DP-DWA: DUAL-PATH DYNAMIC WEIGHT ATTENTION NETWORK WITH STREAMING DFSMN-SAN FOR AUTOMATIC SPEECH RECOGNITION

Dongpeng Ma[1], Yiwen Wang[1], Liqiang He [1], Mingjie Jin[1], Dan Su[1], Dong Yu[2]

[1]Tencent AI Lab, Shenzhen, China
[2]Tencent AI Lab, Bellevue, WA, USA
{dongpengma, poveywang, andylqhe, jackkingjin, dansu, dyu}@tencent.com

ABSTRACT

In multi-channel far-field automatic speech recognition (ASR) scenarios, distortion is introduced when the speech signal is processed by the front end, which damages the recognition performance for the ASR tasks. In this paper, we propose a dual-path network for the far-field acoustic model, which uses voice processing (VP) signal and acoustic echo cancellation (AEC) signal as input. Specifically, we design a dynamic weight attention (DWA) module for combining two signals. Besides, we streamline our best deep feed-forward sequential memory network with self-attention (DFSMN-SAN) acoustic model for real-time requirements. Joint-training strategy is adopted to optimize the proposed approach. We find that with dual-path network, we can achieve a 54.5% relative improvement in character error rate (CER) on a 10,000-hour online conference task . In addition, our proposed method is not affected by the arrangement of different microphone arrays. We achieve a 23.56% relative improvement on a vehicle task, which has an array with two microphones.

Index Terms— dynamic weight attention, streaming, dual-path, acoustic model, joint training

1. INTRODUCTION

Far-field multi-channel ASR is a challenging research topic. In the real scene, background noise significantly reduces target speech intelligibility. Speech enhancement tasks are proposed to address these concerns. In modern ASR, the input signal is generally the enhanced signal, which has better auditory perception for humans. Specifically, for multi-channel situations, signal processing techniques, such as beamforming (BF), is used to generate single-channel speech signals. However, due to the large amount of nonlinear distortion [1], such technique harms the performance of ASR systems.

Acoustic models for multi-channel speech recognition can be divided into two categories. One type is cascading architecture. Map a time-delay feature to substitute beamformer's weight through deep neural network (DNN) network in the first part and use DNN's output for the acoustic model in the second part [2, 3, 4]. The integrated multi-channel techniques achieve better ASR performance as it is straightforward to jointly optimize the model. However, due to the lack of multi-channel data, this kind of method has poor generalization performance. The other type is to use separate modules. Speech enhancement is applied, including localization, beamforming. Normally, Minimum Variance Distortionless Response (MVDR) [5] or multi-channel Wiener filtering (MWF) [6] are chosen for generating single-channel enhanced signal. Then, the enhanced signal is sent to a conventional acoustic model [7, 8]. This method may not be optimal for ASR tasks. Besides, this method suffers from residual noise problems since utterance levels' MVDR filters are not optimal for noise reduction.

In this paper, we propose a dual-path ASR system, combining VP signal and AEC signal. Three main contributions are proposed in this work. Firstly, we propose a dual-path network for the far-field acoustic model, which uses VP signal and AEC signal as input to improve ASR performance. Secondly, we design a dynamic weight attention module for combining two signals, which use AEC and VP as input to dynamically generate weights for AEC and VP. Thirdly, in the frequency domain, only the power spectrum is used. Phase is important for the perceptual quality of speech [9], Time-domain separation has become more and more popular in recent years [10, 11]. Multi-stage signal processing begin to thrive in the recent several years [12, 13]. We propose a two-stage paradigm, the first stage is in the time domain, and the second stage is in the frequency domain, which uses log-mel filterbank as input. We utilize the joint training strategy, including MSE and Connectionist Temporal Classification (CTC) loss to further improve the proposed method. The experiments show that the proposed method outperforms the baseline method by 54.5% relatively in CER on a 10,000-hour online conference task, and outperforms the baseline method by 23.56% relatively in CER on a 10,000-hour vehicle task.

The rest of the paper is organized as follows: Section 2 introduces the conventional ASR system and Section 3 describes the proposed DP-DWA model. Experimental setup is present in Section 4. Results are reported in Section 5. Finally, we draw conclusions in Section 6.

