End-to-End Far-Field Speech Recognition with Unified Dereverberation and Beamforming

Wangyou Zhang[1], *Aswin Shanmugam Subramanian*[2], *Xuankai Chang*[2], *Shinji Watanabe*[2], *Yanmin Qian*[1]

[1]MoE Key Lab of Artificial Intelligence & SpeechLab, Department of Computer Science and Engineering, AI Institute, Shanghai Jiao Tong University, Shanghai
[2]Center for Language and Speech Processing, Johns Hopkins University, USA

wyz-97@sjtu.edu.cn, {aswin, xchang14, shinjiw}@jhu.edu, yanminqian@sjtu.edu.cn

Abstract

Despite successful applications of end-to-end approaches in multi-channel speech recognition, the performance still degrades severely when the speech is corrupted by reverberation. In this paper, we integrate the dereverberation module into the end-to-end multi-channel speech recognition system and explore two different frontend architectures. First, a multi-source mask-based weighted prediction error (WPE) module is incorporated in the frontend for dereverberation. Second, another novel frontend architecture is proposed, which extends the weighted power minimization distortionless response (WPD) convolutional beamformer to perform simultaneous separation and dereverberation. We derive a new formulation from the original WPD, which can handle multi-source input, and replace eigenvalue decomposition with the matrix inverse operation to make the back-propagation algorithm more stable. The above two architectures are optimized in a fully end-to-end manner, only using the speech recognition criterion. Experiments on both spatialized wsj1-2mix corpus and REVERB show that our proposed model outperformed the conventional methods in reverberant scenarios.

Index Terms: Dereverberation, speech separation, overlapped speech recognition, neural beamforming, WPD

1. Introduction

Over the past few years, thanks to the advances in deep learning, significant progress has been made in automatic speech recognition (ASR). Both deep neural network (DNN)/hidden Markov model (HMM) hybrid systems and end-to-end (E2E) systems have attained surprisingly good performance in close-talk scenarios [1–4]. However, it is still a challenging task to recognize speech signals in far-field scenarios, where background noise and reverberation are commonly observed and even interfering speech from other speakers is involved [5, 6]. In recent years, many studies have been focusing on the far-field speech recognition task, including the combination of the speech enhancement frontend and ASR backend [7, 8] and noise robust adaptation approaches [9, 10]. Meanwhile, it is commonly observed that speech processing with multiple microphones usually outperforms the single-microphone one, because additional spatial information can be exploited. Therefore, many existing microphone array signal processing methods can be utilized as the frontend for end-to-end far-field speech recognition, such as the multi-channel Weiner filter [11, 12], minimum variance distortionless response (MVDR) and minimum power distortionless response (MPDR) beamforming [12, 13], multi-frame beamforming [14], etc. In addition, reverberation is also an important problem in real scenarios, which can lead to dramatic degradation in the ASR performance [15]. Various deep learning based methods have been proposed for dereverberation, including DNN based approaches [16–18] incorporating the weighted prediction error (WPE) algorithm [19,20] and complex ideal ratio mask based approach for denoising and dereverberation [21].

In this work, we propose a novel E2E architecture that can perform dereverberation, beamforming and recognition simultaneously. Inspired by the recently developed unified convolutional beamformer for simultaneous denoising and dereverberation, named weighted power minimization distortionless response (WPD) [22, 23], we reformulate WPD by replacing eigenvalue decomposition with an equivalent matrix inverse operation, which makes it differentiable and more stable. The new architecture consists of a frontend and an ASR backend. In the frontend, two novel architectures are explored for joint speech dereverberation, enhancement and separation. In the backend, a joint connectionist temporal classication (CTC) / attention-based encoder-decoder model [24] is used to recognize each separated speech stream. Note that our proposed framework can be used for both single-speaker and multi-speaker scenarios. And in this paper, we mainly focus on the multi-speaker case, which is a more difficult task. It is worth noting that this end-to-end architecture is optimized only based on the final ASR criterion, which was also proven feasible in previous works [17, 25–27]. Our experiments show that our newly proposed method outperformed the conventional end-to-end ASR systems [25, 26, 28] in both single-speaker and multi-speaker reverberant conditions.

2. End-to-End Multi-Channel ASR

This section reviews the end-to-end multi-channel speech recognition system for both single-speaker ($J = 1$) [28] and multi-speaker ($J > 1$) [25, 26] conditions, as shown in Fig. 1. Without loss of generality, we consider the input speech as a mixture of J ($J \geq 1$) different speakers. For simplicity, we consider the noise as the 0-th source ($j = 0$) in the input signal.

