SELF-ADAPTIVE INCREMENTAL MACHINE SPEECH CHAIN FOR LOMBARD TTS WITH HIGH-GRANULARITY ASR FEEDBACK IN DYNAMIC NOISE CONDITION

Sashi Novitasari[1], *Sakriani Sakti*[2,1], *and Satoshi Nakamura*[1]

[1]Nara Institute of Science and Technology, Japan
[2]Japan Advanced Institute of Science and Technology, Japan

ABSTRACT

A common approach for text-to-speech (TTS) in noisy conditions is offline fine-tuning, which is generally utilized on static noises and predefined conditions. We recently proposed a self-adaptive TTS in machine speech chain inference that enables TTS to control its voices in statically and dynamically noisy environments based on auditory feedback from automatic speech recognition (ASR) and speech-to-noise ratio (SNR) recognition. However, that study only investigated the system on synthetic Lombard speech data. Furthermore, the ASR feedback was at a lower granularity based only on the loss of the positive character class. In this paper, we improve the self-adaptive TTS using character-vocabulary level ASR feedback at higher granularity, considering the losses in the positive and negative classes. We focus on a self-adaptive incremental TTS (Adapt-ITTS) with a short-term feedback mechanism that aims for low latency adaptation for dynamically noisy situations. In experiments, our proposed Adapt-ITTS successfully improved intelligibility in noisy conditions based on synthetic and natural Lombard speech data on the Wall Street Journal and Hurricane datasets, respectively.

Index Terms— self-adaptive, incremental, text-to-speech, machine speech chain, text-to-speech, Lombard effect

1. INTRODUCTION

In human spoken communication, speech production and perception are inseparable. This idea is reflected in the human speech chain mechanism, where auditory feedback is passed from mouth to ear, enabling humans to simultaneously speak and listen. This mechanism is essential not only for language acquisition but also for speech monitoring and self-adaptation during everyday communication. For example, auditory feedback makes us aware of environmental sounds that lead to speech adjustment. Notably, in noisy places, humans tend to increase their speaking effort, such as by speaking louder, to make the speech more intelligible. This phenomenon is known as the Lombard effect [1, 2].

Although auditory feedback is very critical for humans, in machines, unfortunately, feedback between text-to-speech (TTS) as the speaking component and an automatic speech recognition system (ASR) as the listening component remains limited. Over the last two decades, ASR and TTS have primarily been performed independently. Several recent works have shown that connecting ASR and TTS by end-to-end differentiable loss is advantageous [3, 4, 5, 6]. However, most of these studies only aimed at TTS that supports ASR as data augmentation during training. After the training is finished, ASR and TTS tasks are still performed separately.

Here we focus on TTS usage during the inference phase. Although a successful neural TTS can produce highly natural speech, it 'speaks' without being able to 'listen'. Furthermore, standard systems are commonly developed by assuming they are operating in ideal clean environments. Consequently, speech intelligibility often decreases in noisy environments. A common approach for such situations is fine-tuning the TTS offline using Lombard speech data [7, 8]. This approach, however, is generally focused on predefined static noise conditions, and self-adaptation during inference in more realistic dynamic environments remains difficult.

We recently proposed a self-adaptive TTS [9, 10] that enables TTS to speak and control its voices in statically and dynamically noisy environments with the non-incremental or incremental mechanism. Our mechanism was inspired by the human speech chain and the ASR-TTS connection in the basic machine speech chain [3, 11]. It allows TTS to dynamically adjust its speech style and improve its intelligibility according to various situations based on auditory feedback from the ASR loss and the speech-to-noise ratio (SNR) predictions. However, our previous study only investigated the system on synthetic Lombard speech data. Furthermore, the ASR feedback was at a lower granularity based only on the loss of the positive character class.

In this work, we propose an advanced self-adaptive TTS mechanism by leveraging ASR feedback at high granularity. Specifically, we use character-vocabulary level ASR feedback based on the losses in the positive and negative classes to enrich the feedback information and improve the TTS speech. We focus on the self-adaptive incremental TTS (Adapt-ITTS) with a short-term feedback mechanism (Fig. 1) that seeks low

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

latency adaptation for a dynamically noisy situation that may change immediately. Since the previous work was only based on a synthetic Lombard dataset, we also investigated TTS on natural Lombard speech in the Hurricane corpus [12].

