MULTI-CHANNEL OVERLAPPED SPEECH RECOGNITION WITH LOCATION GUIDED SPEECH EXTRACTION NETWORK

Zhuo Chen, Xiong Xiao, Takuya Yoshioka, Hakan Erdogan, Jinyu Li, Yifan Gong

Microsoft Corporation, USA

ABSTRACT

Although advances in close-talk speech recognition have resulted in relatively low error rates, the recognition performance in far-field environments is still limited due to low signal-to-noise ratio, reverberation, and overlapped speech from simultaneous speakers which is especially more difficult. To solve these problems, beamforming and speech separation networks were previously proposed. However, they tend to suffer from leakage of interfering speech or limited generalizability. In this work, we propose a simple yet effective method for multi-channel far-field overlapped speech recognition. In the proposed system, three different features are formed for each target speaker, namely, spectral, spatial, and angle features. Then a neural network is trained using all features with a target of the clean speech of the required speaker. An iterative update procedure is proposed in which the mask-based beamforming and mask estimation are performed alternatively. The proposed system were evaluated with real recorded meetings with different levels of overlapping ratios. The results show that the proposed system achieves more than 24% relative word error rate (WER) reduction than fixed beamforming with oracle selection. Moreover, as overlap ratio rises from 20% to 70+%, only 3.8% WER increase is observed for the proposed system.

Index Terms— speech recognition, multi-speaker overlapped speech, audio separation, speech extraction, multi-channel processing

1. INTRODUCTION

Advances in deep learning have brought remarkable improvements to automatic speech recognition in the past decade, especially for close-talking recordings, where the speaker is usually less than 50-cm away from the microphone. In this scenario, the state-of-the art system was claimed to reach the human level performance on certain datasets [1, 2]. However, when the speech is recorded in a far-field setup where the speaker-microphone distance can be greater than 1 m as can be seen in smart speaker or public surveillance scenarios, the task is much more challenging. Even the state-of-the-art recognition systems suffer from insufficient performance.

Compared to the close-talk scenario, there are three additional acoustic challenges in far-field speech recognition. Firstly, the far-field speech usually has lower signal-to-noise ratio (SNR) than close-talk speech. Secondly, lower direct sound-to-reverberation ratio (DRR) is another factor that contributes to the high recognition error rate. Lastly, multi-talker overlapped speech occurs frequently in far-field recording environments, e.g. in meetings where multiple people are talking at the same time. The overlapped speech breaks the fundamental assumption in the modern ASR system that there is only one active speaker at a time, thus making the recognition especially difficult.

To address these three challenges, speech separation methods based on neural networks [3, 4, 5, 6, 7, 8] and beamforming [9, 10, 11, 12] are usually applied. Beamforming utilizes the spatial information collected from multiple microphones to enhance the target speech, while the neural networks learn the regularities in speech magnitude spectra to separate speakers. The separated speech is then passed to the acoustic model for recognition.

Although both approaches are effective to some extent, they suffer from inherent limitations. As a linear filter, the beamforming has limited spatial discrimination and cancellation power for the interfering audio source, especially when the number of microphones is small. And for speech separation networks such as deep clustering (DC) [5] or permutation invariant training network (PIT) [3], because they usually maintain multiple speakers in their internal memory and output simultaneously, their performance and generalization tends to be limited. A more detailed discussion of the neural network based speech separation is given in Section 2.

In this work, we propose a simple yet powerful approach for far-field overlapped speech recognition. Different from the blind separation networks such as DC or PIT where the label permutation ambiguity [4] is mainly handled by specially designed objective function, in the proposed framework a location based angle feature is extracted for each speaker in the speech mixture, and then processed to estimate the ratio mask for each target speaker by a uni-directional long short-term memory (LSTM) recurrent neural network (RNN). The proposed system removes the dependency between the number of mixing speakers and network complexity, thus leading to potentially better reconstruction of the target speaker and the generalization in complex acoustic environments.

The rest of the paper is organized as follows, in Section 2, a brief overview of the neural network based speech separation is discussed. The proposed model is described in detail in Section 3. Section 4 describes the experiment setup. The results are discussed in Section 5, followed by concluding remarks in Section 6

2. OVERVIEW OF MULTI-TALKER SPEECH SEPARATION

The main challenge in overlapped speech separation lies in the label permutation problem [4]. When there are multiple speakers talking simultaneously, the separated output have random orders, which causes ambiguity in pairing with the reference and prevents the data-driven method from having correct gradients.

