Interspeech 2022
18-22 September 2022, Incheon, Korea

DENT-DDSP: Data-efficient noisy speech generator using differentiable digital signal processors for explicit distortion modelling and noise-robust speech recognition

Guo Zixun[1], *Chen Chen*[1], *Chng Eng Siong*[1]

[1]Nanyang Technological University, Singapore

zguo008@e.ntu.edu.sg

Abstract

The performances of automatic speech recognition (ASR) systems degrade drastically under noisy conditions. Explicit distortion modelling (EDM), as a feature compensation step, is able to enhance ASR systems under such conditions by simulating the in-domain noisy speeches from the clean counterparts. Yet, existing distortion models are either non-trainable or unexplainable and often lack controllability and generalization ability. In this paper, we propose a fully explainable and controllable model: DENT-DDSP to achieve EDM. DENT-DDSP utilizes novel differentiable digital signal processing (DDSP) components and requires only 10 seconds of training data to achieve high fidelity. The experiment shows that the simulated noisy data from DENT-DDSP achieves the highest simulation fidelity compared to other baseline models in terms of multi-scale spectral loss (MSSL). Moreover, to validate whether the data simulated by DENT-DDSP are able to replace the scarce in-domain noisy data in the noise-robust ASR tasks, several downstream ASR models with the same architecture are trained using the simulated data and the real data. The experiment shows that the model trained with the simulated noisy data from DENT-DDSP achieves similar performances to the benchmark with a 2.7% difference in terms of word error rate (WER). The code of the model is released online[1].

Index Terms: explicit distortion modelling, DDSP, noise-robust ASR

1. Introduction

The end-to-end automatic speech recognition (ASR) systems have achieved a remarkable success, yet, the systems are prone to noisy conditions. Several datasets [1, 2, 3, 4] containing clean speeches, noisy speeches and the corresponding transcribed texts have been collected and utilized to boost ASR performances under noisy conditions. Yet, such datasets are usually scarce. Explicit distortion modelling (EDM) [5], as an alternative, is able to simulate noisy speeches from their clean counterparts. However, distortion models with high simulation fidelity are hard to be obtained.

To achieve EDM, traditional methods such as parallel model combination (PMC) [6] and vector taylor series (VTS) [7, 8, 9] are widely used to estimate the distortion model before the deep learning era. An intuitive method is proposed in [10] where static noise within the limited collected noisy data are aggregated and added to the clean speeches. Another method [11] adds pre-recorded noise to the clean speeches and passes them through pre-defined codecs to obtain the simulated noisy data. These two static methods [10, 11], however, are untrainable thus cannot be adapted to other circumstances. Recently, GAN-based methods [12, 13] are utilized to achieve EDM and have obtained promising results. We find SimuGAN [12] the closest match to our work. It operates directly on magnitude spectrograms and uses GAN and contrastive learning methods to distort the clean spectrograms. Yet, GAN-based methods contain large amounts of parameters making the distortion models unexplainable and uncontrollable.

Recently with the advent of the differentiable digital signal processing (DDSP) [14], traditional digital signal processors (DSP) become differentiable and trainable and have seen success in the field of speech synthesis [15, 16]. Inspired by the success of DDSP in the synthesis domain, we find DDSP a viable solution to achieve EDM. In this paper, we focus on simulating the VHF/UHF transmitted data (e.g., air traffic control). By comparing the characteristics of the VHF/UHF transmitted data with their clean counterparts, we observe that the VHF/UHF transmitted data are distorted, compressed to a fixed dynamic range, equalized and also contain colored noise. We hence find DDSP capable of achieving such exact conversion.

In this work, we propose a data-efficent noisy speech generator using novel DDSP components (DENT-DDSP) to simulate the VHF/UHF transmitted data. To our best knowledge, DENT-DDSP is the very first model that utilizes DDSP to achieve explicit distortion modelling. It only requires 10 seconds of parallel training data, yet, the trained distortion model achieves high fidelity simulation. As a result, the VHF/UHF transmitted data can easily be simulated from clean speeches. The simulated data can be further used for other downstream tasks (e.g., noise-robust ASR, speech enhancement).