2. ADAPT-ITTS

Adapt-ITTS incrementally synthesizes speech with a fixed incremental unit in a short-term auditory feedback mechanism (Fig. 1). To synthesize speech utterance $\boldsymbol{y} = [y_1, y_2, ..., y_T]$ from sentence text $\boldsymbol{x} = [x_1, x_2, ..., x_S]$, Adapt-ITTS first divides $\boldsymbol{x}$ into $\frac{S}{W}$ segments, each of which consists of W words. Then for each incremental step, Adapt-ITTS takes W-word text and produces corresponding W-word speech, and then slides the input windows to the next segment and repeats the procedure until it reaches the last segment. In addition to text, Adapt-ITTS also takes short-term auditory feedback generated based on the previous speech output. Such feedback aims to capture the system's performance and the environmental conditions, and Adapt-ITTS accordingly uses it to adapt its speech style (normal or Lombard) to improve the speech intelligibility. The feedback employed in Adapt-ITTS consists of ASR loss and SNR from the noisy speech for intelligibility measurement and also the speech power-context information.

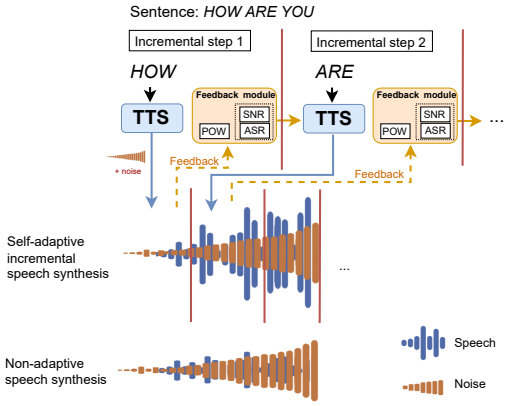

Fig. 1. Adapt-ITTS with low adaptation latency in dynamic noise condition and non-adaptive TTS

2.1. Architecture

2.1.1. Encoder-decoder

The Adapt-ITTS architecture (Fig. 2) is based on the Multi-Speech framework [13] for multi-speaker speech synthesis with the auto-regressive Transformer network. To enable the self-adaptation mechanism, it is extended with feedback modules and a variance adaptor [14, 15] that predicts the speech prosody. The feedback, which is represented as an embedding vector, is denoted as z_{ASR} for ASR loss embedding, z_{SNR} for SNR embedding, and z_{POW} for power-context embedding. These embeddings are generated independently through convolution layers.

All the feedback embeddings are combined with Transformer's encoder output h^e_{trm} along with the variance adaptor output to produce final encoder output h^e:

