FAR-FIELD LOCATION GUIDED TARGET SPEECH EXTRACTION USING END-TO-END SPEECH RECOGNITION OBJECTIVES

Aswin Shanmugam Subramanian[1*], *Chao Weng*[2], *Meng Yu*[2], *Shi-Xiong Zhang*[2], *Yong Xu*[2], *Shinji Watanabe*[1], *Dong Yu*[2]

[1]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA
[2]Tencent AI Lab, Bellevue, WA, USA

ABSTRACT

Target speech extraction is a specific case of source separation where an auxiliary information like the location or some pre-saved anchor speech examples of the target speaker is used to resolve the permutation ambiguity. Traditionally such systems are optimized based on signal reconstruction objectives. Recently end-to-end automatic speech recognition (ASR) methods have enabled to optimize source separation systems with only the transcription based objective. This paper proposes a method to jointly optimize a location guided target speech extraction module along with a speech recognition module only with ASR error minimization criteria. Experimental comparisons with corresponding conventional pipeline systems verify that this task can be realized by end-to-end ASR training objectives without using parallel clean data. We show promising target speech recognition results in mixtures of two speakers and noise, and discuss interesting properties of the proposed system in terms of speech enhancement/separation objectives and word error rates. Finally, we design a system that can take both location and anchor speech as input at the same time and show that the performance can be further improved.

Index Terms— end-to-end speech recognition, target speech extraction, neural beamformer

1. INTRODUCTION

Speech separation which involves extracting individual speech sources from a mixed speech signal has become an important preprocessing step for many speech processing applications. Far-field speech recognition devices, such as Amazon Alexa and Google Home have become omnipresent [1]. These devices are used in challenging home environments [2] and hence combining source separation and denoising with automatic speech recognition (ASR) is becoming crucial [3, 4]. Conventionally, frontend systems to perform speech separation and speech enhancement have been trained with a signal reconstruction objective with the non-overlapped signals as targets. Such systems might not be suitable for (1) real data, and (2) for applications where the target is different from reconstructing a clean signal. End-to-end (E2E) speech recognition [5–7] and other neural advances in ASR have made it easier to train frontend denoising [8–10], dereverberation [11] and speech separation [12–15] systems only based on speech recognition objectives.

Depending on the choice of number of targets, there are broadly two methods for speech separation - (1) unbiased estimation and (2) biased estimation. In unbiased estimation, the blind separation system is usually required to estimate the separation for each source simultaneously with equal importance. Deep clustering (DC) [16] and permutation invariant training (PIT) [17] are two popular techniques for unbiased blind source separation. The PIT method to handle a mixture of two speakers was extended in [15, 18] to train with an end-to-end speech recognition objective to output two individual character sequences. These unbiased estimation approaches suffer from two limitations. First, it has a permutation ambiguity in mapping between the speakers in the mixture and the separated outputs. Second, it has very high memory and computational demands proportional to the number of possible source permutations

In contrast, in biased separation, an additional source of information is assumed to be available, which helps in identifying the target speaker. For example, in [19, 20], the location in terms of the azimuth angle is fed as an angle feature to extract the speaker of interest. In [21–24], an anchor speech utterance consisting of only the target speaker voice is passed to inform the separation system on which speaker to extract. However all these systems require parallel clean speech. Recently, an extension of [21] by training the anchor speech based single channel target speech extraction using an E2E speech recognition objective was proposed in [25]. Including multichannel processing is crucial in tasks like this.

The contributions from this paper are as follows: (1) we propose a location guided multi-channel target speech extraction system with a carefully designed neural network optimized with an ASR objective and only the target speaker's transcription as the target, (2) we also include noise in the data apart from interference and perform simultaneous denoising and separation, (3) we perform a detailed analysis with many possible differentiable beamformers as the unique nature of the problem gives us more possibilities as the steering vector is given as the input and (4) we propose a hybrid target speech extraction system that takes both location and anchor speech to further improve the performance. Additionally, we also evaluate our system in terms of a separation and a perception metric apart from word error rate (WER). This is possible because we use an explainable AI approach to model the network where the intermediate output can be interpreted as the separated signal.