The model is composed mainly of two modules, namely the frontend for speech separation and the backend for ASR. The frontend is a mask-based multi-source neural beamformer. First, the masking network estimates the masks $\mathbf{M}_c^j$ for every source $j \in \{0, 1, \ldots, J\}$ on each channel $c \in \{1, \ldots, C\}$ from the input spectrum $\mathbf{X}_c = (x_{t,f,c})_{t,f} \in \mathbb{C}^{T \times F}$:

$$\mathbf{M} = \left(m_{t,f,c}^j\right)_{t,f,c,j} = \text{MaskEstimator}(\mathbf{X}) \in \mathbb{C}^{T \times F \times C \times (J+1)}, \quad (1)$$

where $m_{t,f,c}^j \in [0, 1]$, T and F represent time and frequency

[†]Shinji Watanabe and Yanmin Qian are the corresponding authors.

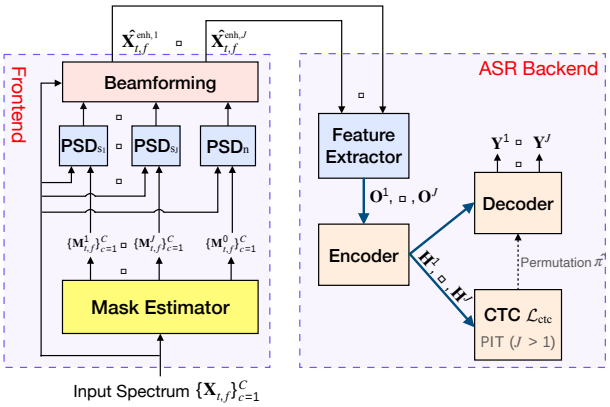

Figure 1: *End-to-End Multi-channel ASR Model.*

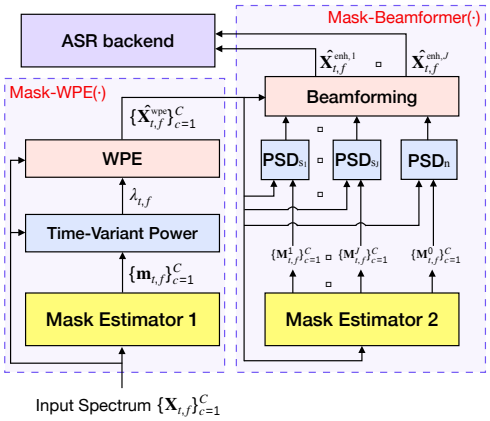

Figure 2: *Proposed end-to-end ASR arch#1: cascaded dereverberation and beamforming frontend.*

dimensions respectively. Second, the multi-source neural beamformer separates the mixture into J streams using the MVDR formalization [12]. The estimated masks are used to compute the cross-channel power spectral density (PSD) matrices $\mathbf{\Phi}^j$ [29–31] and then the time-invariant filter $\mathbf{g}_f^j$ for each speaker j:

$$\mathbf{\Phi}_f^j = \frac{1}{\sum_{t=1}^T m_{t,f}^j} \sum_{t=1}^T m_{t,f}^j \mathbf{x}_{t,f} \mathbf{x}_{t,f}^H \in \mathbb{C}^{C\times C}, \quad (2)$$

$$\mathbf{g}_f^j = \frac{(\sum_{i\neq j} \mathbf{\Phi}_f^i)^{-1} \mathbf{\Phi}_f^j}{\mathrm{Trace}((\sum_{i\neq j} \mathbf{\Phi}_f^i)^{-1} \mathbf{\Phi}_f^j)} \mathbf{u} \in \mathbb{C}^C, \quad (3)$$

where $\mathbf{x}_{t,f} = \{x_{t,f,c}\}_{c=1}^C$, $m_{t,f}^j = \frac{1}{C}\sum_{c=1}^C m_{t,f,c}^j$, $(\cdot)^H$ represents the conjugate transpose, and $\mathbf{u} \in \mathbb{R}^C$ is a vector denoting the reference microphone estimated by an attention mechanism [32]. Finally, the separated speech $\hat{\mathbf{X}}^{\mathrm{enh},j}$ of each speaker j is derived by applying the filters $\mathbf{g}^j$ to the input speech $\mathbf{X}$, from which the log Mel-filterbank feature with global mean and variance normalization (GMVN-LMF$(\cdot)$) is further extracted:

$$\hat{x}_{t,f}^{\mathrm{enh},j} = (\mathbf{g}_f^j)^H \mathbf{x}_{t,f} \in \mathbb{C}, \quad (4)$$

$$\mathbf{O}^j = \text{GMVN-LMF}(|\hat{\mathbf{X}}^{\mathrm{enh},j}|), \quad (5)$$

where $\hat{\mathbf{X}}^{\mathrm{enh},j} \in \mathbb{C}^{T\times F}$, and $\mathbf{O}^j$ is the extracted feature for ASR.