$$z = z_{SPK} + z_{SNR} + z_{ASR}, \tag{1}$$

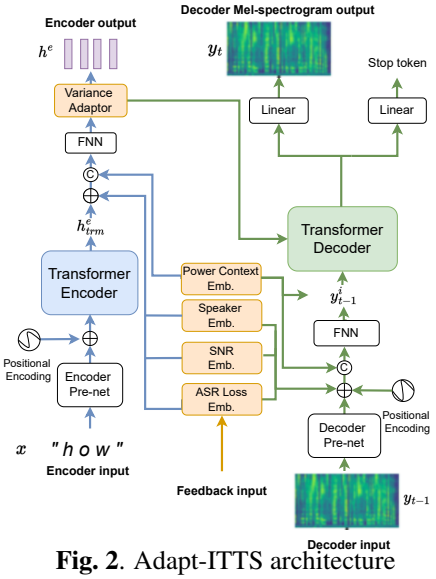

Fig. 2. Adapt-ITTS architecture

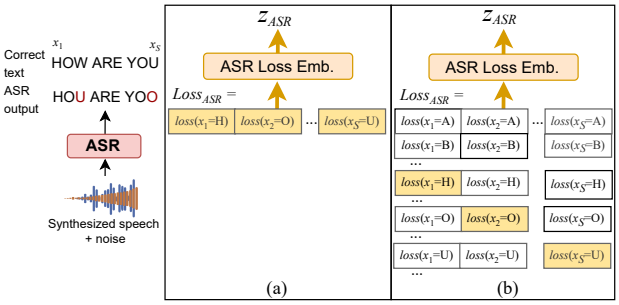

Fig. 3. ASR feedback at (a) low and (b) high granularity

$$h^e = Var\,Adaptor(FNN([h^e_{trm} + z, z_{POW}])), \tag{2}$$

where z_{SPK} represents the speaker embedding. h^e is then passed to the decoder for the cross-attention procedure. The feedback embeddings are also combined with the decoder input for the decoding process:

$$y^i_{t\text{-}1} = FNN([prenet(y_{t\text{-}1}) + PE + z, z_{POW}]). \tag{3}$$

2.1.2. ASR feedback

The ASR feedback represents the TTS speech intelligibility as an ASR loss. The ASR loss is generated by transcribing the noisy synthesized speech using an ASR system (teacher-forcing). As shown in Fig. 3, given a sequence of character-level ASR losses, the ASR-loss embedding module generates embedding vector z_{ASR} for the Adapt-ITTS input:

$$\boldsymbol{p}_x = p(\boldsymbol{x}|\boldsymbol{y}^{noisy})), \tag{4}$$

$$z_{ASR} = ASR\ Loss\ Embedding(Loss_{ASR}(\boldsymbol{x}, \boldsymbol{p}_x)), \tag{5}$$

where $\boldsymbol{y}^{noisy}$ is the synthesized speech combined with the noise. In this paper, we present two approaches on ASR loss calculation for feedback based on the granularity level.

(a) Low granularity

The ASR loss sequence, shown in Fig. 3(a), consists of the character-level losses of the positive character class based on the correct text:

$$Loss_{ASR} = [l_1, l_2, ..., l_S]. \tag{6}$$

Each character loss l_s is calculated with the multi-class cross entropy equation:

$$l_s(x_s, p_{x_s}) = -\sum_{c=1}^{C} \mathbb{1}(x_s = c) * \log p_{x_s}[c]. \quad (7)$$

(b) High granularity

Since the low-granularity ASR feedback only captures the loss on the positive class and discards the others, we newly propose the utilization of ASR feedback in higher granularity to preserve the finer information on the intelligibility, shown in Fig. 3(b). Here the ASR loss-embedding input is a sequence of the ASR loss that consists of the losses from the class of positive and negative characters according to the correct text:

$$Loss_{ASR} = [\boldsymbol{l}_1, \boldsymbol{l}_2, ..., \boldsymbol{l}_S], \quad (8)$$

$$\boldsymbol{l}_s = [l_{s1}, l_{s2}, ..., l_{sC}], \quad (9)$$

where C is the size of the character-vocabulary, $\boldsymbol{l}_s$ is the ASR loss at character index s in the transcription, and l_{sc} is the binary cross entropy loss of character c in the vocabulary at position s in the transcription:

$$l_{sc}(x_s^c, p_{x_s}) = - (x_s^c * \log p_{x_s}[c] + (1 - x_s^c) * \log (1 - p_{x_s}[c])), \quad (10)$$

where $x_s^c = 1$ for the positive character class and $x_s^c = 0$ for the negative class.

2.1.3. SNR feedback

The SNR feedback contains the SNR value between the synthesized speech and the environmental noises. This feedback also aims to capture the environmental conditions. Here we directly estimate the SNR from noisy speech without separating the acoustic sources through a neural network:

$$z_{SNR} = \textit{SNR embedding}(\boldsymbol{y}^{noisy}). \quad (11)$$

In our experiment, we first pre-trained the SNR feedback module for the SNR recognition task.

2.1.4. Power-context feedback

The power-context feedback represents the intensity or the power of the previously synthesized speech segment before it was induced with noise:

$$z_{POW} = \textit{Power-context embedding}(\boldsymbol{y}) \quad (12)$$

This feedback serves as a context cue for Adapt-ITTS to maintain or change the speech style based on environmental conditions, for example, continuing to produce Lombard speech while the noise still exists and changing the speech style when the conditions change. In our experiment, the power-context feedback module was pre-trained as a speech power recognition module before the Adapt-ITTS training.