2. THE PIPELINE APPROACH

The pipeline approach where the frontend location aware target speech extraction is trained independent of the ASR system is explained in this section.

2.1. Input Features

Let $Y_1, Y_2, \cdots, Y_M$ be the M-channel input signal in the short-time Fourier transform domain, with $Y_m \in \mathbb{C}^{T \times F}$, where T is the total number of frames and F is the total number of frequency com-

[*] This work was done while the author was an intern at Tencent AI.

ponents. Similar to [19] and [20], three types of features are concatenated and given as input: (1) spectral, (2) spatial and (3) angle features. Magnitude of the input signal at channel 1 is used as the spectral feature i.e. $|Y_1|$.

Inter-microphone phase difference (IPD) $\mathrm{p}_i(t,f)$ encodes spatial information and it is calculated as,

$$\mathrm{p}_i(t,f) = \frac{1}{M}[\cos\angle(\frac{y_{i_1}(t,f)}{y_{i_2}(t,f)}) + \mathrm{j}\sin\angle(\frac{y_{i_1}(t,f)}{y_{i_2}(t,f)})], i = 1:M, \tag{1}$$

where $y_m(t,f)$ is the input signal at channel m, time t and frequency f, i represents an entry in a microphone pair list defined for calculating the IPD; and i_1 and i_2 are the index of microphones in each pair. We calculate IPD features for M pairs and then concatenate their real and imaginary parts together and hence $P_m \in \mathbb{R}^{T\times 2F}$.

The spectral and spatial features are one way of representing the multichannel input signal but they don't specify the identity of the target. In our task the location in terms of the azimuth angle of the target speaker and all the interfering speakers are known. This crucial information is encoded in the directional angle feature which is formulated with a pre-masking step similar to [19] as,

$$d_n(t,f) = |\sum_{i=1}^{M} \overline{e_n^{i,f}}\,\mathrm{p}_i(t,f)|, \tag{2}$$

$$a_n(t,f) = d_n(t,f) * \mathcal{I}(d_n(t,f) - d_s(t,f))_{s=1:N}, \tag{3}$$

where n is the speaker index, N is the number of speakers, $\mathcal{I}(.)$ is the indicator function that outputs 0 if the input difference is negative for any of the "$s = 1:N$" cases and 1 otherwise and $\overline{e_n^{i,f}}$ represents the conjugate of the steering vector coefficient of speaker n, direction of arrival at microphone pair i, and frequency f which can be calculated from the azimuth angle. The angle feature for each speaker $A_n \in \mathbb{R}^{T\times F}$ are concatenated in such a way the target speaker's is always placed first.

2.2. Mask Based Beamformer

A neural network that outputs a spectral mask is trained by using the groundtruth clean target speaker's signal as the target and takes the three types of features defined in Section 2.1 as input after concatenation. The masking network is trained separately and its output is fed to a beamformer in a pipeline like [19]. The phase sensitive magnitude spectral approximation (PS-MSA) loss [26] $\mathcal{L}^{\mathrm{PS}}$ is used to optimize the masking network and it is defined below:

$$\mathcal{L}^{\mathrm{PS}} = \frac{1}{T*F}\sum_{t,f} ||w_{\mathrm{S}}(t,f)|y_1(t,f)| - |s_1(t,f)|\max(\cos(\theta_{y_1}(t,f) - \theta_{s_1}(t,f)), 0)||^2, \tag{4}$$

where $w_{\mathrm{S}}(t,f) \in [0,1]$ is the output target speech mask of the masking network, $|s_1(t,f)|$ and $\theta_{s_1}(t,f)$ are the magnitude and phase spectrum of the target speech signal and $\theta_{y_1}(t,f)$ is the phase spectrum of the input signal at time t, frequency f and channel 1. For estimating a beamforming filter, we will also need a residual spectral mask that corresponds to the interference and noise in the signal. Since the masking network is optimized to output only the speech mask in Eq (4), we calculate the residual mask from the speech mask as,