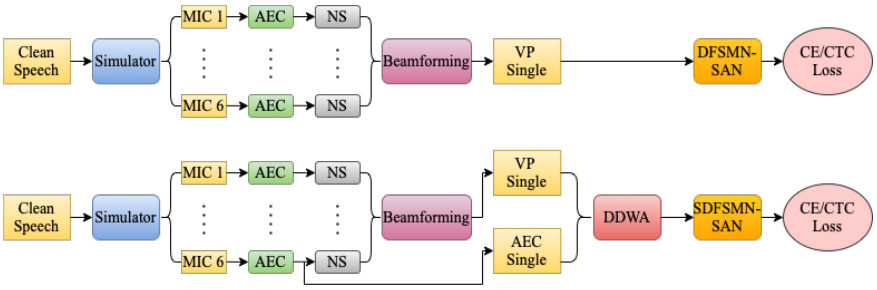


Fig. 1. Comparison of conventional acoustic model and proposed DP-DWA based model. The figure above is the structure of the conventional ASR acoustic model, while the figure below depicts the process of the proposed DP-DWA model.

2. CONVENTIONAL ASR ACOUSTIC SYSTEM

This section describes the conventional ASR system for online conferences, while the proposed improved version will be introduced in the next section. The conventional far-field ASR acoustic system is depicted in Fig.1. The whole progress is divided into two parts, including the signal processing part and the acoustic model part. Voice processed (VP) signal is obtained after the signal processing module. The extracted acoustic feature is then sent to DFSMN-SAN acoustic model. Generally, the conventional scheme includes three blocks, the AEC part, the noise suppression (NS) part, and the BF part for the online stage. The goal of each part is to improve the sound quality of the speech signal. However, the cascade model seriously damages the spectral structure of the speech signal, which harms the performance of the ASR system.

For the training stage, the total procedure includes the simulation part, AEC part, NS part, BF part, and acoustic part. Specifically, consider a clean signal $\hat{s}$, for an M-size microphone array, the i-th channel's input is simulated using the corresponding room impulse responses (RIRs) h_i. Noise, booming noise, etc, are added to simulate real scenery. Then, for each channel $i, (i = 1, 2, ..., M)$, the simulated signal y_i is simulated and added up with noise. For each channel, after the AES, NS part, the beamforming process is used to produce a single output speech signal. Then, fbank feature is extracted from the enhanced single channel signal, which is sent to DFSMN-SAN acoustic model.

3. PROPOSED MODEL STRUCTURE

In this section, we propose a dual-path network for better utilizing the VP signal. Dual-path refers to one channel of VP signal and one reference signal after AEC progress with channel p, which is depicted in Fig.1. One advantage of using the dual-path input signal is that the model utilizes both signals, which can make full use of both channels' advantages and make up for the shortcomings, which comes to the target clean speech signal with less spectrum distortion.

3.1. Model Description

In this part, we propose a time-frequency domain hybrid dual-path speech recognition system, which can take advantages of both method to make up for the shortcomings, to achieve the purpose of fine-resolution, strong modeling ability, and low computing power consumption.

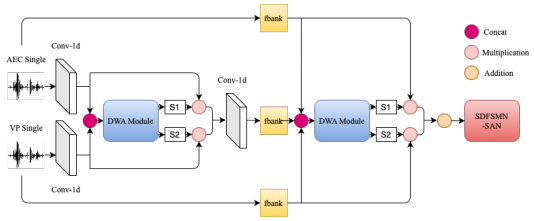


Fig. 2. DP-DWA network structure

The procedure of generating simulation signal is done by simulation, AEC, NS, and BF module, which is the same as the description in Section 2. To simplify, we choose one fixed channel's signal after AEC processing as the reference signal. The two inputs signals are the selected channel's preprocessing signal and the VP signal. Conv-1d operation is chosen to replace short-time Fourier transform. The dual-path spectrum is concatenated and then sent to the dynamic weight attention (DWA) block to get two vectors $s1$ and $s2$. DWA block will be introduced in the next part, whose purpose is to get two dynamic vectors as two dynamic weights. The two vectors will be multiplied by the corresponding spectrum feature. Then, symmetrically, a conv-1d operation is followed behind. This operation is symmetrical, similar to the STFT and inverse short-time Fourier transform (iSTFT). The acoustic feature is extracted from the 1-dimension signal generated after the conv-1d operation. The above-mentioned procedure models the time-domain signal.