The backend is a joint CTC/attention-based encoder-decoder model [24] for single-channel speech recognition. First, the encoder transforms the feature $\mathbf{O}^j = \{\mathbf{o}_1^j, \ldots, \mathbf{o}_T^j\}$ of each speaker j into a high-level representation $\mathbf{H}^j = \{\mathbf{h}_1^j, \ldots, \mathbf{h}_L^j\}$ $(L \leq T)$ with subsampling. Then, the representation is processed by the attention-based decoder to generate the output token sequences $\mathbf{Y}^j = \{y_1^j, \ldots, y_N^j\}$. The ASR process is formulated as follows:

$$\mathbf{H}^j = \text{Encoder}(\mathbf{O}^j), \quad (6)$$

$$\mathbf{y}_n^j \sim \text{Attention-Decoder}(\mathbf{H}^j, \mathbf{y}_{n-1}^j), \quad (7)$$

where $\mathbf{y}_n^j$ is the posterior probability vector for the n-th token. Note that in the multi-speaker case, i.e. $J > 1$, in order to solve the label ambiguity problem, the permutation invariant training (PIT) technique [33–36] is further applied in the CTC module to determine the order of the label sequences. The whole model is optimized with only the ASR loss $\mathcal{L}$ combining the attention and CTC losses with the determined label sequence order.

3. End-to-End ASR with Unified Frontend

In this section, we introduce the proposed multi-channel speech recognition architecture for coping with the reverberant speech. First, we describe the mask-based WPE model for multi-channel dereverberation. Then, we show a cascade integration method which incorporates the mask-based WPE model followed by the model introduced in last Section, where the WPE filter coefficients are estimated for each speaker. Furthermore, another frontend architecture extending the WPD beamformer [22] is designed, which unifies the dereverberation and beamforming modules with our new formulation.

3.1. Mask-based WPE model

The mask-based WPE algorithm [16] is introduced in this subsection. First, the input spectrum $\mathbf{X} = (\mathbf{x}_{t,f})_{t,f}$ is fed into a neural network to estimate a time-frequency mask $\mathbf{m} = (m_{t,f,c})_{t,f,c}$, as formulated below:

$$\mathbf{m} = \text{MaskEstimator}(\mathbf{X}) \in \mathbb{R}^{T\times F\times C}, \quad (8)$$

$$\mathbf{x}_{t,f} = \mathbf{x}'_{t,f} + \mathbf{n}_{t,f} \approx \sum_{j=1}^J \mathbf{v}_f^j s_{t,f}^j + \mathbf{n}_{t,f} \in \mathbb{C}^C, \quad (9)$$

where $\mathbf{x}'_{t,f}$ is the direct path and early reflection of the source signal, $s_{t,f}^j$ is the j-th source signal, $\mathbf{v}_f = [v_f^{(0)}, v_f^{(1)}, \cdots, v_f^{(C-1)}]^T \in \mathbb{C}^C$ is the steering vector, and $\mathbf{n}_{t,f} \in \mathbb{C}^C$ is the noise and late reverberation of source signals.

With the estimated mask, the time-variant power $\lambda_{t,f}$ of the input signal can be estimated by Eq. (10), and then the signal can be dereverberated via a standard WPE procedure:

$$\lambda_{t,f} = \frac{1}{C}\sum_{c=1}^C \frac{m_{t,f,c}}{\frac{1}{T}\sum_{\tau=1}^T m_{\tau,f,c}} |x_{t,f,c}|^2 \in \mathbb{R}, \quad (10)$$

$$\hat{\mathbf{x}}_{t,f}^{\mathrm{wpe}} = \mathrm{WPE}(\mathbf{x}_{t,f}, \lambda_{t,f}), \quad (11)$$