$$w_{\mathrm{V}}(t,f) = 1 - w_{\mathrm{S}}(t,f). \tag{5}$$

The estimated masks are used to compute the speech and residual power spectral density (PSD) matrices $\mathbf{\Phi}_{\mathrm{S}}(f) \in \mathbb{C}^{M\times M}$ and $\mathbf{\Phi}_{\mathrm{V}}(f) \in \mathbb{C}^{M\times M}$ at frequency f as follows:

$$\mathbf{\Phi}_k(f) = \sum_{t=1}^{T} w_k(t,f)\mathbf{y}(t,f)\mathbf{y}^{\mathrm{H}}(t,f) \text{ where } k \in \{\mathrm{S},\mathrm{V}\},. \tag{6}$$

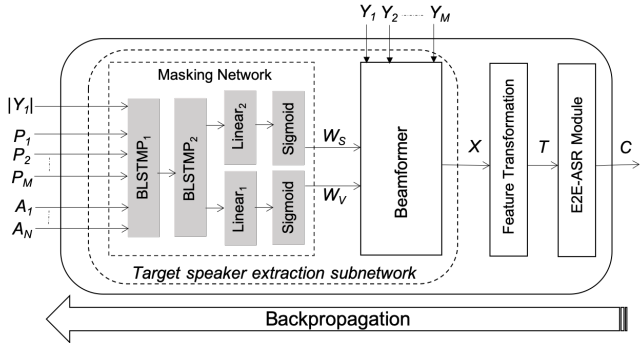


Fig. 1: Proposed End-to-End Architecture which takes the spectral $\{Y_m\}_{m=1}^{M}$, spatial $\{P_m\}_{m=1}^{M}$ and angle $\{A_n\}_{n=1}^{N}$ features as input to extract the target speech signal X with an ASR objective

where $^{\mathrm{H}}$ denotes the conjugate transpose. The PSD matrices are used to compute the target speech extracting minimum variance distortionless response (MVDR) beamforming filter $\mathbf{b}_{\mathrm{MVDR}}(f) \in \mathbb{C}^M$ at frequency f as follows:

$$\mathbf{b}_{\mathrm{MVDR}}(f) = \frac{\mathbf{\Phi}_{\mathrm{V}}(f)^{-1}\mathbf{\Phi}_{\mathrm{S}}(f)}{\mathrm{Trace}(\mathbf{\Phi}_{\mathrm{V}}(f)^{-1}\mathbf{\Phi}_{\mathrm{S}}(f))}\mathbf{u}, \tag{7}$$

where $\mathbf{u} \in \{0,1\}^M$ is a one-hot vector to choose a reference microphone. The separated target speech signal $x(t,f) \in \mathbb{C}$ is extracted by applying the beamforming filter estimated in Eq. (7) on the input signal as:

$$x(t,f) = \mathbf{b}^{\mathrm{H}}_{\mathrm{MVDR}}(f)\mathbf{y}(t,f). \tag{8}$$

We call the beamformer used here as **MVDR-1** as only the speech mask $w_{\mathrm{S}}(t,f)$ comes directly from the masking network.

3. END-TO-END TARGET SPEAKER ASR

3.1. Joint Optimization

We propose to connect the masking network and the beamformer with an E2E ASR module and train all the parameters of the combined network only based on the text transcription of the target speaker given as character sequence $C = (c_1, c_2, \cdots)$ as shown in Figure 1, unlike Eq. (4) which requires the target signal. As the beamformer is included in the computational graph, reference vector $\mathbf{u}$ introduced in Eq. (7) is softly estimated based on an attention mechanism such that $\sum_m u_m = 1$ and also included in the network [8]. The separated signal X by beamforming as in Eq. (8) is passed through feature transformation and the ASR module as follows,

$$T = \mathrm{MVN}(\mathrm{Log}(\mathrm{MelFilterbank}(|X|))) \tag{9}$$

$$C = \mathrm{ASR}(T). \tag{10}$$

Log Mel Filterbank transformation is applied on the magnitude of X and utterance based mean-variance normalization (MVN) is performed to produce an input T that is suitable for ASR.