To handle this problem two families of algorithm were proposed in recent years, namely the blind speech separation [5, 10, 3, 4] and informed speech extraction [13, 14, 15]. The two families handle the permutation from different perspective, and both achieved high quality separation performance.

In blind separation, usually the only observation is the mixture audio. To address the permutation problem, a specially designed

network objective function is usually applied. The most representative ones are deep clustering (DC) and permutation invariant training (PIT). In DC [5], the objective function is designed to focus on the local affinities between time-frequency bins, with no global assignment of the source, thus avoiding the label ambiguity. The PIT systems [3] diminish the ambiguity by exhaustively searching all possible configurations of output-to-reference pairing and find the optimal one for network optimization. Several updates were proposed based on those models, including the end-to-end optimization [16] and the multi-channel extension [10, 17].

As no further information is available to distinguish the individual speakers, all speakers in the mixture are handled with equal emphasis in blind separation systems. In other words, the blind separation system is usually required to estimate the separation for each source simultaneously. Therefore, the blind separation can be viewed as an "unbiased" separation. The main limitation of this family is that the separation for each source is usually sub-optimal, due to the equal consideration for all of them simultaneously. And because of the same reason, their separation performance drops significantly when more speakers are involved [7].

In contrast, in the informed speech extraction systems, an additional source of information was assumed available, which helps to identify each involving speaker and remove the uncertainty in permutation from the input feature perspective. Several clues have been shown to be helpful. In [13, 14, 18], speaker identity features extracted from an additional enrollment utterance has been shown useful for separation. In [15, 19], vision clue has been explored. And in [20, 21], the location based clue has also been shown to be beneficial.

When an additional clue is provided, during the separation, the network usually has a clear bias toward certain speakers. Therefore this type of speech extraction can be viewed as "biased separation". The limitation for this family is also obvious: when the bias signal can not provide sufficient bias, the system will fail to separate speakers. For example in a visual bias system, the separation system will entirely fail if the face cannot be detected, or occluded by other speakers.

3. INFORMED SPEECH EXTRACTION

3.1. Speech extraction network

In this work, following the informed speech extraction approach, we utilize the location information as the bias signal, and propose a system to extract the target speaker out of the speech mixture.

We assume that the location of each speaker in the speech mixture is known. This assumption can be achieved through various possibilities in real world applications, such as surveillance video, indoor GPS, or from the sound source localization system as proposed in [22, 23].

A schematic diagram of the proposed system is shown in Fig. 1. The proposed model adopts a similar framework as mask learning systems [24, 25, 26], where a mask is estimated for the target speaker through a neural network. And depending on the usage, the mask is used to mask out the interfering sources, or used for mask based beamforming. In the proposed system, three different types of features are calculated from the multi-channel recordings, which are referred to as spectral, spatial, and angle features.

As with the traditional mask estimation network [24, 25, 26], the spectral feature aims to extract the spectral structure in speech, such as harmonics and pitch continuity. To increase the discrimination between speakers, a set of fixed beamformers as in [20, 7] was applied to pre-process the multi-channel recordings, and the magnitude

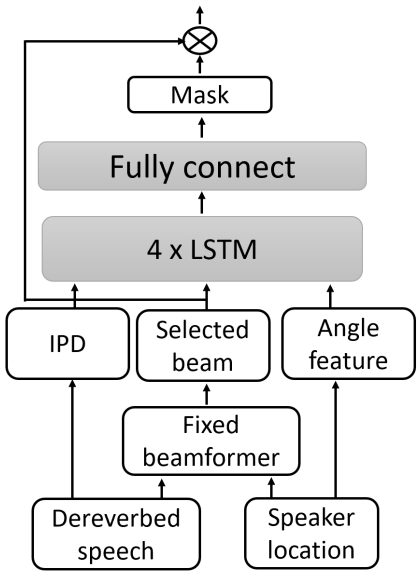

Fig. 1. Proposed informed speech extraction network.

spectrogram of the beam that points to the target speaker was used to compute the spectral feature.