3.2. Cascaded dereverberation and beamforming

One straightforward way to enable dereverberation in the multi-channel ASR system in Section 2 is the cascade integration of the mask-based WPE model and the neural beamformer like [27]. As illustrated in Fig. 2, the multi-channel input speech mixture is first fed into the mask-based WPE model, which is composed of a mask estimator and a WPE filter. Then the dereverberated speech is processed by the beamformer introduced in Section 2 to generate the enhanced single-channel speech of J speakers for speech recognition. The frontend process can be formulated as follows:

$$\hat{\mathbf{X}}^{\mathrm{enh}} = \text{Mask-Beamformer}(\text{Mask-WPE}(\mathbf{X})), \quad (12)$$

where $\hat{\mathbf{X}}^{\mathrm{enh}} = \{\hat{\mathbf{X}}^{\mathrm{enh},j}\}_{j=1}^J \in \mathbb{C}^{T\times F\times J}$ is the set of the separated speech from all speakers, Mask-Beamformer$(\cdot)$ and Mask-WPE$(\cdot)$ denote the respective modules in Fig. 2. The

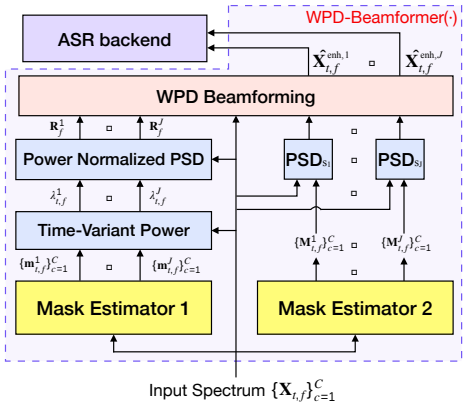

Figure 3: *Proposed end-to-end ASR arch#2: unified dereverberation and beamforming frontend.*

ASR backend here is the same as that described in Section 2.

3.3. Unified dereverberation and beamforming

The original WPD beamformer [22] aims to eliminate the late reverberation and noise from the noisy signal, while keeping the direct signal undistorted. It combines the ideas of WPE and MPDR beamformer [13], and optimizes their filters at the same time, with the constrained optimization objective below:

$$\bar{\mathbf{w}} = \arg\min_{\bar{\mathbf{w}}} \sum_t \frac{\left|\bar{\mathbf{w}}_f^H \bar{\mathbf{x}}_{t,f}\right|^2}{\lambda_{t,f}} \quad \text{s.t.} \quad \mathbf{w}_{0,f}^H \mathbf{v}_f = v_f^{(\text{ref})} \quad (13)$$

where $\bar{\mathbf{w}} = [\mathbf{w}_{0,f}^T, \mathbf{w}_{D,f}^T, \mathbf{w}_{D+1,f}^T, \cdots, \mathbf{w}_{K+D-1,f}^T]^T \in \mathbb{C}^{C(K+1)}$ is the WPD filter coefficient, $\bar{\mathbf{x}}_{t,f} = [\mathbf{x}_{t,f}^T, \mathbf{x}_{t-D,f}^T, \mathbf{x}_{t-D-1,f}^T, \cdots, \mathbf{x}_{t-K-D+1,f}^T]^T \in \mathbb{C}^{C(K+1)}$ is the concatenation of input signals of current and previous frames, D is the delay parameter, K is the number of filter taps, $v_f^{(\text{ref})}$ is the value of the steering vector at the reference channel, and $\lambda_{t,f}$ is the power of the desired signal as in Eq. (10).

By solving the above constrained optimization problem, we can calculate the WPD filter $\bar{\mathbf{w}}_f$ and the enhanced signal $\hat{\mathbf{X}}^{\text{enh}}$ by the following formulas:

$$\bar{\mathbf{w}}_f = \frac{\mathbf{R}_f^{-1}\bar{\mathbf{v}}_f}{\bar{\mathbf{v}}_f^H \mathbf{R}_f^{-1} \bar{\mathbf{v}}_f} \left(v_f^{(\text{ref})}\right)^* \in \mathbb{C}^{C(K+1)}, \quad (14)$$

$$\mathbf{R}_f = \sum_{t=D}^{T} \frac{\bar{\mathbf{x}}_{t-D,f} \bar{\mathbf{x}}_{t-D,f}^H}{\lambda_{t,f}} \in \mathbb{C}^{C(K+1)\times C(K+1)}, \quad (15)$$

$$\hat{\mathbf{X}}_{t,f}^{\text{enh}} = \bar{\mathbf{w}}_f^H \bar{\mathbf{x}}_{t,f} \in \mathbb{C}, \quad (16)$$