2.2. Training and Inference

Adapt-ITTS is trained in a clean condition to produce normal speech and in noisy conditions to produce Lombard speech [10]. In training, ASR-loss embedding is generated from the synthesized speech. SNR embedding is generated from the audio in training material whose acoustic environment represents the environment condition before adaptation starts. The power-context embedding is generated from the clean speech in the training data.

In inference, Adapt-ITTS incrementally synthesizes speech by taking the feedback from the earlier incremental step (Fig. 1). Since the adaptation might be delayed by an incremental step due to the nature of the mechanism, Adapt-ITTS also applies a speech power post-adaptation to reduce the adaptation delay. It modifies the speech power incrementally on the M-ms units by first estimating the SNR and noise levels using the SNR feedback module and modifying the next M-ms segment to reach a certain SNR.

3. EXPERIMENTAL SETTING

3.1. Dataset

3.1.1. WSJ

We used the normal speech and synthetic Lombard speech data in the static noise condition based on the Wall Street Journal corpus [16] to initialize our model, which originally consisted of 81 hours of multi-speaker normal speech in clean condition. The Lombard speech was generated by modifying the original speech prosody attributes into Lombard-like prosody by a method that is identical as a previous work [10].

3.1.2. Hurricane

The Hurricane speech corpus [12], which was introduced in the Hurricane Challenge 2013 [17] for speech enhancement and synthesis in noisy conditions, consists of normal and Lombard speech recorded from a single, native British-English male speaker in the static noise condition. During training, we followed similar data partitions as in a related work [8]. We also created the noisy speech for SNR-embedding generation during training by combining noise sounds with normal speech. We used the same noise sounds (babble and white noises) and SNR (0 dB and -10 dB) as for our WSJ data. The SNR here is relative to the normal speech intensity in the Hurricane data. Since the data size was small, all the Hurricane-based TTS's parameters were initialized using TTS trained on the WSJ dataset.

3.2. Model Configuration

The Adapt-ITTS Transformer configuration followed the same configuration as in a previous work [10]. The Adapt-ITTS input was a character sequence, and the output was in 80 dimensions of Mel-spectrogram. The speech signal was generated from the predicted Mel-spectrogram using the CBHG (1-D Convolution Bank + Highway + bidirectional GRU) and Griffin-Lim algorithm, as in the Tacotron framework [18]. We composed the Adapt-ITTS incremental unit follows: the main input three words, the look-back input that included the previous ten words, and the look-ahead input that includes the next two words. The incremental unit for the power post-adaptation was 200 ms, and the target SNR was 20 dB. The configuration of the feedback modules was also identical as the one utilized in a previous work. All the feedback modules were trained using short speech segments to match the incremental unit in the Adapt-ITTS.

Table 1. TTS performance on WSJ dataset.

System	ASR CER(%) ↓			STOI (%) ↑	
	Clean	Static noise	Dynamic noise	Static noise	Dynamic noise
Baseline standard TTS (delay = 17 words)					
Base-TTS	18.32	73.81	46.25	43.88	72.05
Finetune-TTS	14.82	29.70	19.08	77.32	91.78
Adapt-TTS (delay = 17 words)					
ASR:Low granularity	13.52	24.88	17.79	81.83	90.60
ASR:High granularity	5.90	20.38	15.52	80.30	90.65
Adapt-ITTS (delay = 3 words)					
ASR:Low granularity	14.42	25.82	20.55	87.93	94.66
ASR:High granularity	14.04	23.19	17.48	89.04	97.20
Topline					
Natural speech	7.43	15.96			

Table 2. TTS performance on Hurricane dataset.