The spatial feature models the correlation between the multi-channel signal and carries the spatial location information of the speech sources. Although [27] shows that such information could potentially be discovered through large scale training, it is more useful and efficient to directly extract the spatial information, as suggested in [28]. Following the same recipe as in [28], an inter-microphone phase difference (IPD) is calculated as the spatial feature as follows

$$\text{IPD}_{i,tf} = \angle(\frac{y_{i,tf}}{y_{1,tf}}), i = 2...M \tag{1}$$

where y refers to the observed data in the frequency domain, M denotes the number of microphones being used, i indexes the microphone, and tf indexes time-frequency bins. $y_{i,tf}$ is the i-th channel complex spectrum of the mixture signal at time frame t and frequency bin f. $\angle(\cdot)$ outputs the angle of the input argument. The IPD feature captures the relative phase difference between microphones, which reflects the time difference of arrival (TDOA), i.e., the spatial sound field information. An utterance-level normalization was applied to IPD in the same way as in [10]. The IPDs between the first microphone and all the other microphones are concatenated to be used as the final IPD feature.

As discussed in the Section 2, the bias signal helps to create the target specific feature for later extraction. In this work, we utilize the speaker location to form the bias signal, which we refer to as an angle feature. To get the angle feature, we first form the steering vector for the direction-of-arrival (DOA) of each speaker. Then, the cosine distance between the steering vector and the complex spectrum of each channel that is normalized with respect to the first microphone is calculated as follows:

$$A_{n,tf} = \sum_{i=1}^{M} \frac{e_n^{i,f} \frac{y_{i,tf}}{y_{1,tf}}}{\left| e_n^{i,f} \frac{y_{i,tf}}{y_{1,tf}} \right|} \tag{2}$$

where n is the speaker index, $e_n^{i,f}$ is the steering vector coefficient for speaker n's DOA at microphone i and frequency bin f. Intuitively, the angle feature lets the network to attend the sound coming from the direction of a certain speaker. The idea of the angle feature is similar to beamforming while it is different from the traditional beamforming in that non-linear processing is performed with a neural network.

An additional pre-masking step was applied to increase the discrimination resolution between different angles. Motivated by the sparsity property of speech spectrogram where most of the time-frequency bins are dominated by a single speaker or noise, each bin is turned on (i.e. set to non-zero value) at most once among all the speakers. Specifically, for the angle feature of a speaker, we only keep the bins that has the maximum value among all speakers and set the rest to 0, as suggested in eqn. 3, where $\mathcal{I}(\cdot)$ is the indicator function that outputs 0 if the input is negative and 1 otherwise.

$$A_{n,tf} = A_{n,tf} * \mathcal{I}(A_{n,tf} - A_{m,tf}), m = 1...N \tag{3}$$

Finally, a neural network is trained to recover the voice of the target speaker through masking using a signal reconstruction loss function: $Loss = \sum_{tf} \|x_{tf} - m_{tf} * y_{tf}\|^2$, where x is the clean spectrogram of target speaker and m is the estimated mask and y is the input selected beam. To form the clean reference for target speaker, the same fixed beamforming is firstly applied on the clean version of the target speech, and then the beam pointing to that speaker is selected based on the location information to form the reference spectrogram.

During testing, dereverberation is also performed by using the weighted prediction error (WPE) method [29, 30], before the speech extraction, to further improve the robustness against reverberation.

3.2. Multi-pass Mask Update

Although enjoying the advantage of low latency and high robustness, the fixed beamforming used to extract spectral features as shown in Fig. 1 usually has less interference-cancelling power than adaptive beamforming such as minimum variance distortionless response (MVDR) beamformer when the signal statistics are abundant. To fully utilize the benefits of beamforming, we propose a second pass strategy. The speaker masks estimated by the network are used to build a mask-based MVDR beamformer for each speaker, and the beamformed signal are used to replace the beam selected based on speaker's location in Fig. 1 to generate the second-pass speaker masks. This process can be repeated until the estimated masks converge.

To build mask-based MVDR, a target mask and interfering mask are required for each speaker. We use the mask from the first pass as the target mask. The average of all other speakers' masks and the inverse target mask ($m_{\overline{n}} = 1 - m_n$) is used as the interfering mask. Following [12], a threshold of 0.6 was applied on both target and interfering mask to increase the robustness of the estimated beamforming filters.