More specifically, DENT-DDSP consists of two signal chains: an audio signal chain and a noise signal chain. Each signal chain consists of trainable DDSP components connected in series. The audio signal chain is able to distort, compress and equalize the clean speeches. The noise signal chain is able to transform the input white noise to the desired spectral characteristics. The outputs from the two signal chains are added to form the final simulated noisy speech. In the audio signal chain, two novel DDSP components are proposed: waveshaper and computionally-efficient dynamic range compressor (DRC). The experiment has shown that the proposed computationally-efficient DRC drastically improves the computational efficiency while maintaining the simulation fidelity.

Extensive experiments have shown that DENT-DDSP achieves the highest simulation fidelity among all distortion models in terms of MSSL. Moreover, it has shown strong generalization abilities over unseen training data and outperforms other GAN-based models. To further validate whether the labor-intensive way of collecting the VHF/UHF transmitted data can be replaced by simulating these data from DENT-

[1]https://github.com/guozixunnicolas/DENT-ddsp

Copyright © 2022 ISCA

10.21437/Interspeech.2022-763

DDSP, a downstream noise-robust ASR model trained using simulated data from DENT-DDSP is benchmarked against a same model trained with real noisy data. The experiment has shown that the model trained with simulated data from DENT-DDSP has achieved similar WER to the benchmark model with 2.7% difference and has outperformed the best baseline model by 7.3% in terms of WER.

2. Model Architecture

The model architecture of DENT-DDSP is shown in Figure 1. DENT-DDSP contains 2 parallel signal chains. The audio signal chain contains a waveshaper, a computationally-efficient DRC and an equalizer and receives clean speech: $s_{in}(t)$ as input. The noise signal chain contains an equalizer and receives white noise: $n_{in}(t)$ as input. In each signal chain, the input signal will pass through each DDSP component in series. The simulated noisy speech: $s_{simulated}(t)$ will be formed by adding the audio signal chain output: $s_{out}(t)$ and the weighted noise signal chain output:$n_{out}(t)$ following Equation 1. λ is a weighting and non-trainable parameter and is set to 1 during training. However, it can be adjusted during the generation phase to simulate noisy data with different SNRs in order to achieve data augmentation. A spectral loss function: MSSL (multi-scale spectral loss) [14] will be calculated between the simulated noisy speech: $s_{simulated}(t)$ and the real noisy data to update the model's parameters via backpropagation. We calculate MSSL using the following FFT sizes: 2048, 1024, 512, 256, 128, 64, to reflect the spectral distance in different spectral resolutions.

$$s_{simulated}(t) = s_{out}(t) + \lambda \cdot n_{out}(t) \tag{1}$$

We then formulate the DDSP components in the signal chains and introduce their functionalities. For simplicity, we represent each component's input as $x(t)$ and output as $y(t)$. $x_{dB}(t)$ and $y_{dB}(t)$ represents the input and output in the dB scale.

2.1. Waveshaper

The waveshaper is able to introduce a distortion effect to the input signal as shown in Equation 2. The distortion gain $g_{distort}$ represents the amount of distortion added to the input signal. With $g_{distort}$ close to 0, the transfer function of a waveshaper is almost linear hence little distortion is applied. However, with an increasing $g_{distort}$, the transfer function becomes non-linear and saturates quickly. Such non-linearity introduces the distortion to the input signal. The distortion gain $g_{distort}$ is set to be a trainable parameter.

$$y(t) = \frac{2}{\pi}\arctan(g_{distort} \cdot \frac{\pi}{2} \cdot x(t)) \ , \ g_{distort} > 0 \tag{2}$$

2.2. Computationally-efficient DRC

Dynamic range compressors(DRC) [17] are widely used in music production to limit the dynamic ranges of different audio tracks. By observation, the VHF/UHF transmitted data are compressed to a fixed dynamic range in the time domain. We hence find DRC a suitable DDSP component to achieve such limiting behaviour. A detailed survey [17] has introduced different kinds of DRCs and we have chosen a hard-knee DRC as our DDSP backbone. We then formulate our novel computationally-efficient DRC in Equation 3-7.