where $\mathbf{R}_f$ is the power normalized covariance matrix, $\bar{\mathbf{v}}_f = [\mathbf{v}_f^T, \mathbf{0}, \cdots, \mathbf{0}]^T \in \mathbb{C}^{C(K+1)}$, and $(\cdot)^*$ denotes complex conjugate. While the original WPD can perform denoising and dereverberation simultaneously with an elegant formulation, it is only designed for speech enhancement of the single-speaker input. In addition, the steering vector $\bar{\mathbf{v}}$ in Eq. (14) is needed for calculating the beamformer weights, which requires the direction information of the sound source or needs to be approximated by eigenvalue decomposition of a complex matrix [23].

Based on the above formulation, we first derive another equivalent formula that no longer requires the steering vector $\bar{\mathbf{v}}_f$, and then extend the original WPD to the multi-speaker case. Consider the padded speech signal $\tilde{\mathbf{x}}_{t,f}'^T = [\mathbf{x}_{t,f}'^T, \mathbf{0}, \cdots, \mathbf{0}]^T \in \mathbb{C}^{C(K+1)}$, it is easy to derive from Eq. (9) that:

$$\tilde{\mathbf{x}}_{t,f}' = \bar{\mathbf{v}}_f s_{t,f}, \quad (17)$$

$$(\mathbf{\Phi}_{\tilde{\mathbf{x}}'\tilde{\mathbf{x}}'})_f = \sum_{t=1}^{T} \frac{m_{t,f} \tilde{\mathbf{x}}_{t,f}' \tilde{\mathbf{x}}_{t,f}'^H}{\sum_{\tau=1}^{T} m_{\tau,f}} = \bar{\mathbf{v}}_f \phi_f \bar{\mathbf{v}}_f^H = \bar{\mathbf{v}}_f \bar{\mathbf{v}}_f^H \phi_f, \quad (18)$$

$$v_f^{(\text{ref})} = \bar{\mathbf{v}}_f^T \bar{\mathbf{u}}, \quad (19)$$

where $(\mathbf{\Phi}_{\tilde{\mathbf{x}}'\tilde{\mathbf{x}}'})_f \in \mathbb{C}^{C(K+1)\times C(K+1)}$ is the cross-channel PSD matrix of the padded speech signal $\tilde{\mathbf{x}}_{t,f}'$, $\bar{\mathbf{u}} = [\mathbf{u}^T, \mathbf{0}, \cdots, \mathbf{0}]^T \in \mathbb{R}^{C(K+1)}$ and $\mathbf{u} \in \mathbb{R}^C$ is the reference vector denoting the reference microphone estimated by an attention mechanism. Substitute Eq. (17) – (19) into Eq. (14), we can derive that:

$$\bar{\mathbf{w}}_f = \frac{\mathbf{R}_f^{-1}(\mathbf{\Phi}_{\tilde{\mathbf{x}}'\tilde{\mathbf{x}}'})_f}{\text{Trace}\left[\mathbf{R}_f^{-1}(\mathbf{\Phi}_{\tilde{\mathbf{x}}'\tilde{\mathbf{x}}'})_f\right]} \bar{\mathbf{u}}. \quad (20)$$

This new formula is equivalent to Eq. (14), but no longer requires the steering vector for calculating the filter weights.

Furthermore, we can easily extend WPD to the multi-speaker case. For each speaker j, the corresponding covariance matrix $\mathbf{R}_f^j$ can be derived from Eq. (15) and (10), where the estimated mask $\mathbf{m}^j$ for speaker j is used for both dereverberation and beamforming. Then the WPD beamformer for speaker j can be calculated from Eq. (20). Finally, the separated speech of each speaker can be derived from Eq. (16), using the corresponding WPD beamformer. Note that the masks for calculating $\mathbf{R}_f^j$ and $\mathbf{\Phi}_{\tilde{\mathbf{x}}'\tilde{\mathbf{x}}'}^j$ for each speaker j can be either shared using a single mask estimator or estimated by two separate mask estimators. The WPD based architecture is illustrated in Fig. 3.

4. Experiments

To make our experimental results comparable to previous results of MIMO-Speech [25, 26], we evaluated the proposed methods on the same spatialized wsj1-2mix dataset as in [25, 26], which consists of two sub-datasets: anechoic and reverberant. The reverberation time (RT_{60}) of the reverberant data ranges from 200 ms to 600 ms. In each sub-dataset, the duration of the spatialized speech for training, development and evaluation is 98.5 hr, 1.3 hr and 0.8 hr respectively. We also adopt the multi-condition training in [25, 37], i.e. include the WSJ train_si284 in training to improve the performance. We also test our methods for single-speaker speech recognition on the REVERB dataset [38], which uses 2-channel simulated reverberant data for training and 8-channel real data for evaluation.