With the masks, the spatial covariance matrices for the target and interference are estimated as

$$\Phi_{n,f} = \frac{1}{\sum_t m_{n,tf}} \sum_t m_{n,tf} Y_{tf} Y_{tf}^H \tag{4}$$

$$\Phi_{\overline{n},f} = \frac{1}{\sum_t m_{\overline{n},tf}} \sum_t m_{\overline{n},tf} Y_{tf} Y_{tf}^H \tag{5}$$

where $Y_{tf} = [y_{1,tf}, ..., y_{M,tf}]^T$ is the observed mixture Fourier coefficients at time frequency bin tf, $\Phi_{n,f}$ and $\Phi_{\overline{n},f}$ refers to the spatial covariance for the target and interfering sources and $m_{n,tf}$, $m_{\overline{n},tf}$ represents the target and interfering masks, respectively. The MVDR weights can be obtained as

$$\mathbf{w}_{n,f} = \frac{\Phi_{\overline{n},f}^{-1} \Phi_{n,f} \mathbf{e}}{\text{tr}(\Phi_{\overline{n},f}^{-1} \Phi_{n,f})} \tag{6}$$

where $\mathbf{e} = [1, \mathbf{0}_{M-1}]^T$ and $\mathbf{0}_{M-1}$ is a row vector with $M-1$ zeros. $\mathbf{w}_{n,f}$ is the MVDR weight vector for speaker n at frequency bin f. The beamformed signal is obtained as $u_{n,tf} = \mathbf{w}_{n,f}^H Y_{tf}$.

The MVDR step achieves better spatial discrimination for target speakers compared to the fixed beamformer. To further improve the performance, the MVDR beamformed signal is feed back to the network for second pass processing, with spatial and angle feature unchanged. The full diagram of the multi-pass procedure is shown in Fig. 2

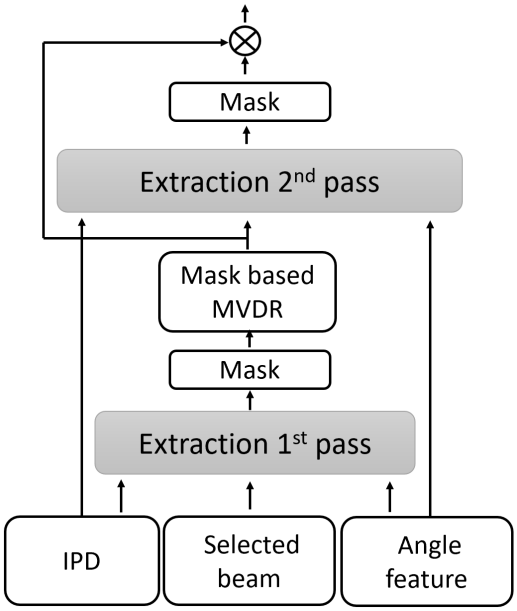

Fig. 2. The multi-pass update strategy.

The multi-pass update further improves the separation performance by utilizing the adaptive beamforming. However, such improvement comes with the price of additional computational cost and potentially longer latency, as the adaptive beamformer usually requires a long observation window to estimate the spatial covariances robustly, while the masking on fixed beamforming approach can generate instantaneous result. Although in theory the iterative updates of MVDR and mask can be carried on for many passes for the best results, as suggested in [31], its computational complexity also increases proportionally, which would be problematic in real-time applications. Therefore in this work, we only use two passes, i.e. the whole process finishes after the mask estimation on MVDR beamformed audio as shown in Fig. 2.

3.3. Model analysis

Compared with the multi-channel (MC) blind speech separation network such as MC-PIT and MC-DC, the proposed system has two advantages, making it appropriate for real-world application.

Firstly, the proposed framework removes the dependency between network computation and the acoustic complexity, which is the main limitation for the blind separation network. In blind separation systems, the network is trained to maintain the separation of all participating sources, which usually results in sub-optimum reconstruction for each individual source. Moreover, due to the same reason, the computational complexity grows exponentially with the number of speakers. For example, when there are more than 3 speakers in the mixture [7], it is extremely difficult for blind separation to maintain high quality reconstruction for any participant. In contrast, the proposed system is trained to only recover the target speaker, treating all other speakers as one interference. This architecture allows the network to focus on a specific speaker and may achieve better performance. This design also increases the system robustness and generalization when there are more speakers. In the experiments, we observed that although the network was only trained on two-speaker mixture, it works robustly for even six-speaker mixture, when the difference between target speaker and interfering speaker angles is more than 20 degree. This property makes the proposed model especially suitable for public recording with large number of simultaneous speakers, such as those in restaurants or malls.