Equation 3 calculates the amplitude reduction gain $g(t)$ based on the compression threshold T and the compression ratio R. Equation 4-6 shows the proposed efficient implementation of gain smoothing using a novel companding operation. The amplitude reduction gain $g(t)$ is firstly downsampled via a linear interpolation by a downsampling factor ds_factor in Equation 4. The downsampled gain function $g_d(t)$ is then smoothed by α_A and α_R which represents attack and release time constant respectively in Equation 5. The smoothed gain function $g_{ds}(t)$ is then upsampled with overlapping hanning windows by an upsampling factor us_factor in Equation 6. ds_factor and us_factor are set to be equal such that $g_{es}(t)$ will have the same shape as $g(t)$. $g_{ds}(t)$ is the main performance bottleneck in gain calculation since the operation is auto-agressive and recursive with time complexity of $O(L/ds_factor)$ where L is the total number of audio samples. Hence, the time complexity will increase linearly with a decreasing ds_factor. We will later prove the companding operation in Equation 4 and 6 drastically improve the efficiency while maintaining the simulation quality in Section 3.3. Finally, the output $y_{dB}(t)$ is obtained in Equation 7 with an additional makeup gain g_{makeup}. The following parameters are trainable: T, R, α_A, α_R, g_{makeup}.

$$g(t) = \begin{cases} 0 & x_{dB}(t) \leq T \\ \dfrac{x_{dB}(t) - T}{R} & x_{dB}(t) > T \end{cases} \tag{3}$$

$$g_d(t) = downsample(g(t), ds_factor) \tag{4}$$

$$g_{ds}(t) = \begin{cases} \alpha_A g_{ds}(t-1) + (1-\alpha_A) g_d(t) & g_d(t) > g_{ds}(t-1) \\ \alpha_R g_{ds}(t-1) + (1-\alpha_R) g_d(t) & g_d(t) \leq g_{ds}(t-1) \end{cases} \tag{5}$$

$$g_{es}(t) = upsample(g_{ds}(t), us_factor) \tag{6}$$

$$y_{dB}(t) = x_{dB}(t) + g_{es}(t) + g_{makeup} \tag{7}$$

2.3. Equalizer(EQ)

By observation, the VHF/UHF transmitted data are band-limited and equalized in the frequency domain. To achieve this, we adopt the equalizer (EQ) implementation from [14] where the EQ is implemented using a linear time invariant FIR filter. For each EQ, the trainable parameters are set to be the magnitude of the filter frequency response: FR_{mag} and the number of the frequency bins are set to be 1000. Additionally, for the EQ in the noise signal chain, the noise amplitude is made trainable in order to adjust the volume of the filtered noise.

3. Experiments and results

3.1. Dataset

The Robust automatic transcription of speech (RATS) project [1] has collected a parallel dataset which contains clean speeches, the corresponding VHF/UHF transmitted noisy speeches and the transcribed texts. To obtain the dataset, pre-recorded conversational speeches are broadcasted over 8 radio channels and the transmitted noisy audio are captured concurrently. We select Channel A from the RATs dataset which contains 57.4 hours of data for training and testing. To address the data scarcity problem mentioned in Section 1, only less than 60 seconds of parallel audio are selected as training data. During training, the data are batched into 1-second chunks.