Fbank features are extracted from the VP signal and the AEC signal. For frequency-domain signal, the model maintains the AM part to get two dynamic weights. The only difference is the method for concatenate operation. To get the better utility of time-domain signal, we concatenate the three-path fbank

feature as input and output the final fbank feature, which is finally sent to the acoustic recognition part.

3.2. Dynamic Weight Attention Module

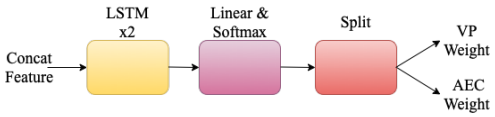


Fig. 3. Dynamic weight attention module

The attention module is to get two dynamic weight vectors, which are used as dynamic weight attention (DWA) for combining dual-path information. DWA model consists of three blocks, where features concatenated in the frequency domain are sent to the LSTM, Linear, Softmax and Split block, as in Fig.3. Finally, multiply two dynamic weight vectors with the dual paths' feature vectors and add them together to get a feature vector. The result is the dynamic mixture of the two parts. Dual paths' features after conv-1d operation are sent to the first DWA module. Similarly, three fbank features are concatenated and sent to the second DWA module.

3.3. Streaming Acoustic Model

Details of DFSMN-SAN are mentioned in [14]. DFSMN model needs the context of the whole utterance, which is sent to SAN part afterward. This solution doesn't meet the streaming requirements. In this work, a segmentation block is added where chunk-wise information is sent to the network. Specifically, the segmentation block splits a sequential input into overlapped chunks. Symmetrically, a merge block is added to convert the chunks back to the sequential output. To retain more historical information, LSTM is added in the SAN part, which is used to learn the information among the historical chunks. LSTM's output is then concatenated with SAN's output. With this strategy, the model only needs to accumulate a chunk of speech inputs to obtain the corresponding acoustic score and obtain the ASR result. The batch size for LSTM is chunk size, which makes the two parts correspond.

4. EXPERIMENTAL SETUP

4.1. Datasets

On the online conference task, the training corpus is collected from several different application domains, all in Mandarin. Our Device is a linear microphone array, contains 6 microphones, the microphone spacing is 26cm. We use simulation tools to convert single-channel speech to multi-channel speech. In order to improve system robustness, RIRs and noises are added to the speech. RIRs are created with different rectangular room sizes, speaker positions. Noises are collected from the meeting room office with our microphone array, which including music, keyboard, air-conditioning wind noise, etc. Together, they contain a total of 10,000-hour speeches. On the vehicle task, The training corpus is obtained in a similar way to the online conference scene.

On the online conference task, to evaluate the performance of our proposed method, we report performance on the simulation test set which consists of 1.44 hours' hand-transcribed anonymized utterances extracted from reading speech (1001 utterances). We refer it as Read. The test sets are like the training sets to obtain multi-channel speech through simulation tools with noises and RIRs added. On the vehicle task, we have three test sets. One speaker and multi speakers are simulated test sets. Real is a real test set, recorded in highway scene, containing 4.54 hours' recording data.

4.2. Training Setup

Feature vectors used for the SDFSMN-SAN part are 40-dimensional log-Mel filterbank energy features with first-order and second-order derivatives. Log-mel filterbank energy features are computed with a 25ms window and shifted every 10ms. A global mean and variance normalization is applied for each frame. All the experiments are based on the CTC learning framework. The CI-syllable-based acoustic modeling method is chosen for CTC learning as mentioned in [15]. The target labels of CTC learning are defined to include 1394 Mandarin syllables, 39 English phones, and a blank.

4.3. Joint Training Strategy

To further optimize the learning goals of the network, we use a joint-training strategy for training. During the training stage, the acoustic model and proposed DP-DWA model are trained separately. Cross entropy (CE) loss and MSE loss are chosen for each part. Then the two models are put together for joint training, remaining the loss functions mentioned above. Finally, CTC loss is chosen for further optimization. For stable CTC learning, gradients are clipped to $[-1.0, 1.0]$. All models are trained in a distributed manner using BMUF [16] optimization with 8 Tesla P40 GPUs.