For feature extraction, the short-time Fourier transform (STFT) is performed with a 16-kHz sampling rate and a 25-ms Hann window with a 10-ms stride, and the spectral feature's dimension is $F = 257$. After the frontend processing, 80-dimensional log Mel-filterbank features are extracted from the enhanced spectrum of each separated speech, where the global mean and variance normalization is applied using the statistics from the single-speaker WSJ1 training set. The number of channels for training in our experiments is $C = 2$. But it can be extended to an arbitrary number of channels as described in [32].

4.1. Experimental Setup

All our proposed end-to-end multi-channel speech recognition models are implemented based on the ESPnet framework [39].[1] The AdaDelta optimizer with $\rho = 0.95$ and $\epsilon = 10^{-8}$ is used for training. The data in both anechoic and reverberant conditions are used for training.

[1] Our experimental setup is available at `https://github.com/Emrys365/espnet/blob/wsj1_mix_spatialized/egs/wsj1_mix_spatialized/asr1/`.

Table 1: *Performance (WER [%]) of the proposed arch1 / arch2 models with different numbers of filter taps (K) and microphones (C) on the spatialized reverberant wsj1-2mix eval set.*

K \ C	2	4	6
1	28.87 / 27.44	17.95 / 16.67	14.92 / **13.97**
3	27.62 / 26.42	16.65 / 15.95	14.63 / 14.23
5	**21.88** / 25.54	15.93 / **15.72**	15.46 / 16.81
7	26.62 / 25.68	16.09 / 16.32	18.67 / 22.40
10	26.64 / 25.81	18.67 / 19.55	27.79 / 36.28

Table 2: *Performance evaluation on the spatialized reverberant wsj1-2mix corpus.*

Model	dev WER (%)	eval WER (%)
baseline (RNN backend) [25]	34.98	29.99
+ Nara-WPE [26]	24.45	17.67
baseline (Transformer backend) [26]	32.95	28.01
+ Nara-WPE [26]	**19.17**	**15.24**
proposed arch 1	19.37	14.63
proposed arch 2[2]	**18.34**	**13.97**

In the mask-based WPE module, the mask estimation network is a 3-layer bidirectional long-short term memory with projection (BLSTMP) network with 300 cells in each direction. The number of iterations for performing mask-based WPE is set to 1. The prediction delay D and the number of taps K is set to 3 and 5 respectively. The mask estimators in both the MVDR beamformer and the WPD beamformer are 3-layer BLSTMP networks with 512 cells. Note that in all our experiments except Section 4.3, we used a shared mask estimator instead of two separate ones in Fig. 3. In the ASR module, following the configurations in [26], we use the CNN-Transformer based encoder, which consists of 2 CNN blocks and 12 Transformer layers, and the 6-layer Transformer-based decoder. The self-attention in all Transformer layers has the same configuration as in [26], i.e. 4 heads and 256 dimensions. As for decoding, a word-level language model [40] trained on the official text data included in the WSJ corpus is used. The interpolation factor between CTC and attention losses is set to 0.2.

For experiments on REVERB, the network configuration and experimental conditions are the same as [28]. The reference microphone is fixed as the second channel. Both dereverberation and denoising subnetworks are trained to predict two dimensional time-frequency masks.

4.2. Evaluation of the proposed architectures for multi-speaker speech recognition

Since our proposed architectures can be tested flexibly with different numbers of microphones C and filter taps K, even if the model is trained with fixed C and K, we first evaluate the performance of the proposed two architectures with different numbers of filter taps and microphones for inference on the reverberant wsj1-2mix dataset. The results are presented in Table 1. We can observe that the performance can be significantly improved when more microphones are available. The number of filter taps closer to the training setup ($K = 5$) usually leads to better results, but with more microphones, using fewer filter taps may also provide enough information for dereverberation and increases the stability of the operations in Eq. (20). The best performance for arch 1 and arch 2 is achieved with $C = 6, K = 3$ and $C = 6, K = 1$ respectively.