Another advantage of the proposed model lies in its flexibility in choosing network architecture thanks to the reference signal. As shown in [3, 16], the Bi-directional LSTM is important to high quality performance for blind separation networks. This is because in blind separation (e.g. permutation invariant training for two speaker separation), the only clue for separation is from the local continuity within the speech, and each source can be assigned to either output. Therefore, to maintain the global coherence, a long window of observation with future data is usually required. In contrast, the proposed network utilize the angle feature as additional reference, which helps to maintain both the local and global coherence. Thus it enables more flexibility in architecture choice.

On the other hand, the proposed system also suffers from two limitations that need to be improved in further works. Firstly, as the biased signal is calculated entirely from the spatial information, when the speakers are close, e.g. less than 20 degree, the location bias will not be sufficient to distinguish target speaker, and the permutation ambiguity will again hinder the separation process. A possible solution to this could be combination with additional bias signal e.g. speaker ID or separation network. Additionally, as the proposed network only extracts one speaker each time, the total computation is proportional to the number of participants in the mixture, which is more expensive than the blind separation system, where the network is evaluated only once to obtain the voice of all speakers. This problem could be alleviated by using pruning mechanism based on sound source localization, i.e. only run the extraction when a sound is detected.

3.4. Comparison with other systems

The idea of using location information for multi-channel speech separation has been explored in the signal processing community. Most of the methods investigated are based on independent component analysis or spatial clustering which requires prior knowledge of the speakers and assumes that all the speakers do not move (e.g., [32]). [20, 21] also explore the idea of using spatial features to enhance the neural network mask estimation, by feeding the concatenated beam of each speaker. The limitation of this strategy is that the number of speakers has to be pre-defined before the network training, which makes it difficult to be used for real world applications. The IPD features were discussed in [10, 17], and served as the spatial feature for blind speech separation network. And a similar multi-pass strategy was discussed in [12, 31], in a single speaker denoising task. To the best of our knowledge, the proposed framework is the first system for multi-channel overlapped speech recognition with location based features that achieves high quality separation on real recorded data.

4. EXPERIMENT

4.1. Network training

The full evaluation pipeline consists of two parts: the speech extraction part and speech recognition part. In this work, the two parts were trained separately and work in a sequential manner. The multi-channel recording was firstly processed by the proposed speech extraction network. After the extraction step, the masked beam was converted back to the time domain, and then processed by the acoustic model for recognition.

A simulated multi-channel overlapped speech dataset was created for the training of the extraction network. The training set consists of 300 hours of mixture speech. The clean utterance was randomly sampled from libri-train-360 set in Libri speech dataset [33]. For each mixture sample, two clean utterances from different speakers were sampled and combined with various overlap ratio and with random signal to interference ratio between -2.5 and 2.5 dB.

The same array geometry as in [7] was employed for both recording and simulation The image method was applied to generate the room impulse response, where the room size was picked from 2m to 20m in length and width, 2$\sim$5m in height with random T60 between 0.1 and 0.9s. The locations of the microphone array and speaker were randomly selected. We ensured the minimum angle between two speakers was no less than 20 degree. The directional and isotropic noise were added to each sample. For directional noise, the noise data was sampled from MUSAN [34] and Chime-3 [35] noise dataset. We randomly added up to 4 directional noise to each training sample, with random SNR between 5dB to 30dB and random location for each noise source. For the isotropic noise, we use the method in [36] to simulate the spatially correlated white noise, with random SNR between 20 to 40dB. All the data has sampling rate of 16kHz. All feature for speech extraction was calculated with 32ms window and 10ms frame shift.

The speech extraction network used in our experiments consisted of 4 LSTM layers, each with 1,024 memory cells. A fully connected layer with sigmoid activation function was added to generate the mask.

For ASR, we trained an acoustic model on 7,000 hours of transcribed spontaneous speech, which were collected from various sources, both public (e.g., Switchboard and Fisher) and private (e.g., Microsoft Research lecture talks). The model input was 40-channel Mel filterbank energies compressed with 10th-root nonlinearity. The model consisted of four 1,024-unit LSTM layers. It was trained with a cross entropy criterion, followed by sequence training. Decoding was performed with a dictionary of 240K words and our internal trigram language model built for conversational tasks.