3.2. Differentiable Beamformers

Additional to MVDR-1, we also explore other beamformers in the joint optimization approach. The mask $w_{\mathrm{V}}(t,f)$ can be estimated in the masking network unlike Eq. (5) and we refer to this beamformer as **MVDR-2**. We also try the following linearly constrained minimum variance (**LCMV**) beamformer which takes the steering vectors as input and requires only the residual spectral mask from

Table 1: WER (%) on our simulated data comparing the ASR performance of pipeline and E2E approaches

Frontend Type	*Row ID*	*Auxiliary Input*		*Beamformer Method*	*Post-filter*	*Fine-tune*	*Dev Avg.*	*Eval*								*Avg.*
		Location Angle	*Anchor Speech*					*HI*	*HN*	*LI*	*LN*					
											1-15	*16-45*	*46-90*	*91-180*	*Avg.*	
Pipeline	1	✓	✗	MVDR-1	✗	✗	16.0	29.5	28.7	12.5	31.4	13.1	10.2	8.8	13.7	21.1
E2E	2	✓	✗	MVDR-1	✗	✗	15.2	28.2	28.1	12.8	30.3	13.0	9.8	9.7	13.7	20.7
	3	✓	✗	MVDR-2	✗	✗	12.5	23.4	22.2	10.6	25.9	9.3	8.6	6.8	10.8	16.8
	4	✓	✗	LCMV	✗	✗	17.0	27.9	28.2	14.8	33.5	13.8	10.0	9.0	14.2	21.3
	5	✓	✗	GDR	✗	✗	13.7	24.9	24.9	11.9	29.4	11.3	8.0	8.7	12.3	18.5
	6	✓	✗	MVDR-2	✓	✗	13.3	23.4	24.4	11.5	26.5	10.0	7.6	6.3	10.7	17.5
	7	✓	✗	MVDR-2	✓	✓	**11.8**	20.7	22.0	10.6	27.0	8.1	7.7	6.0	10.2	15.9
Pipeline	8	✗	✓	MVDR-2	✗	✗	21.4	39.8	39.3	20.3	30.2	23.2	20.2	15.6	21.1	30.1
E2E	9	✗	✓	MVDR-2	✗	✗	30.8	47.3	46.1	24.4	38.0	30.4	30.2	27.5	30.6	37.1
	10	✓	✓	MVDR-2	✗	✓	**11.8**	22.6	21.6	**10.2**	26.0	10.2	8.5	7.6	11.3	16.4
	11	✓	✓	MVDR-2	✓	✗	12.9	22.5	24.1	11.5	**23.8**	9.5	8.1	7.5	10.6	17.2
	12	✓	✓	MVDR-2	✓	✓	**11.8**	**20.3**	**21.2**	10.6	24.6	**7.8**	**7.0**	**6.0**	**9.6**	**15.4**

Table 2: Comparison using SDR & PESQ metrics. PESQ is well correlated with WER in Table 1

Frontend Type	*Auxiliary Input*		*Beamformer Method*	*Post-filter*	*Fine-tune*	*SDR*						*PESQ*					
	Location Angle	*Anchor Speech*				*Dev Avg.*	*Eval*					*Dev Avg.*	*Eval*				
							HI	*HN*	*LI*	*LN*	*Avg.*		*HI*	*HN*	*LI*	*LN*	*Avg.*
Input	-	-	-	-	-	-2.93	-6.62	-6.34	-0.36	-1.04	-3.59	1.68	1.49	1.46	1.78	1.82	1.64
Pipeline	✓	✗	MVDR-1	✗	✗	**5.76**	2.25	2.08	7.12	7.07	4.63	2.24	2.01	1.99	2.32	2.31	2.16
	✓	✗	MVDR-1	✗	✗	4.18	1.07	0.51	6.01	6.13	3.43	2.25	2.03	2.01	2.36	2.36	2.19
E2E	✓	✗	MVDR-2	✗	✗	4.80	1.75	1.55	6.57	6.45	4.08	2.31	2.10	2.07	2.40	2.40	2.24
	✓	✗	LCMV	✗	✗	-0.13	-1.66	-1.67	1.18	1.57	0.15	2.21	2.06	2.03	2.29	2.30	2.17
	✓	✗	MVDR-2	✓	✓	2.29	2.05	1.86	2.58	1.86	2.09	1.50	1.39	1.40	1.56	1.54	1.47
	✓	✓	MVDR-2	✗	✓	5.75	2.43	**2.74**	**7.45**	**7.15**	**4.94**	**2.36**	**2.13**	**2.14**	**2.45**	**2.44**	**2.29**
	✓	✓	MVDR-2	✓	✓	2.75	**2.50**	2.30	3.37	3.41	2.90	1.52	1.39	1.40	1.62	1.59	1.50

