FPs (NoFP), speech with FPs synthesized from text with ground-truth FPs (TrueFP), the FP part of speech with ground-truth FPs (FP part in TrueFP), the linguistic part of speech with ground-truth FPs (Linguistic part in TrueFP), speech with predicted FPs synthesized from text with predicted FPs (PredFP), the FP part of speech with predicted FPs (FP part in PredFP), and the linguistic part of speech with predicted FPs (Linguistic part in PredFP). The results are listed in Table 4(a). We can see that the proposed method improves the naturalness of the linguistic parts of speech with both ground-truth and predicted FPs. In synthetic speech with predicted FPs, the naturalness of the linguistic parts improves while that of the FP parts is degraded, but the naturalness of the entire speech improves. While the naturalness of the speech with predicted FP significantly improves, there is no significant difference in the evaluation of the entire speech for the synthesized speech with ground-truth FPs.

Next, we conducted an absolute evaluation by five-scale mean opinion score (MOS) tests for naturalness. Since this evaluation aims to investigate the naturalness of FP-containing synthetic speech, we only evaluated speech with ground-truth and predicted FPs. A total of 100 listeners participated in each test, and each listener evaluated 12 speech samples. The results are shown in Table 4(b). Compared to the baseline model, the proposed model improves the naturalness of synthetic speech with both ground-truth and predicted FPs.

5.3.3. Discussion

The results in Table 4(a) show that while the consistency guarantee for energy improved the naturalness, that for pitch and energy did not. This might be because, in spontaneous speech, pitch has a large difference between training and evaluation data, and the direct loss to pitch causes overtraining. In Table 4(a) while naturalness improved with PredFP, there was no significant difference with TrueFP, suggesting that this method trains a model to focus more on improving speech with predicted FPs than that with ground-truth ones. We can also see that while there was no significant difference in the FP parts of TrueFP, those of PredFP were degraded. This might be because the consistency loss is only for the linguistic parts, and thus the predicted FP parts cannot be trained. Our future work will focus on improving the naturalness by data augmentation of FP-included speech.

6. CONCLUSION

We investigated the quality degradation of a spontaneous speech synthesis model and proposed a training method for this model that guarantees consistency of the linguistic part between synthesized speech with and without FPs. Our experimental results demonstrated that the proposed method improves the naturalness of the entire synthetic speech with predicted FPs and the linguistic parts of synthetic speech with ground-truth and predicted FPs. However, FP parts of synthesized speech were still not improved. Therefore, our future work will focus on improving the quality of FPs by data augmentation of FP-included speech.


7. REFERENCES

[1] Qicong Xie, Xiaohai Tian, Guanghou Liu, Kun Song, Lei Xie, Zhiyong Wu, Hai Li, Song Shi, Haizhou Li, Fen Hong, Hui Bu, and Xin Xu, “The multi-speaker multi-style voice cloning challenge 2021,” in Proc. ICASSP, Jun. 2021, pp. 8613–8617.

[2] Merlijn Blaauw, Jordi Bonada, and Ryunosuke Daido, “Data efficient voice cloning for neural singing synthesis,” in Proc. ICASSP, May 2019, pp. 6840–6844.

[3] Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou, “Neural voice cloning with a few samples,” in Advances in NeurIPS, Dec. 2018, pp. 10019–10029.

[4] Jordi Adell, Antonio Bonafonte, and David Escudero-Mancebo, “On the generation of synthetic disfluent speech: local prosodic modifications caused by the insertion of editing terms,” in Proc. INTERSPEECH, Sep. 2008, pp. 2278–2281.

[5] Rasmus Dall, “Statistical parametric speech synthesis using conversational data and phenomena,” PhD dissertation of the University of Edinburgh, 2017.

[6] Yuzi Yan, Xu Tan, Bohan Li, Guangyan Zhang, Tao Qin, Sheng Zhao, Yuan Shen, Wei-Qiang Zhang, and Tie-Yan Liu, “Adaptive text to speech for spontaneous style,” in Proc. INTERSPEECH, Aug. 2021, pp. 4668–4672.

[7] Siyang Wang, Joakim Gustafson, and Éva Székely, “Evaluating sampling-based filler insertion with spontaneous tts,” in Proc. LREC, Jun. 2022, pp. 385–392.

[8] Yoshihiro Yamazaki, Yuya Chiba, Takashi Nose, and Akinori Ito, “Filler prediction based on bidirectional lstm for generation of natural response of spoken dialog,” in Proc. GCCE, Oct. 2020, pp. 360–361.

[9] Wilhelm J.M. Levelt, “Monitoring and self-repair in speech,” Cognition, vol. 14, no. 1, pp. 41–104, 1983.

[10] Schriberg Elisabeth, “Preliminaries to a theory of speech disfluencies,” Unpublished PhD dissertation, University of California, Berkeley, 1994.

[11] Howard Maclay and Charles E. Osgood, “Hesitation phenomena in spontaneous english speech,” WORD, vol. 15, no. 1, pp. 19–44, 1959.

[12] Herbert H Clark and Jean E Fox Tree, “Using uh and um in spontaneous speaking,” Cognition, vol. 84, no. 1, pp. 73–111, 2002.

[13] Jennifer E. Arnold, Michael K. Tanenhaus, Rebecca J. Altman, and Maria Fagnano, “The old and thee, uh, new: Disfluency and reference resolution,” Psychological Science, vol. 15, no. 9, pp. 578–582, 2004.

[14] Kengo Ohta, Masatoshi Tsuchiya, and Seiichi Nakagawa, “Construction of spoken language model including fillers using filler prediction model,” in Proc. INTERSPEECH, Aug. 2007, pp. 1489–1492.

[15] Marcus Tomalin, Mirjam Wester, Rasmus Dall, Bill Byrne, and Simon King, “A lattice-based approach to automatic filled pause insertion,” in Proc. DISS, Aug. 2015.

[16] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, “Fastspeech 2: Fast and high-quality end-to-end text to speech,” arXiv preprint arXiv:2006.04558, 2020.