Then we compare the performances of our proposed models with the baseline models. In Table 2, the baselines are the MIMO-Speech models with RNN backend (row 1) and Transformer backend (row 3) from our previous study [25, 26]. Since these baseline models do not contain a dereverberation module, we also introduce two enhanced baselines (row 2 & 4), i.e. MIMO-Speech with iterative Nara-WPE[3] preprocessing. We ran Nara-WPE with 10 filter taps for 5 iterations to preprocess both training and evaluation data for the baseline models. By comparing the four baselines and our proposed two architectures with best results taken from Table 1, we can observe that both proposed models combining neural dereverberation and beamforming in the end-to-end structure achieve comparable results to the best baseline ones. Note that our models do not need an iterative process compared to the Nara-WPE preprocessing baselines. Finally, our proposed arch 2 model based on WPD outperforms all baseline methods.

Table 3: *ASR performance on REVERB evaluation real dataset comparing unified and cascade filtering with $K = 5$ & $C = 8$.*

Frontend	Near WER (%)	Far WER (%)
WPE + MVDR	10.8	13.6
proposed arch 2	**8.9**	**11.1**

4.3. Effectiveness of unified filtering for single source robust speech recognition

We also evaluated our methods in the single source condition on REVERB. We first trained a single source multi-channel E2E ASR model, which is a variant of a cascaded architecture in Section 3.2 based on the WPE and MVDR frontend [28]. Then, we replace the cascaded frontend with the proposed unified WPD frontend, and compare both frontends. Using our WPD unified filter gives a significant improvement in performance over the cascade configuration, as shown in Table 3. This shows that using our unified frontend is also effective for single source data.

These results indicate that our proposed E2E multi-channel speech recognition model is a powerful method for applications in reverberant single-speaker and multi-speaker scenarios.

5. Conclusion

In this paper, we proposed an end-to-end multi-channel far-field speech recognition framework with unified dereverberation and beamforming, which is capable of performing speech dereverberation, separation and recognition simultaneously. The whole model is optimized via only the ASR criterion but can still learn relatively good dereverberation and separation skills. Two novel frontend architectures are explored, and promising performance is achieved on the spatialized wsj1-2mix corpus compared to the previous MIMO-Speech model. Experimental results on REVERB dataset also demonstrate the effectiveness of our proposed WPD based architecture.

6. Acknowledgement

This work was supported by the China NSFC project No. U1736202. Experiments have been carried out on the PI supercomputers at Shanghai Jiao Tong University. We would like to thank the NTT Communication Laboratories for the use of their DNN-WPE module[4] for our implementation.

[2] The arch 2 model in Table 2 was trained on the basis of a pretrained MIMO-Speech model, since the direct training of arch 2 currently does not work well due to numerical instability issues.

[3] https://github.com/fgnt/nara_wpe

[4] https://github.com/nttcslab-sp/dnn_wpe

7. References

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The Microsoft 2017 conversational speech recognition system," in *Proc. IEEE ICASSP*, 2018, pp. 5934–5938.

[3] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *Proc. IEEE ICASSP*, 2018, pp. 4774–4778.

[4] S. Karita *et al.*, "A comparative study on transformer vs RNN in speech applications," in *Proc. IEEE ASRU*, 2019, pp. 449–456.

[5] R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M. L. Seltzer, H. Zen, and M. Souden, "Speech processing for digital home assistants: Combining signal processing with deep-learning techniques," *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 111–124, 2019.

[6] S. Watanabe, M. Mandel, J. Barker, and E. Vincent, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *arXiv preprint arXiv:2004.09249*, 2020.

[7] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. ISCA Interspeech*, 2016, pp. 545–549.

[8] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *Proc. IEEE ICASSP*, 2018, pp. 5739–5743.

[9] A. Narayanan and D. Wang, "Joint noise adaptive training for robust automatic speech recognition," in *Proc. IEEE ICASSP*, 2014, pp. 2504–2508.

[10] T. Tan, Y. Qian, H. Hu, Y. Zhou, W. Ding, and K. Yu, "Adaptive very deep convolutional residual network for noise robust speech recognition," *IEEE/ACM Trans. ASLP.*, vol. 26, no. 8, pp. 1393–1405, 2018.

[11] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, 2004.

[12] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 2, pp. 260–276, 2009.

[13] H. L. Van Trees, *Optimum array processing: Part IV of detection, estimation, and modulation theory*. John Wiley & Sons, 2004.

[14] Z.-Q. Wang, H. Erdogan, S. Wisdom, K. Wilson, and J. R. Hershey, "Sequential multi-frame neural beamforming for speech separation and enhancement," *arXiv preprint arXiv:1911.07953*, 2019.

