Vol.2021-NL-251 No.24
Vol.2021-SLP-139 No.24
2021/12/3

Improving Intelligibility of Synthesized Speech in Noisy Condition with Dynamically Adaptive Machine Speech Chain

Sashi Novitasari[1,2,a)] Sakriani Sakti[3,1,2,b)] Satoshi Nakamura[1,2,c)]

Abstract: This paper focuses on the machine speech chain mechanism for improving the intelligibility of synthesized speech in noisy conditions. Our proposed text-to-speech synthesis (TTS) system synthesizes a speech by adapting to the situation. It will speak with a Lombard effect and high intelligibility in noisy conditions by processing auditory feedback that consists of speech-to-signal ratio (SNR) and automatic speech recognition (ASR) system loss. Our experiments show that auditory feedback improves TTS intelligibility in noisy environments.

Keywords: text-to-speech, machine speech chain inference, Lombard effect, dynamic adaptation

1. Introduction

Standard text-to-speech synthesis (TTS) systems suffer from speech intelligibility degradation in noisy places because they only learn how to speak without learning how to listen and understand the situation. On the other hand, humans speak louder to enhance their speech audibility in noisy places, a phenomenon known as the Lombard effect [1]. Lombard effect not only includes the change in the speech intensity but also the pitch and speed [2]. This could be done thanks to auditory feedback (speech chain mechanism) from mouth to ear that enables speakers to monitor their speech and improve it when necessary.

Inspired by the human speech chain mechanism, a machine speech chain [3] was proposed as a semi-supervised learning method for automatic speech recognition (ASR) and TTS systems by connecting them with a closed-feedback loop. This framework is able to improve ASR and TTS in the training with unpaired speech-text data. However in inference, the feedback connection is removed, thus, these systems are still unable to adapt to their environment.

In this work, we propose a machine speech chain mechanism for TTS inference in noisy places. Our TTS (Fig. 1(a)) synthesizes speech with a Lombard effect dynamically to improve the speech intelligibility given the auditory feedback in utterance level. Here the auditory feedback consists of the speech-to-noise ratio (SNR) as the speech and noise intensity measurement and the ASR loss as the speech intelligibility measurement in noise.

1 Nara Institute of Science and Technology, Ikoma-shi, 630–0192 Japan
2 RIKEN, Center for Advanced Intel- ligence Project AIP, Ikoma-shi, 630–0192 Japan
3 Japan Advanced Institute of Science and Technology, Nomi-shi, 923–1292 Japan
a) sashi.novitasari.si3@is.naist.jp
b) ssakti@jaist.ac.jp
c) s-nakamura@is.naist.jp

2. Proposed TTS in Speech Chain Framework

The proposed TTS is a multi-speaker Transformer TTS [4] extended with auditory feedback components (ASR-loss embedding and SNR embedding) and a variance adaptor (Fig. 1(b)). Given character sequence $x = [x_1, x_2, ..., x_S]$ with length S, TTS generates the corresponding speech Mel-spectrogram $y = [y_1, y_2, ..., y_T]$ with a length T and the prosody based on the auditory feedback in SNR (Z_{SNR}) and ASR loss (Z_{ASR}) embedding. In noisy situations, the proposed TTS performs a dynamic adaptation in several feedback iterations until the ASR loss converges. In this work, we constructed three TTS systems with different feedback configurations. All systems are trained using normal speech and Lombard speech in various noisy conditions.

2.1 TTS with SNR feedback

TTS generates the speech based on the text input and the SNR feedback embedding. In noisy conditions, TTS re-synthesizes speech to achieve a higher SNR ($\geq$ 20 dB) by integrating the SNR embedding into the encoder output and the decoder input. We implement the SNR embedding module using convolution network layers (Fig. 1(d)), which generate the embedding Z_{SNR} from noisy speech features y^{noisy}.

2.2 TTS with SNR-ASR feedback

TTS synthesizes a speech based on the text input and feedback as the SNR and ASR-loss embedding. ASR-loss embedding (Fig. 1(c)) represents the speech intelligibility measurement in the presence of noise. ASR-loss embedding vector Z_{ASR} is generated by transcribing the noisy TTS speech using ASR and feeding the ASR recognition loss to the embedding module.

2.3 TTS with SNR-ASR feedback and variance adaptor

In addition to the SNR and ASR-loss embedding modules, the proposed TTS also applies a variance adaptor [5] that guides the