3.2. Experiment setup

3.2.1. Training data selection and model comparison

To obtain the best performing and most efficient distortion model, several DENT-DDSP models are trained and compared using different amounts of parallel training data from 10 to 60

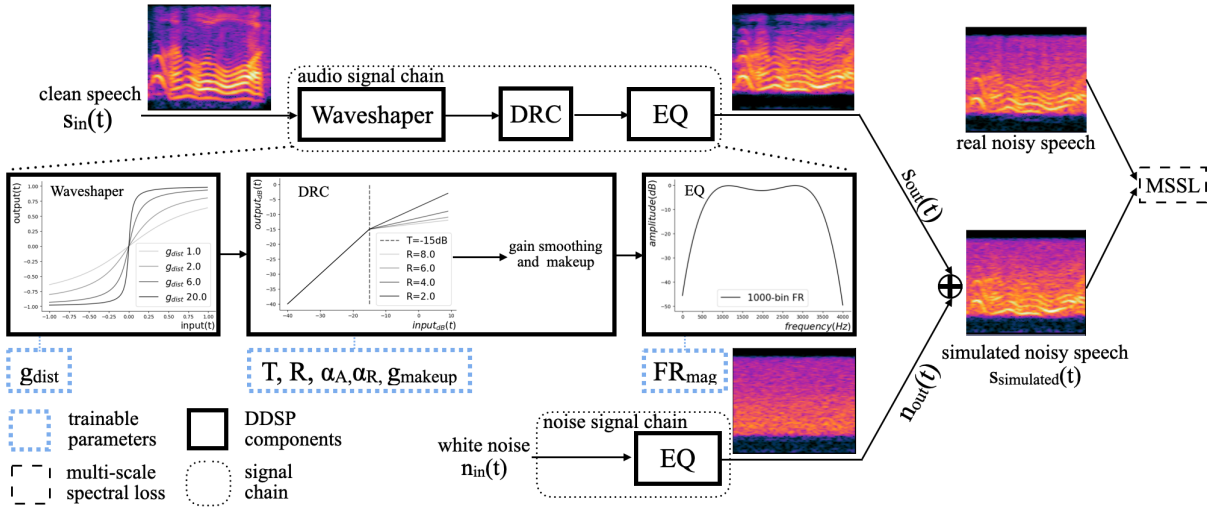


Figure 1: *DENT-DDSP architecture. DRC stands for dynamic range compressor and EQ stands for equalizer. The parameters of the DDSP components are trainable and controllable.*

seconds with various speech to total ratio: $s2t$. $s2t$ is defined as the duration of active speech over the total duration in each 1-second chunk. Active speech duration is calculated based on the clean audio energy in a sliding window with a threshold of $50dB$. By analyzing the dataset, with $s2t < 0.4$, the data contains mostly non-speech audio or silence. Hence, we only select data with $0.4 \leq s2t < 1.0$. To acquire the desired training data, the 1-second data chunks with the desired $s2t$ are randomly chosen and aggregated to the desired total duration. To evaluate the simulation fidelity, testing clean speeches are input to the trained distortion model and the MSSL[14] is calculated between the real and simulated noisy data. We compare the proposed DENT-DDSP with the following baseline distortion models in terms of simulation fidelity:

- Clean augment model[10]: simulated noisy speeches are obtained by adding the aggregated stationary noise from RATs Channel A to the clean speeches.
- G.726 augment model[11]: The clean speeches are first passed through the G.726 codec. The aggregated stationary noise from RATs Channel A is added subsequently to the codec outputs.
- Codec2 augment model[11]: Codec choice in the G.726 augment model is replaced to Codec2 with 700 bits/s.
- SimuGAN[12]: simulated noisy speeches are obtained by feeding the clean speeches to a trained SimuGAN.

3.2.2. Noise-robust automatic speech recognition (ASR)

A downstream noise-robust ASR task is performed to validate the effectiveness of DENT-DDSP. The 57.4-hour parallel data from Channel A are split into 3 folds [18]: 44.3-hour data for training; 4.9-hour data for validation and 8.2-hour data for testing. We adopted the conformer-based [19] dual-path ASR system from [12] with 12 conformer layers in the encoder and 6 transformer layers in the decoder. Such dual-path ASR architecture is specifically designed for noise-robust ASR [20, 21]. We pre-train a language model [22] with 2 RNN layers using the existing transcribed texts and utilize it during text decoding. To benchmark the performance, a benchmark model is firstly trained with real noisy data. To address the data scarcity problem aforementioned in the real world, the benchmark model only incorporates the single-path setting in [12] where only real noisy data and the transcribed text are input to the network . Another set of model is trained using the simulated noisy data from various distortion models using the dual-path setting in [12]. These sets of models will be tested on a common test set consists of real noisy data and the performances in terms of WER will be compared.