[15] M. Wölfel and J. W. McDonough, *Distant speech recognition*. John Wiley & Sons, 2009.

[16] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online WPE dereverberation." in *Proc. ISCA Interspeech*, 2017, pp. 384–388.

[17] J. Heymann, L. Drude, R. Haeb-Umbach, K. Kinoshita, and T. Nakatani, "Joint optimization of neural network-based WPE dereverberation and acoustic model for robust online ASR," in *Proc. IEEE ICASSP*, 2019, pp. 6655–6659.

[18] T. Taniguchi, A. S. Subramanian, X. Wang, D. Tran, Y. Fujita, and S. Watanabe, "Generalized weighted-prediction-error dereverberation with varying source priors for reverberant speech recognition," in *Proc. IEEE WASPAA*, 2019, pp. 288–292.

[19] T. Yoshioka, T. Nakatani, K. Kinoshita, and M. Miyoshi, "Speech dereverberation and denoising based on time varying speech model and autoregressive reverberation model," in *Speech Processing in Modern Communication*, 2010, pp. 151–182.

[20] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE/ACM Trans. ASLP.*, vol. 20, no. 10, pp. 2707–2720, 2012.

[21] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. ASLP.*, vol. 25, no. 7, pp. 1492–1501, 2017.

[22] T. Nakatani and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation," *IEEE Signal Processing Letters*, vol. 26, no. 6, pp. 903–907, 2019.

[23] ——, "Simultaneous denoising and dereverberation for low-latency applications using frame-by-frame online unified convolutional beamformer," *Proc. ISCA Interspeech*, pp. 111–115, 2019.

[24] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. IEEE ICASSP*, 2017, pp. 4835–4839.

[25] X. Chang, W. Zhang, Y. Qian, J. Le Roux, and S. Watanabe, "MIMO-Speech: End-to-end multi-channel multi-speaker speech recognition," in *Proc. IEEE ASRU*, 2019, pp. 237–244.

[26] ——, "End-to-end multi-speaker speech recognition with transformer," in *Proc. IEEE ICASSP*, 2020, pp. 6129–6133.

[27] A. S. Subramanian, X. Wang, M. K. Baskar, S. Watanabe, T. Taniguchi, D. Tran, and Y. Fujita, "Speech enhancement using end-to-end speech recognition objectives," in *Proc. IEEE WASPAA*, 2019, pp. 229–233.

[28] A. S. Subramanian, X. Wang, S. Watanabe, T. Taniguchi, D. Tran, and Y. Fujita, "An investigation of end-to-end multichannel speech recognition for reverberant and mismatch conditions," *arXiv preprint arXiv:1904.09049*, 2019.

[29] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi *et al.*, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. IEEE ASRU*, Dec. 2015, pp. 436–443.

[30] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE ICASSP*, Mar. 2016, pp. 196–200.

[31] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks," *Proc. ISCA Interspeech*, pp. 1981–1985, 2016.

[32] T. Ochiai, S. Watanabe, T. Hori, and J. R. Hershey, "Multichannel end-to-end speech recognition," in *Proc. ICML*, 2017, pp. 2632–2641.

[33] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE ICASSP*, 2017, pp. 241–245.

[34] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. ASLP.*, vol. 25, no. 10, pp. 1901–1913, 2017.

[35] H. Seki, T. Hori, S. Watanabe, J. Le Roux, and J. R. Hershey, "A purely end-to-end system for multi-speaker speech recognition," in *Proc. ACL*, Jul. 2018, pp. 2620–2630.

[36] W. Zhang, X. Chang, Y. Qian, and S. Watanabe, "Improving end-to-end single-channel multi-talker speech recognition," *IEEE/ACM Trans. ASLP.*, vol. 28, pp. 1385–1394, 2020.

[37] T. Ochiai, S. Watanabe, and S. Katagiri, "Does speech enhancement work with end-to-end ASR objectives?: Experimental analysis of multichannel end-to-end ASR," in *Proc. MLSP*. IEEE, 2017, pp. 1–6.

[38] K. Kinoshita *et al.*, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[39] S. Watanabe *et al.*, "ESPnet: End-to-End Speech Processing Toolkit," in *Proc. ISCA Interspeech*, 2018, pp. 2207–2211.

[40] T. Hori, J. Cho, and S. Watanabe, "End-to-end speech recognition with word-based RNN language models," in *Proc. IEEE SLT*, 2018, pp. 389–396.