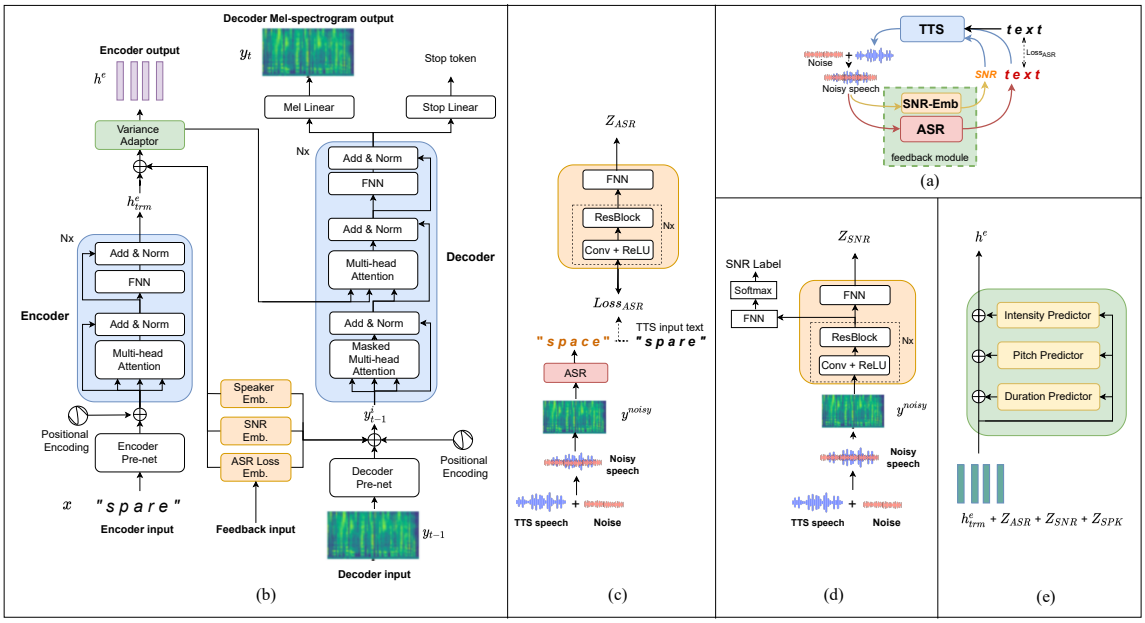

Fig. 1 (a) Proposed TTS in machine speech chain based on (b) Transformer architecture with encoder-decoder extended with (c) ASR-loss embedding, (d) SNR embedding, and (e) variance adaptor.

Table 1 Speech intelligibility measure (CER %) at different SNR levels using multi-condition training ASR.

System	Clean	SNR 0	SNR -10
Baseline TTS			
Standard TTS	18.32	70.54	77.07
+ Rule-based modification into Lombard speech	18.32	44.68	57.86
+ Fine tuning with Lombard speech (SNR 0)	13.19	32.71	53.35
+ Fine tuning with Lombard speech (SNR -10)	14.26	24.47	40.62
+ Fine tuning with Lombard speech (SNR 0 + SNR -10)	13.40	28.12	46.13
Proposed TTS			
TTS in speech chain framework	18.32	70.54	77.07
+ SNR feedback	11.58	22.82	42.00
+ SNR-ASR feedback	12.55	16.11	25.61
+ SNR-ASR feedback + variance adaptor	11.99	14.70	24.96
Topline (human natural speech)			
Natural speech	7.43	22.17	58.81
+ Rule-based modification into Lombard speech	7.43	13.24	15.15
Natural Lombard speech	7.43	11.46	20.46

prosody adaptation by predicting the speech intensity, duration, and pitch (Fig. 1(e)). Prosody attributes are predicted from the encoder Transformer output h^e_{trm} fused with feedback embedding and then utilized to produce the encoder output h^e.

3. Experiment

We experimented on Wall Street Journal (WSJ) dataset [6]. Here we also recorded natural Lombard speech in noisy conditions with a single male speaker. The noises in the recording were simulated by generating noises of SNR 0 dB and SNR -10 dB based on WSJ clean speech data first. Based on the prosody attribute changes observed in the recorded Lombard speech, we constructed synthetic Lombard WSJ speech by modifying the original WSJ speech pitch, duration, and intensity into a target SNR 20 dB. The original WSJ speech and the synthetic Lombard WSJ speech were used for TTS training and testing. Our baselines are (1) the standard TTS, (2) the standard TTS with the rule-based speech modification into the Lombard speech, and (3) the standard TTS fine-tuned to Lombard speech [7]. The rule-based speech modification into Lombard speech applied the same method as the synthetic Lombard WSJ construction method. The topline is the natural human speech.

The proposed SNR feedback and ASR feedback mechanism significantly improved the TTS speech intelligibility in noisy conditions, shown in Table 1. SNR feedback made the TTS aware of the environmental noise and speak with the Lombard effect, and ASR feedback helped the TTS improve the speech intelligibility further. Here the variance adaptor guided the speech prosody adaptation and resulted in a better intelligibility enhancement. These results reveal that the machine can also dynamically adapt in several loops; listen to its voice in a noisy environment and then speak louder to improve it. For further information on speech samples, see the following reference: https://sites.google.com/view/lombard-dynamic-tts/home.

4. Conclusions

We constructed a dynamically adaptive machine speech chain inference framework to improve TTS intelligibility in noisy conditions. Our results reveal that dynamic adaptation with auditory feedback is important not only for humans but also for machines to generate a highly audible speech in various conditions.

Acknowledgments Part of this work is supported by JSPS KAKENHI Grant Numbers JP21H05054 and JP21H03467.

References

[1] H. Lane and B. Tranel, "The Lombard sign and the role of hearing in speech," *Journal of Speech and Hearing Research*, vol. 14, no. 4, pp. 677–709, 1971.

[2] T. Letowski, T. Frank, and J. Caravella, "Acoustical properties of speech produced in noise presented through supra-aural earphones," *Ear and hearing*, vol. 14, pp. 332–338, 1993.

[3] A. Tjandra, S. Sakti, and S. Nakamura, "Machine speech chain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 976–989, 2020.

[4] M. Chen, X. Tan, Y. Ren, J. Xu, H. Sun, S. Zhao, and T. Qin, "MultiSpeech: Multi-speaker text to speech with Transformer," in *Proc. INTERSPEECH*, 2020, pp. 4024–4028.

[5] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," *ArXiv*, vol. abs/2006.04558, 2020.

[6] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[7] D. Paul, M. P. Shifas, Y. Pantazis, and Y. Stylianou, "Enhancing speech intelligibility in text-to-speech synthesis using speaking style conversion," in *Proc. INTERSPEECH*, 2020, pp. 1361–1365.