3.3. Results

3.3.1. Effect of different amounts of training data with various $s2t$ and model comparison

Table 1 shows the MSSL and between the real and simulated audio using different amounts of training data with various $s2t$. In general, MSSL decreases with fewer amounts of training data and increasing $s2t$. The best DENT-DDSP model is hence obtained using 10 seconds of training data with $0.8 \leq s2t < 1.0$. Comparing DENT-DDSP with other distortion models, we observe that DENT-DDSP is trainable, contains much fewer trainable parameters and requires much fewer amount of training data. Moreover, noisy audio simulated by DENT-DDSP has the lowest MSSL which indicates DENT-DDSP has the highest simulation fidelity among all baseline models. Readers are strongly encouraged to listen to the simulated noisy speeches and compare them with the real noisy data from[2]. To visualize the simulation fidelity, example spectrograms of the clean audio, audio simulated from various distortion models and the real noisy data are plotted in Fig 2.

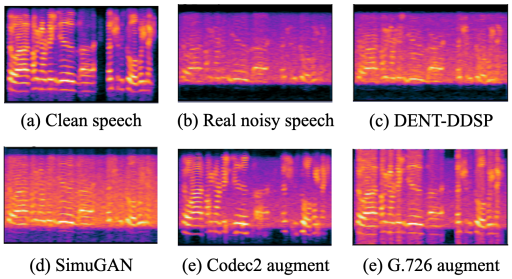
3.3.2. Effectiveness of the companding operation in the proposed computationally-efficient DRC

Table 2 shows the effectiveness of the proposed companding operation in the novel computationally-efficient DRC. With decreasing ds_factors using the best DENT-DDSP obtained, the training time increases linearly, however, without obvious improvements on the MSSL. This justifies the effectiveness of the companding operation which is able to increase computational efficiency while maintaining the generation quality.

[2]https://guozixunnicolas.github.io/DENT-DDSP-demo/

Table 1: *Simulation quality of DENT-DDSP using different amounts of training data with various s2t and model comparison*

model	no. of trainable param.	s2t	amount of training data	MSSL
DENT-DDSP (ours)	2k	[0.8, 1, 0)	60-sec parallel	0.184
		[0.6, 0.8)	60-sec parallel	0.187
		[0.4, 0.6)	60-sec parallel	0.186
		[0.8, 1, 0)	40-sec parallel	0.177
		[0.6, 0.8)	40-sec parallel	0.187
		[0.4, 0.6)	40-sec parallel	0.188
		[0.8, 1, 0)	20-sec parallel	0.170
		[0.6, 0.8)	20-sec parallel	0.189
		[0.4, 0.6)	20-sec parallel	0.189
		[0.8, 1, 0)	10-sec parallel	**0.170**
		[0.6, 0.8)	10-sec parallel	0.181
		[0.4, 0.6)	10-sec parallel	0.189
SimuGAN	14M	-	10-min unparallel	0.173
Clean augment	non-trainable	-	-	0.192
G.726 augment	non-trainable	-	-	0.197
Codec2 augment	non-trainable	-	-	0.197



(a) Clean speech (b) Real noisy speech (c) DENT-DDSP

(d) SimuGAN (e) Codec2 augment (e) G.726 augment

Figure 2: *Magnitude spectrogram comparison between clean audio, audio simulated from various distortion models and the real noisy data.*