5. EXPERIMENTAL RESULTS AND ANALYSIS

5.1. Model Architecture Experiments

On the online conference task, different variations of model structures are compared. The systems' performance is evaluated by WER. As in Table 1, the results show that the VP signal (C2) is worse than the AEC signal (C1). That's because in order to get a better listening feeling on the online conference scene, the noise and reverberation are eliminated very cleanly, resulting in some damage to the human voice. Results show that the proposed dynamic weight attention model achieves better performance than C1 and C2 with attention

Table 1. WER Results (%) for different models

MODEL	Read	Model size(M)
AEC Single (C1)	10.86	35.92
VP Single (C2)	18.31	35.92
C1 and C2 Concat	10.11	36.48
C1 and C2 with Attention [17]	10.02	38.32
C2 same size with C4 (C3)	22.65	45.39
C1 and C2 with Ours (C4)	8.97	45.84
C3 with Joint Training (C5)	**8.32**	45.84

Table 2. WER Results (%) for vehicle experiment

MODEL	Single Speaker	Multi Speaker	Real
AEC (C1)	23.92	39.26	14.50
VP (C2)	26.61	35.31	12.85
MIC(C3)	58.52	-	-
C1 C2 with Ours	**18.88**	**26.70**	**11.57**

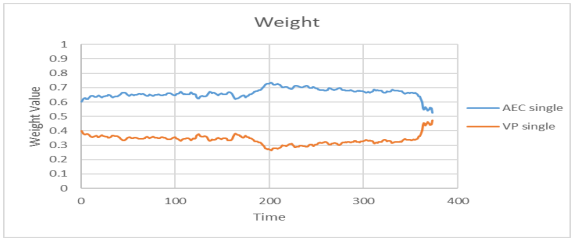
module [17]. Besides, further experiments explored the benefits of joint training. Result (C5) show that our proposed method achieves the best performance, which outperforms the baseline method by 54.5% relatively compare with VP results (C2). The results show the importance of considering both the time domain and frequency domain for performance improvement. The model size of the acoustic model which we used is 35.92M. DP-DWA model size is 9.92M. DP-DWA plus acoustic model size is 45.84M. By increasing the depth of the acoustic model, we get an acoustic model of the same number of parameters as C4. Results (C3) show that the model with the same number of parameters is worse than the C4.

5.2. Experiments on vehicle task

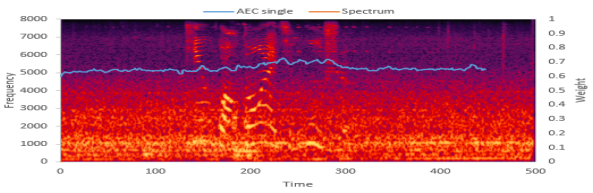
Results on the vehicle task are shown in Table 2. In the vehicle experiments, two microphones with 11.8cm intervals are selected. One speaker and multi speakers are simulation test sets. Real is a real test set, recorded in highway scene. Under single speaker conditions, results (C2) are worse after VP operation. The reason is that RIRs used in the One speaker test set are simulated, and our fixed beamforming weight in VP is developed on real RIRs. Results (C2) are getting better with multi speakers, as interfering people are eliminated by beamforming. Our proposed DP-DWA model achieves the best performance, Obtained a relative improvement of 29.04% 29.04% 9.96% respectively compare with VP results (C2). The method we propose can ignore the influence of array arrangement.

5.3. Why DP-DWA works

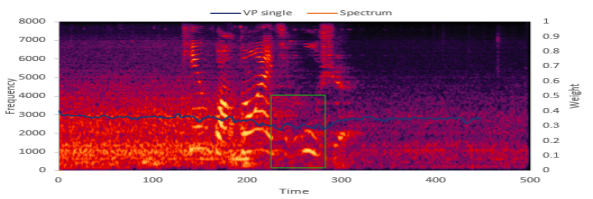
In order to explore why DP-DWA works, we get the attention weight of AEC and VP single from the vehicle task model, as shown in Fig.4(a). The attention weights are obtained from the Real test set, as shown in Table 2. It can be seen that the weight changes as the input signal changes. The AEC weight has a greater value than the VP weight. We analyze the probable reason is that the AEC signal has less nonlinear distortion than the VP signal. Signals with better quality will have greater weight. As shown in the green box in Fig.4(c), the VP signal is damaged by single processing. AEC signal has better quality than the VP signal, the weight of the AEC signal is increased, as shown in Fig.4(b).