4.2. Evaluation process

For evaluation, we collected eight recordings from the regular meeting inside Microsoft.

Among all recordings, two meetings are recorded from the conversation between real human speakers, in one Microsoft meeting room that has a T60 around 0.4s. All speaker are required to wear

System & Meeting	MT01	MT02	RP01	RP02	RP03	RP04	RP05	RP06
Close talk	21.2	24.4	21.97	22.3	22.65	22.32	21.44	22.1
Raw recording	43.3	48.2	52.3	58.87	64.32	71.08	77.31	83.96
WPE + fixed beamforming	33.7	36.4	37.86	41.24	42.87	48.21	50.03	55.79
Speech extraction	31.5	34.4	35.38	38.26	38.94	42.09	45.51	49.86
Mask MVDR	32.0	34.0	32.58	34.86	36.46	37.06	39.38	42.71
Multi-pass extraction	30.8	33.3	32.00	32.4	32.63	32.82	34.29	35.81

Table 1. Word error rates (%) on real and replayed meetings.

a close talk microphone, and the microphone array was placed on the table, surrounded by participants. The speakers were free to discuss any topic, and the direction of the speakers was labeled by a moderator. Each meeting has around 30 minutes in length.

Another six recordings were collected from the replaying of the close talk recording from $MT02$, through loudspeakers. The replay was taken place in another meeting room with similar room size, but with around 0.5s T60. For each speaker, the total number of words spoken in the meeting remain unchanged across meeting with different overlap ratio.

Different overlap ratio was applied in the replayed data. We follow the definition of overlapping ratio (OVR) in [37], i.e. $OVR = \frac{L_{overlap}}{L_{total}}$, where $L_{overlap}$ is the total length of the overlapped speech and L_{total} total speech length.

To create the overlapped recording, we firstly edit the close talk recording from $MT02$, by random shifting the onset of each utterance from each speaker. Note that the utterance sequence for each speaker was not changed. We also avoid the overlap from the same speaker, i.e. a short interval was added to two consecutive utterances of one speaker, if they overlapped in time. Then the edited close talk speech was replayed through six loudspeakers, and recorded with the same microphone array. All six replayed recordings used the same microphone and speaker locations, i.e. the only difference among them is the overlap ratio.

Note that in the overlap ratio calculation, we didn't differentiate overlap type with different number of mixing speaker. Therefore, the overlap contains various mixing speakers from 2 to 6. For example, in $RP06$, the ratio from 2 speaker mixture to 5 speaker mixture is 70.3%, 27.0%, 7.5% and 0.4%. The full configuration of each meeting is given in Table 2.

The main difference between real meeting and replayed meeting lies in the speaker head movement. In real meeting $MT01$ and $MT02$, we can observed a frequent head movement for around 5 to 10 degree. This phenomenon potentially increases the data variation and processing difficulty. However, since the movement is not significant, we believe that the replayed data posed similar realistic challenge.

The word error rate(WER) was used as the evaluation metric.

ID	Words	OVR(%)	Speaker	Average angle
MT01	6096	16.3	4	25,190,270,310
MT02	5158	16.0	6	85,151,170,212,235,311
RP01	5158	20.0	6	0,60,90,157,205,270
RP02	5158	30.1	6	0,60,90,157,205,270
RP03	5158	40.0	6	0,60,90,157,205,270
RP04	5158	50.2	6	0,60,90,157,205,270
RP05	5158	60.6	6	0,60,90,157,205,270
RP06	5158	70.3	6	0,60,90,157,205,270

Table 2. The configuration for recorded meetings.

The data was evaluated per utterance and the average WER was reported. We reported the evaluation with six setups. Firstly, we reported the WER on the close talk and raw recording, which is the upper limit and the original performance on those data. Then a fixed beamforming with dereveberation was used as the baseline, as it is the general setup for current commercial speakers. We use the beamformer from [7]. The beam to each speaker was selected for that speaker. Finally, the performance of proposed speech extraction and two of its variation were reported.