the masking network as the target speech PSD matrix is not required here like Eq.-(7).

$$\mathbf{b}_{\text{LCMV}}(f) = \mathbf{\Phi}_{\text{V}}(f)^{-1}\mathbf{Q}(f)(\mathbf{Q}(f)^{\text{H}}\mathbf{\Phi}_{\text{V}}(f)^{-1}\mathbf{Q}(f))^{-1}\mathbf{r}, \quad (11)$$

where $\mathbf{Q}(f) \in \mathbb{C}^{M*N}$ is a matrix with steering vectors of the target and interfering speakers at frequency f, $\mathbf{r} \in \{0,1\}^N$ is a one-hot vector to choose the target speaker. We also try an interesting hybrid Generalized Distortionless Response (**GDR**) beamformer [27] formulated as follows:

$$\mathbf{b}_{\text{GDR}}(f) = \beta(f)\mathbf{b}_{\text{MVDR}}(f) + (1-\beta(f))\mathbf{b}_{\text{LCMV}}(f), \quad (12)$$

where $\beta(f) \in [0,1]$ is the frequency dependent trade-off factor between noise reduction with $\mathbf{b}_{\text{MVDR}}(f)$ in Eq. (7) and interference reduction with $\mathbf{b}_{\text{LCMV}}(f)$ in Eq. (11). The estimation of the distortion weight parameter vector $\boldsymbol{\beta} = [\beta(f)]_{f=1}^{F}$ is incorporated inside the network as follows:

$$\boldsymbol{\beta} = \text{Sigmoid}(\text{Linear}([|\mathbf{r}_{\text{S}}|^{\text{T}}, |\mathbf{r}_{\text{V}}|^{\text{T}}]^{\text{T}})), \quad (13)$$

where $\text{Linear}(\cdot)$ is an affine transformation with learnable parameters. Features $\mathbf{r}_{\text{V}}$ and $\mathbf{r}_{\text{S}}$ are obtained from the PSD matrices:

$$\mathbf{r}_k = \frac{1}{(M-1)^2}\sum_{m=1}^{M}\sum_{m'=1}^{M}[\phi_k(f,m,m')]_{f=1}^{F} \text{ where } k \in \{\text{S},\text{V}\}, \quad (14)$$

where $\phi_k(f,m,m')$ is m-m' entry of the PSD matrix $\mathbf{\Phi}_k(f)$. An optional postfiltering step can be included after the beamformer by performing an elementwise multiplication of the beamformed signal with the estimated target speech mask.