(a) Weight



(b) Spectrum of AEC



(c) Spectrum of VP

Fig. 4. Weight and Spectrum

6. CONCLUSION

In this paper, we propose a dual-path ASR system, combining VP signal and AEC signal. We design a dynamic weight attention module for combining two signals, which use AEC and VP as inputs to dynamically generate weights for AEC and VP. We propose a two-stage paradigm, the first stage is in the time domain, while the second stage is in the frequency domain, which uses a log-mel filterbank as input. The joint training strategy plays an important role in further improving the performance of the two parts. The experiments show that the proposed system achieves a significant improvement compared to existing acoustic models for the different arrangements of microphone arrays. The future work will continue to explore the performance of our method in complex acoustic scenes and verify the superiority of our method.

7. REFERENCES

[1] Zhuohuang Zhang, Yong Xu, Meng Yu, Shi-Xiong Zhang, Lianwu Chen, and Dong Yu, "Adl-mvdr: All deep learning mvdr beamformer for target speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6089–6093.

[2] Wu Minhua, Kenichi Kumatani, Shiva Sundaram, Nikko Ström, and Björn Hoffmeister, "Frequency domain multi-channel acoustic modeling for distant speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6640–6644.

[3] Tara N Sainath, Ron J Weiss, Kevin W Wilson, Bo Li, Arun Narayanan, Ehsan Variani, Michiel Bacchiani, Izhak Shafran, Andrew Senior, Kean Chin, et al., "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, 2017.

[4] Tsubasa Ochiai, Shinji Watanabe, Takaaki Hori, John R Hershey, and Xiong Xiao, "Unified architecture for multichannel end-to-end speech recognition with neural beamforming," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1274–1288, 2017.

[5] Matthias Wolfel and John McDonough, "Minimum variance distortionless response spectral estimation," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 117–126, 2005.

[6] Ann Spriet, Marc Moonen, and Jan Wouters, "Spatially pre-processed speech distortion weighted multi-channel wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, 2004.

[7] Thomas Hain, Lukáš Burget, John Dines, Philip N Garner, František Grézl, Asmaa El Hannani, Marijn Huijbregts, Martin Karafiat, Mike Lincoln, and Vincent Wan, "Transcribing meetings with the amida systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 486–498, 2011.

[8] Takuya Higuchi, Nobutaka Ito, Shoko Araki, Takuya Yoshioka, Marc Delcroix, and Tomohiro Nakatani, "Online mvdr beamformer based on complex gaussian mixture model with spatial prior for noise robust asr," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 780–793, 2017.

[9] Ke Tan and DeLiang Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2019.

[10] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[11] Yi Luo, Zhuo Chen, and Takuya Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 46–50.

[12] Andong Li, Wenzhe Liu, Xiaoxue Luo, Guochen Yu, Chengshi Zheng, and Xiaodong Li, "A simultaneous denoising and dereverberation framework with target decoupling," *arXiv preprint arXiv:2106.12743*, 2021.

[13] Nils L Westhausen and Bernd T Meyer, "Dual-signal transformation lstm network for real-time noise suppression," *arXiv preprint arXiv:2005.07551*, 2020.

[14] Zhao You, Dan Su, Jie Chen, Chao Weng, and Dong Yu, "Dfsmn-san with persistent memory model for automatic speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7704–7708.

[15] Zhongdi Qu, Parisa Haghani, Eugene Weinstein, and Pedro Moreno, "Syllable-based acoustic modeling with ctc-smbr-lstm," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 173–177.

[16] Kai Chen and Qiang Huo, "Scalable training of deep learning machines by incremental block training with intra-block parallel optimization and blockwise model-update filtering," in *2016 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2016, pp. 5880–5884.

[17] Xuan Ji, Meng Yu, Jie Chen, Jimeng Zheng, Dan Su, and Dong Yu, "Integration of multi-look beamformers for multi-channel keyword spotting," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7464–7468.