3.3.3. Generalization ability

To reflect the generalization ability of the trainable distortion models: DENT-DDSP and SimuGAN over different types of audio, non-speech audio are used to test the distortion models. The non-speech audio (available online[2]) include ambient noise, guitar and piano music and synthesized audio. The spectrograms of the testing audio and the simulated noisy audio from the two models are shown in Figure 3. By observation, data simulated by DENT-DDSP have consistent noise spectrums and the audio are distorted to the desired spectral behaviour: the spectrogram becomes blurred with certain frequencies being dampened. Yet, data simulated by SimuGAN fail to produce such consistent distortion behaviour. Moreover, data simulated by SimuGAN either contain artefacts or become unintelligible. Hence, DENT-DDSP shows stronger generalization ability compared to SimuGAN due to the use of explainable DDSP components.

3.3.4. Benchmarking the noise-robust ASR task

To compare against the benchmark model, several additional models are obtained as follows: Data simulated by DENT augment model is obtained by setting $\lambda = 0.79, 1.26$ in Equation 1

Table 2: *Training time and MSSL comparison among DENT-DDSPs with different* $ds_factors$

ds factor	training time	MSSL
16	7min	0.170
8	13.5min	0.169
4	25.5min	0.170
2	60.2min	0.169

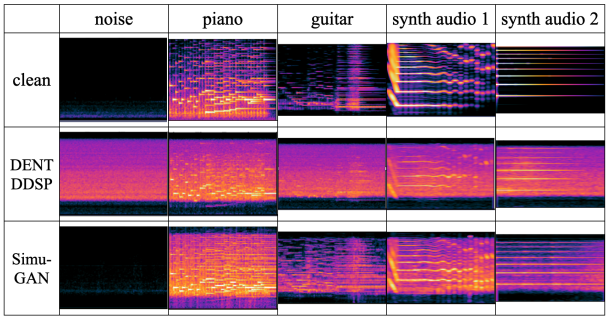


Figure 3: *Noisy data simulated by DENT-DDSP and SimuGAN using unseen non-speech data*

in a trained DENT-DDSP model so that the noise gain of the $n_{out}(t)$ become $\pm 2dB$; Additionally, a clean model is trained only using clean speeches. Table 3 shows the WER of these models tested on a common test set consists of real noisy data. By observation, the DENT augment model achieves the closest WER compared to the benchmark model with a difference of 2.7% and outperforms the best baseline model: SimuGAN by 7.3%. This proves that DENT-DDSP is able to simulate noisy data with similar characteristics to the real noisy data and offers an alternative to simulate the VHF/UHF transmitted data thus greatly reducing the effort to collect these data manually on a large scale.

Table 3: *WER comparison using same ASR model with simulated data from different distortion models and real noisy data*

model	noisy speech source	WER
Clean	-	93.4%
Clean augment	clean speech + aggregated noise	73.8%
G.726 augment	g726 codec speech + aggregated noise	75.6%
Codec2 augment	codec2 speech + aggregated noise	83.2%
SimuGAN	SimuGAN simulated	65.9%
DENT-DDSP (ours)	DENT simulated	66.4%
DENT augment (ours)	DENT simulated with noise gain $= \pm 2dB$	**58.6%**
Benchmark	real noisy data	55.9%

4. Conclusion

We propose a fully explainable and controllable model DENT-DDSP based on novel DDSP components which only requires 10 seconds of training data to achieve explicit distortion modelling. Besides the existing DDSP components, we propose 2 novel DDSP components: waveshaper and computationally-efficient DRC which can be easily integrated to other DDSP models. Experiments have shown that data simulated by DENT-DDSP achieve the highest simulation fidelity among all distortion models. Moreover, the noise-robust ASR system trained with the data simulated by DENT-DDSP achieves similar WER compared to the benchmark trained with real noisy data with a 2.7% difference in terms of WER. It has also achieved 7.3% WER improvement over the best baseline distortion model.

5. Acknowledgement

This work is supported by ST Engineering Mission Software Services Pte. Ltd (Research Collaboration No. REQ0149132). The computational resource was partially provided by the National Super Computing Centre, Singapore.