3.3. Combination of Anchor Speech as Auxiliary Information

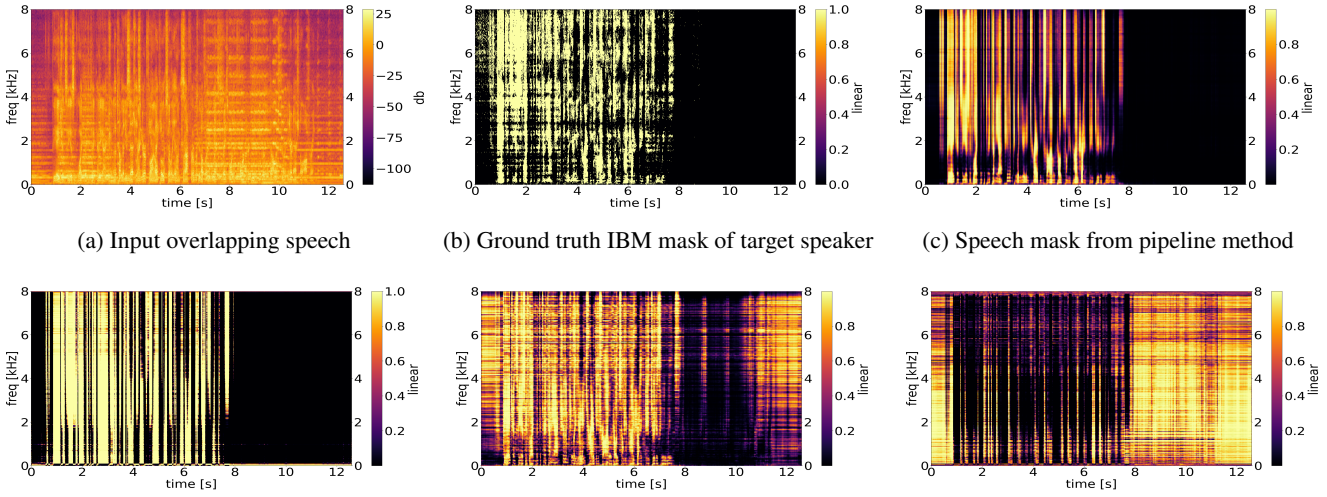
Additional to giving the location to find the target, a pre-stored example anchor speech utterance can also be passed to aid in the identification of the target in the mixed signal. We use a multi-channel extension of [25] and follow the same speaker adaptation procedure. The magnitude spectrum of the anchor speech is passed through an embedding network $G(.)$ and then time averaged to get an embedding. An adaptation layer is introduced between the first and second recurrent layers of the masking network. The output of the first layer of the masking network is scaled with the embedding using an element wise multiplication and the result is passed as input to the next layer of the masking network.

4. EXPERIMENTS

4.1. Data & Setup

We simulated the data using clean speech from wall street joirnal (WSJ) corpus [28]. The subset WSJ0 "si_tr_s", "dt_05" and "et_05" were used for training, development and evaluation respectively. For each utterance, we mixed a single noise source and interference utterance from a different speaker within the same set, so the resulting simulated data is the same size as the original clean data with 12776, 1206 and 651 utterances for training, development and each type of evaluation set respectively. A circular microphone array with 6 mics and diameter of 7cm was used. Six microphone pairs - (1, 4), (2, 5), (3, 6), (1, 2), (3,4) and (5, 6) were used to compute the IPD defined in Eq (1). Room impulse responses (RIR) were generated using image method [29] randomly from 3,000 different room configurations with the size (length-width-height) ranging from 3m-3m-2.5m to 8m-10m-6m. The reverberation time T60 is sampled in a range of 0.05s to 0.5s. 3 Point sources - target, interference and noise and the microphone-array are randomly located in the room. The SIR was randomly chosen from the set -5, 0, 5 dB and SNR from 0, 5, 10, 20 dB. Four types of evaluation data by fixing either SNR or SIR was simulated to get a better picture of the noise and interference robustness of the methods - (1) HI (SIR is -5 dB), (2) LI (SIR is 5 dB), (3) HN (SNR is 0 dB) and (4) LN (SNR is 20 dB).

Our implementation is based on ESPnet [30]. The masking



(a) Input overlapping speech (b) Ground truth IBM mask of target speaker (c) Speech mask from pipeline method

(d) Speech mask from E2E MVDR-1 (e) Speech mask from E2E MVDR-2 (f) Noise Mask from E2E MVDR-2

Fig. 2: Examining the spectral masks generated by the pipeline and E2E methods for an eval file with SNR-0dB and SIR-0dB where the noise source is music. The masks generated with E2E MVDR-2 in (e) and (f) was by giving both auxiliary information as input with postfiltering.