System	ASR CER(%) ↓			STOI (%) ↑	
	Clean	Static noise	Dynamic noise	Static noise	Dynamic noise
Baseline standard TTS (delay = 7 words)					
Finetune-TTS	8.95	32.85	25.95	73.12	88.21
Adapt-TTS (delay = 7 words)					
ASR:Low granularity	10.53	24.91	21.46	82.29	93.73
ASR:High granularity	9.36	23.98	20.58	82.33	93.77
Adapt-ITTS (delay = 3 words)					
ASR:Low granularity	13.77	28.82	28.24	79.91	85.86
ASR:High granularity	14.92	27.86	28.27	80.37	86.58
Topline					
Natural speech	6.85	22.28	16.05	73.67	88.68

4. RESULT AND DISCUSSION

We evaluated the TTS speech intelligibility in the static and dynamic noise conditions based on the ASR character error rate (CER) and the short-term objective intelligibility measure (STOI) [19]. The ASR CER was calculated by transcribing noisy speech using an utterance-level ASR in the Speech-Transformer framework [20] trained on clean and noisy speech based on the WSJ and Hurricane data. Then STOI measures the temporal envelope correlation between the speech signals before and after they are combined with noise. Our baseline was a standard non-incremental TTS that was trained on normal speech (Base-TTS) and fine-tuned on Lombard speech data [8] (Finetune-TTS). The baseline's architecture was identical as the Adapt-ITTS without feedback modules or a variance adaptor. We also compared the Adapt-ITTS to the self-adaptive non-incremental TTS (Adapt-TTS)[10] that synthesized the speech using utterance-level feedback in a re-speaking manner. The Adapt-TTS architecture was also identical as Adapt-ITTS's without using the power-context embedding module. The topline was natural human speech.

4.1. Static Condition

We evaluated Adapt-ITTS in the clean and static noise conditions shown in Table 1 for the WSJ data results and Table 2 for the Hurricane data results. The experiment in clean conditions was conducted without noise, and the noisy condition experiment was done using noise from the SNR 0 dB or -10 dB conditions according to the dataset.

In static conditions, the high-granularity ASR feedback improved the Adapt-ITTS performance. ASR CER was reduced by 2.63% on WSJ and 0.96% on Hurricane. The performance difference between the datasets might have been affected by the acoustic characteristics and the length of the speech utterances. Adapt-ITTS performed with lower input latency and adaptation latency, which in our setting was three words of approximately 1 sec. On the other hand, although Adapt-TTS achieved the best intelligibility, it requires a higher delay than Adapt-ITTS due to the utterance-level feedback mechanism. Adapt-ITTS with high-granulated ASR feedback successfully approached the Adapt-TTS speech intelligibility with shorter delay.

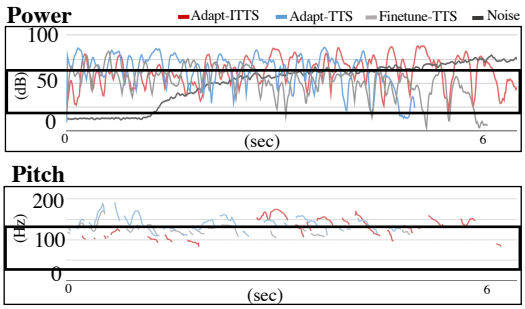

Fig. 4. Speech power and pitch in dynamic noise condition

4.2. Dynamic Condition

In this experiment, the noise intensity was increased or decreased between the clean and noisy conditions. The highest noise intensity was based on noise from the SNR-10 dB condition. For the baseline and the topline, the noise was combined with the static Lombard speech since no dynamic Lombard speech was available.

The high-granulated ASR feedback successfully improved the Adapt-ITTS performance in dynamically noisy conditions, in which the ASR CER could be reduced by 3.07%. The Adapt-ITTS speech prosody adapted accordingly to the noise change, as shown in Fig. 4 from the model trained on the Hurricane dataset. In this example with the increasing noise, Adapt-ITTS initially produced normal speech and then, after the noise power reached a certain point, it produced Lombard speech with the increased speech power and pitch. On the other hand, the Adapt-TTS and Finetune-TTS spoke loudly initially, but they were unable to adapt during the middle of utterance to cope with the increasing noise. Although the best intelligibility was achieved by the Adapt-TTS, it has to wait for one utterance to finish. The Adapt-ITTS, on the other hand, was more realistic because it adapted within 1 sec and improved the speech intelligibility.