For MVDR and multi-pass experiment, within each meeting, we firstly process the whole meeting for each speaker with the extraction network and collect the mask. Note that when the angle between a speaker pair is less than 30 degree, the pre-masking step was not performed for that close interfering speaker. Then for each speaker, the utterances were clustered according to their DOA. Within each group, we ensure that the maximum DOA difference is less than 20 degree. The the spatial covariance matrix for each group was estimated and used for estimating MVDR weights. Therefore, in the multi-pass setup, the earlier utterance might benefit from the statistics form later ones, i.e. the system is offline. In contrast, since the speech extraction network used the LSTM network, the first pass can be viewed as online processing since no future data was used.

5. RESULT AND DISCUSSION

The evaluation result is shown in Table 1. From the result, we can see that the proposed model and its variation significantly outperformed by on average 27.6% than the beamforming baseline in all meeting. This advantage becomes stronger as more overlapped speech involved, from 15.4% to 35.8%. The improvement is more significant when comparing with signal channel data, which is on average 51% relatively.

As the same mask was used in both the speech extraction and the mask based MVDR, the direct comparison between them is available here. The mask-based MVDR achieved better performance in replayed data, while similar performance was observed in real meeting. This is natural as in real meetings, the small head movement influenced the accumulation of statistics for MVDR, while the masking step doesn't require the history accumulation, i.e. no theoretical latency. It has been shown in previous work that the masking step could result in distortion[12], and this can be fixed through the AM-retraining or joint training. Then the speech extraction network could have better performance, making it more suitable for real world applications.

For replayed meetings, an obvious trend is that the performance gap increase as overlap ratio rises. In the multi-pass system, only 3.81% absolute increase (12% relative), is observed between RP01 and RP06. This observation suggests that the speech overlapping problem can be well addressed with the proposed model. Moreover, as discussed above, we didn't specify the mixing type. Thus a large ratio of more than 3 speaker mixture can be observed in high overlap

ratio meetings. Although trained with only two speaker mixture, we observed that the estimated mask can robustly handle different types of overlapping, from single speaker segments to six speaker mixture across meetings.

When comparing the speech extraction in different passes, the second pass system has shown a clear benefit in terms of performance. This confirmed the benefits of adaptive beamforming. Since the second pass can be considered as an offline process, it has more flexibility in more advanced models, such as the bi-directional LSTM, which can potentially further improve the performance.

Finally, although the proposed system largely improved the recognition rate, a clear gap can still be observed between the results of close talking and the multi-pass extraction, which suggests the room for improvement in far-field speech recognition. Since all steps involved in the processing are differential, an obvious solution is the joint training, i.e. a network containing components, namely first pass, MVDR, second pass and AM. This possibility will be further explored in the future work.

6. CONCLUSION

In this work, we introduced a multi-channel overlapped speech recognition system. In the system, three different features were formed, representing respectively the spectral, spatial and location feature for each participating speaker. The features were then processed through an LSTM neural network, outputting the clean mask for the target speaker. To further enhance the separation and recognition, a multi-pass update strategy that includes an adaptive MVDR beamforming and neural network re-evaluation step was introduced. The proposed system was evaluated with various real recorded meetings with different overlap ratios. The result showed that the proposed system significantly outperformed the baseline. Only minor recognition degradation can be observed when the overlap ratio for each meeting was increased from 20% to 70%.

7. REFERENCES

[1] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Michael L Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig, "Toward human parity in conversational speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 12, pp. 2410–2423, 2017.

[2] George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, et al., "English conversational telephone speech recognition by humans and machines," *arXiv preprint arXiv:1703.02136*, 2017.

[3] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 241–245.

[4] Zhuo Chen, Yi Luo, and Nima Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 246–250.

[5] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 31–35.

[6] Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R Hershey, "Single-channel multi-speaker separation using deep clustering," *arXiv preprint arXiv:1607.02173*, 2016.

[7] Zhuo Chen, Jinyu Li, Xiong Xiao, Takuya Yoshioka, Huaming Wang, Zhenghao Wang, and Yifan Gong, "Cracking the cocktail party problem by multi-beam deep attractor network," in *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*. IEEE, 2017, pp. 437–444.

[8] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, Jesper Jensen, Morten Kolbaek, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 10, pp. 1901–1913, 2017.

[9] Manohar N Murthi and Bhaskar D Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 221–239, 2000.

[10] Takuya Yoshioka, Hakan Erdogan, Zhuo Chen, and Fil Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018.