network consists of two output gate projected bidirectional long short-term memory (BLSTMP) recurrent layers with 771 and 514 as the cell and projection dimensions respectively. The E2E-ASR system has a 6-layer VGG-BLSTMP encoder with 320 units for the BLSTMs and a single layer LSTM decoder with 300 units trained with a joint CTC/attention criteria [7]. The ASR network is initialized with a pretrained model that used clean "si_tr_s" utterances from both WSJ0 and WSJ1. For the pipeline approach, the masking network is first trained with the PS-MSA loss defined in Eq (4), and then it is freezed while fine tuning the ASR network. The first channel is fixed as the reference while beamforming for the pipeline methods. The frontend pipeline models were trained with adam for up to 30 epochs with a patience of 3 epochs and the E2E models were trained with adadelta for up to 15 epochs with a patience of 3 epochs. Attention/CTC joint ASR decoding was performed with score combination with a word-level recurrent language model from [31] trained on the text data from WSJ. Three different clean utterances of the target speaker are randomly chosen and concatenated as the anchor speech for each mixed utterance. Three linear layers with rectified linear unit activation after the first two layers were used for the embedding network $G(.)$ in Section 3.3.

4.2. Results & Discussion

The WER results comparing the pipeline method with the ASR objective E2E methods for all the evaluation sets are given in Table 1. For the LN set, addtionally results at different target-interference angle differences are also shown. Overall, the proposed location-aware end-to-end method with angle features as auxiliary input (rows 1-6) outperforms the pipeline method except for LCMV beamformer defined in Eq (11) (row 4). Among the four beamfomers, the performance improves most significantly with MVDR-2 beamformer (row 3). Introducing a combination of fine tuning and post filtering improves the performance of MVDR-2 further (row 7). For fine tuning, the masking network of the E2E model is initialized from the frontend of the pipeline model and then its parameters are fine tuned with the ASR criteria. The pipeline system using only anchor speech based on [21] (row 8) (The feedforward layer for estimating the noise mask alone is optimized with ASR criteria) performs worse than the location based method (row 1). The combination of both auxiliary inputs with postfiltering mechanism and fine tuning (row 12) gives the best performance for most sets with considerable improvements in the low angle difference of "1-15" in LN set and the most challenging HI and HN sets.

The source to distortion ratio (SDR) and perceptual evaluation of speech quality (PESQ) scores computed with the dry signal as the reference are shown in Table 2 for some systems from Table 1. Among the methods using only the angle feature, the SDR scores are best for the pipeline system and PESQ is best for E2E MVDR-2. E2E system combining both auxiliary features with fine tuning works best overall in terms of both metrics for all sets. Like shown in [11], PESQ correlates well with WER when the beamformed signal is used directly. Introducing postfiltering makes the PESQ scores to be lower than the input. LCMV severely degrades the SDR score but the WER results are in general better than the pipeline method. This shows that SDR doesn't give a good indication of ASR performance.

The spectral masks generated for an evaluation example using different methods are shown in Figure 2. The mask generated by the pipeline MVDR-1 method is close to the ground truth ideal binary mask (IBM). The mask from the E2E MVDR-1 lacks the fine structure compared to the pipeline method. The E2E MVDR-2 method with both auxiliary features and postfiltering has a good formant structure. Some audio examples for demonstration are given in `https://sas91.github.io/E2E-LGASR.html`

5. SUMMARY AND CONCLUSIONS

This paper proposes a multichannel target speech extraction method using an end-to-end ASR objective, and experimentally showed that these methods give better performance compared to the corresponding pipeline methods. We also proposed an extension to combine speaker identity information from both location angle and anchor speech, which further improves performance. Our future work will be to explore approaches to improve the performance on cases where the angle difference between the target and interference is small.

6. REFERENCES

[1] Reinhold Haeb-Umbach, Shinji Watanabe, Tomohiro Nakatani, Michiel Bacchiani, et al., "Speech processing for digital home assistants," *IEEE Signal processing magazine*, vol. 36, no. 6, pp. 111–124, Nov 2019.

[2] Bo Li, Tara Sainath, Arun Narayanan, Joe Caroselli, et al., "Acoustic modeling for Google Home," in *Interspeech*, 2017, pp. 399–403.

[3] Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal, "The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Interspeech*, 2018, pp. 1561–1565.

[4] Naoyuki Kanda, Christoph Boeddeker, Jens Heitkaemper, Yusuke Fujita, et al., "Guided Source Separation Meets a Strong ASR Backend: Hitachi/Paderborn University Joint Investigation for Dinner Party ASR," in *Interspeech*, 2019, pp. 1248–1252.