5. CONCLUSION

We proposed an Adapt-ITTS with high-granulated ASR feedback for a self-adaptive speech synthesis in noisy conditions. Adapt-ITTS adapts the speech style based on noise conditions in real time using short-term feedback in an incremental mechanism. The utilization of the proposed ASR feedback successfully improved Adapt-ITTS intelligibility in noisy conditions. Examples in https://sites.google.com/view/adapt-lombard-tts/home.

6. REFERENCES

[1] Harlan Lane and Bernard 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] Maëva Garnier, Nathalie Henrich, and Danièle Dubois, "Influence of sound immersion and communicative interaction on the Lombard effect," *J. Speech Lang. Hear. Res.*, vol. 53, no. 3, pp. 588–608, 2010.

[3] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura, "Listening while speaking: Speech chain by deep learning," in *Proc. IEEE ASRU*, 2017, pp. 301–308.

[4] Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro Moreno, Yonghui Wu, and Zelin Wu, "Speech recognition with augmented synthesized speech," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 996–1002.

[5] Shigeki Karita, Shinji Watanabe, Tomoharu Iwata, Marc Delcroix, Atsunori Ogawa, and Tomohiro Nakatani, "Semi-supervised end-to-end speech recognition using text-to-speech and autoencoders," in *IEEE ICASSP*, 2019, pp. 6166–6170.

[6] Zhehuai Chen, Andrew Rosenberg, Yu Zhang, Heiga Zen, Mohammadreza Ghodsi, Yinghui Huang, Jesse Emond, Gary Wang, Bhuvana Ramabhadran, and Pedro J. Moreno, "Semi-Supervision in ASR: Sequential MixMatch and Factorized TTS-Based Augmentation," in *Proc. Interspeech 2021*, 2021, pp. 736–740.

[7] Tuomo Raitio, Antti Suni, Martti Vainio, and Paavo Alku, "Analysis of HMM-based Lombard speech synthesis," in *Proc. Interspeech 2011*, 2011, pp. 2781–2784.

[8] Dipjyoti Paul, Muhammed P.V. Shifas, Yannis Pantazis, and Yannis Stylianou, "Enhancing speech intelligibility in text-to-speech synthesis using speaking style conversion," in *Proc. Interspeech 2020*, 2020, pp. 1361–1365.

[9] Sashi Novitasari, Sakriani Sakti, and Satoshi Nakamura, "Dynamically adaptive machine speech chain inference for TTS in noisy environment: Listen and speak louder," in *Proc. Interspeech 2021*, 2021, pp. 4124–4128.

[10] Sashi Novitasari, Sakriani Sakti, and Satoshi Nakamura, "A machine speech chain approach for dynamically adaptive lombard tts in static and dynamic noise environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2673–2688, 2022.

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

[12] Martin Cooke, Catherine Mayo, and Cassia Valentini-Botinhao, "Hurricane natural speech corpus, [sound]," `https://doi.org/10.7488/ds/140`, 2013.

[13] Mingjian Chen, Xu Tan, Yi Ren, Jin Xu, Hao Sun, Sheng Zhao, and Tao Qin, "MultiSpeech: Multi-speaker text to speech with transformer," in *Proc. Interspeech 2020*, 2020, pp. 4024–4028.

[14] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "FastSpeech: Fast, robust and controllable text to speech," in *Advances in Neural Information Processing Systems*, 2019, pp. 3165–3174.

[15] Yi Ren, C. Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and T. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," *ArXiv*, vol. abs/2006.04558, 2020.

[16] Douglas B Paul and Janet M Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*, 1992, pp. 357–362.

[17] Martin Cooke, C Mayo, and Cassia Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane Challenge," in *Proc. Interspeech 2013*, 2013, pp. 3552–3556.

[18] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech 2017*, 2017, pp. 4006–4010.

[19] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[20] L. Dong, S. Xu, and B. Xu, "Speech-Transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. ICASSP 2018*, 2018, pp. 5884–5888.