[11] Hakan Erdogan, John R Hershey, Shinji Watanabe, Michael I Mandel, and Jonathan Le Roux, "Improved mvdr beamforming using single-channel mask prediction networks.," in *Interspeech*, 2016, pp. 1981–1985.

[12] Xiong Xiao, Shengkui Zhao, Douglas L Jones, Eng Siong Chng, and Haizhou Li, "On time-frequency mask estimation for mvdr beamforming with application in robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP),*

2017 IEEE International Conference on. IEEE, 2017, pp. 3246–3250.

[13] Marc Delcroix, Katerina Zmolikova, Keisuke Kinoshita, Atsunori Ogawa, and Tomohiro Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018.

[14] Kateřina Žmolíková, Marc Delcroix, Keisuke Kinoshita, Takuya Higuchi, Atsunori Ogawa, and Tomohiro Nakatani, "Learning speaker representation for neural network based multichannel speaker extraction," in *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*. IEEE, 2017, pp. 8–15.

[15] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *arXiv preprint arXiv:1804.03619*, 2018.

[16] Zhehuai Chen and Jasha Droppo, "Sequence modeling in unsupervised single-channel overlapped speech recognition," ICASSP, 2018.

[17] Zhong-Qiu Wang, Jonathan Le Roux, and John R Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," 2018.

[18] Katerina Zmolikova, Marc Delcroix, Keisuke Kinoshita, Takuya Higuchi, Atsunori Ogawa, and Tomohiro Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures," in *Interspeech*, 2017.

[19] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "The conversation: Deep audio-visual speech enhancement," *arXiv preprint arXiv:1804.04121*, 2018.

[20] Zhuo Chen, Yoshioka Takuya, Xiao Xiong, Li Jinyu, L. Seltzer Michael, and Gong Yifan, "Efficient integration of fixed beamformers and speech separation networks for multi-channel far-field speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018.

[21] Lauréline Perotin, Romain Serizel, Emmanuel Vincent, and Alexandre Guérin, "Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings," in *ICASSP 2018-IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[22] Xiong Xiao, Shengkui Zhao, Xionghu Zhong, Douglas L Jones, Eng Siong Chng, and Haizhou Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2814–2818.

[23] Fabio Vesperini, Paolo Vecchiotti, Emanuele Principi, Stefano Squartini, and Francesco Piazza, "A neural network based algorithm for speaker localization in a multi-room environment," in *Machine Learning for Signal Processing (MLSP), 2016 IEEE 26th International Workshop on*. IEEE, 2016, pp. 1–6.

[24] Zhuo Chen, Yan Huang, Jinyu Li, and Yifan Gong, "Improving mask learning based speech enhancement system with restoration layers and residual connection," in *Proc. Interspeech*, 2017.

[25] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[26] Felix Weninger, John R Hershey, Jonathan Le Roux, and Björn Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proceedings 2nd IEEE Global Conference on Signal and Information Processing, GlobalSIP, Machine Learning Applications in Speech Processing Symposium, Atlanta, GA, USA*, 2014.

[27] 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.

[28] Takuya Yoshioka, Hakan Erdogan, Zhuo Chen, Xiao Xiong, and Fil Alleva, "Recognizing overlapped speech in meetings: a multichannel separation approach using neural networks," in *Proc. Interspeech*, 2018.

[29] Takuya Yoshioka and Tomohiro Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[30] Bo Li, Tara Sainath, Arun Narayanan, Joe Caroselli, Michiel Bacchiani, Ananya Misra, Izhak Shafran, Hasim Sak, Golan Pundak, Kean Chin, et al., "Acoustic modeling for google home," *INTERSPEECH-2017*, pp. 399–403, 2017.

[31] Liu Yuzhou, Ganguly Anshuman, Kamath Krishna, and Kristjansson Trausti, "Neural network based time-frequency masking and steering vector estimation for two-channel mvdr beamforming," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018.

[32] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 9, pp. 1913–1928, 2013.

[33] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

[34] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[35] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, "The third chimespeech separation and recognition challenge: Dataset, task and baselines," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 504–511.

[36] Emanuël AP Habets and Sharon Gannot, "Generating sensor signals in isotropic noise fields," *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3464–3470, 2007.

[37] Özgür Çetin and Elizabeth Shriberg, "Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: insights for automatic speech recognition," in *Ninth International Conference on Spoken Language Processing*, 2006.