[5] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, et al., "End-to-end attention-based large vocabulary speech recognition," in *ICASSP*, 2016, pp. 4945–4949.

[6] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*, 2016, pp. 4960–4964.

[7] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, et al., "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[8] Tsubasa Ochiai, Shinji Watanabe, Takaaki Hori, John R Hershey, et al., "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.

[9] J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink, et al., "Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system," in *ICASSP*, 2017, pp. 5325–5329.

[10] Bo Li, Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, et al., "Neural network adaptive beamforming for robust multichannel speech recognition," in *Interspeech*, 2016, pp. 1976–1980.

[11] Aswin Shanmugam Subramanian, Xiaofei Wang, Murali Karthick Baskar, Shinji Watanabe, et al., "Speech enhancement using end-to-end speech recognition objectives," in *WASPAA*, 2019, pp. 229–233.

[12] Shane Settle, Jonathan Le Roux, Takaaki Hori, Shinji Watanabe, et al., "End-to-end multi-speaker speech recognition," in *ICASSP*, 2018, pp. 4819–4823.

[13] Tobias Menne, Ralf Schlüter, and Hermann Ney, "Speaker adapted beamforming for multi-channel automatic speech recognition," in *IEEE SLT Workshop*, 2018, pp. 535–541.

[14] Yanmin Qian, Xuankai Chang, and Dong Yu, "Single-channel multi-talker speech recognition with permutation invariant training," *Speech Communication*, vol. 104, pp. 1 – 11, 2018.

[15] X. Chang, Y. Qian, K. Yu, and S. Watanabe, "End-to-end monaural multi-speaker ASR system without pretraining," in *ICASSP*, 2019, pp. 6256–6260.

[16] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *ICASSP*, 2016, pp. 31–35.

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

[18] Hiroshi Seki, Takaaki Hori, Shinji Watanabe, Jonathan Le Roux, et al., "A purely end-to-end system for multi-speaker speech recognition," in *Proc. of 56th ACL (Volume 1: Long Papers)*, 2018, pp. 2620–2630.

[19] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, et al., "Multi-channel overlapped speech recognition with location guided speech extraction network," in *IEEE SLT Workshop*, 2018, pp. 558–565.

[20] Fahimeh Bahmaninezhad, Jian Wu, Rongzhi Gu, Shi-Xiong Zhang, et al., "A Comprehensive Study of Speech Separation: Spectrogram vs Waveform Separation," in *Interspeech*, 2019, pp. 4574–4578.

[21] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, et al., "Compact network for speakerbeam target speaker extraction," in *ICASSP*, 2019, pp. 6965–6969.

[22] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, et al., "Single channel target speaker extraction and recognition with speaker beam," in *ICASSP*, 2018, pp. 5554–5558.

[23] Jun Wang, Jie Chen, Dan Su, Lianwu Chen, et al., "Deep extractor network for target speaker recovery from single channel speech mixtures," in *Interspeech*, 2018, pp. 307–311.

[24] Guanjun Li, Shan Liang, Shuai Nie, Wenju Liu, et al., "Direction-aware speaker beam for multi-channel speaker extraction," in *Interspeech*, 2019, pp. 2713–2717.

[25] Marc Delcroix, Shinji Watanabe, Tsubasa Ochiai, Keisuke Kinoshita, et al., "End-to-End SpeakerBeam for Single Channel Target Speech Recognition," in *Interspeech*, 2019, pp. 451–455.

[26] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *ICASSP*, 2015, pp. 708–712.

[27] M. Souden, J. Benesty, and S. Affes, "A study of the lcmv and mvdr noise reduction filters," *IEEE Transactions on Signal Processing*, vol. 58, no. 9, pp. 4925–4935, Sep. 2010.

[28] 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*. Association for Computational Linguistics, 1992, pp. 357–362.

[29] Jont B Allen and David A Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[30] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, et al., "ESPnet: End-to-end speech processing toolkit," in *Interspeech*, 2018, pp. 2207–2211.

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