6. References

[1] D. Graff, K. Walker, S. Strassel, X. Ma, K. Jones, and A. Sawyer, "The RATS collection: Supporting HLT research with degraded audio data," in *Proc. of the 9th Int. Conf. on Lang. Resour. and Eval. (LREC)*, 2014, pp. 1970–1977.

[2] S. Badrinath and H. Balakrishnan, "Automatic speech recognition for air traffic control communications," *Transportation Research Record*, vol. 2676, no. 1, pp. 798–810, 2022.

[3] T. Pellegrini, J. Farinas, E. Delpech, and F. Lancelot, "The airbus air traffic control speech recognition 2018 challenge: towards atc automatic transcription and call sign detection," in *Proc. Int. Speech Commun. Assoc.(Interspeech)*, 2018, pp. 2993–2997.

[4] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: telephone speech corpus for research and development," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 1992, pp. 517–520 vol.1.

[5] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 22, no. 4, pp. 745–777, 2014.

[6] M. J. F. Gales, "Model-based techniques for noise robust speech recognition," in *Ph.D. thesis, University of Cambridge*, 1995.

[7] Y. Minami and S. Furui, "A maximum likelihood procedure for a universal adaptation method based on hmm composition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 1995, pp. 129–132.

[8] A. Sankar and C.-H. Lee, "Robust speech recognition based on stochastic matching," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 1995, pp. 121–124.

[9] A. Acero, l. Deng, T. Kristjansson, and J. Zhang, "Hmm adaptation using vector taylor series for noisy speech recognition," 2000, pp. 869–872.

[10] D. Ma, G. Li, H. Xu, and E. S. Chng, "Improving code-switching speech recognition with data augmentation and system combination," in *Proc. Asia-Pacific Signal and Inf. Process. Assoc. Annu. Summit and Conf. (APSIPA ASC)*, 2019, pp. 1308–1312.

[11] M. Ferràs, S. Madikeri, P. Motlicek, S. Dey, and H. Bourlard, "A large-scale open-source acoustic simulator for speaker recognition," *IEEE Signal Processing Letters*, vol. 23, no. 4, pp. 527–531, 2016.

[12] C. Chen, N. Hou, Y. Hu, S. Shirol, and E. S. Chng, "Noise-robust speech recognition with 10 minutes unparalleled in-domain data," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4298–4302.

[13] H. Hu, T. Tan, and Y. Qian, "Generative adversarial networks based data augmentation for noise robust speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5044–5048.

[14] J. H. Engel, L. Hantrakul, C. Gu, and A. Roberts, "DDSP: differentiable digital signal processing," in *in Proc. Int. Conf. on Learn. Representations(ICLR)*, 2020.

[15] G. Fabbro, V. Golkov, T. Kemp, and D. Cremers, "Speech synthesis and control using differentiable dsp," 2020. [Online]. Available: https://arxiv.org/abs/2010.15084

[16] J. Alonso and C. Erkut, "Latent space explorations of singing voice synthesis using ddsp," in *Proc. of 18th Sound and Music Computing Conf.*, 2021.

[17] D. Giannoulis, M. Massberg, and J. Reiss, "Digital dynamic range compressor design—a tutorial and analysis," *J. of the Audio Eng. Soc.(AES)*, vol. 60, pp. 399–408, 2012.

[18] C. Chen, Y. Hu, N. Hou, X. Qi, H. Zou, and E. S. Chng, "Self-critical sequence training for automatic speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 3688–3692.

[19] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Int. Speech Commun. Assoc.(Interspeech)*, 2020, pp. 5036–5040.

[20] Y. Hu, N. Hou, C. Chen, and E. S. Chng, "Interactive feature fusion for end-to-end noise-robust speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6292–6296.

[21] Y. Hu, N. Hou, C. Chen, and E. Chng, "Dual-path style learning for end-to-end noise-robust speech recognition," in *preprint arXiv2203.14838*, 2022.

[22] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang, "A comparative study on transformer vs rnn in speech applications," *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec 2019.