ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446168

A SPATIAL LONG-TERM ITERATIVE MASK ESTIMATION APPROACH FOR MULTI-CHANNEL SPEAKER DIARIZATION AND SPEECH RECOGNITION

Feng Ma[1], *Yanhui Tu*[2], *Maokui He*[1], *Ruoyu Wang*[1],
Shutong Niu[1], *Lei Sun*[2], *Zhongfu Ye*[1], *Jun Du*[1,*], *Jia Pan*[2], *Chin-Hui Lee*[3]

[1] University of Science and Technology of China, China
[2] iFlytek Research, China [3] Georgia Institute of Technology, USA

ABSTRACT

Deep learning (DL)-based speaker diarization methods have proven powerful performance comparing to traditional clustering-based methods for multi-talker speech diarization and recognition in far-field scenes. However, most DL-based approaches cannot utilize the spatial information well due to the poor robustness to unknown array topology and acoustic scenario. In this paper, a spatial long-term iterative mask estimation (SLT-IME) method is proposed to improve the performance of speaker diarization in various real-world acoustic scenarios. First, the complex angular central gaussian mixture model (cACGMM) with diarization results as initial values is used to estimate the presence probability of each speaker at each time-frequency bin, namely speaker masks, in a long-term chunk. Then, the speaker masks are converted to speaker activities according to the threshold, which deliver the diarization information of which speaker is active and when. Finally, the estimated speaker activity can also serve as the initial input for the diarization system, resulting in improved ASR performance. Experimental results on the CHiME-7 three datasets (CHiME-6, DiPCo, Mixer 6) show proposed method can improve diarization and recognition systems performance simultaneously. It also plays a key role in the ensemble system that achieves the best performance in the main track of CHiME-7 DASR Challenge.

Index Terms— Speaker diarization, multi-channel speech enhancement, iterative mask estimation, CHiME-7 Challenge

1. INTRODUCTION

Automatic speech recognition (ASR) in distant-talking scenarios based on the use of microphone arrays has become an important part of everyday life with the emergence of speech-enabled applications on multi-microphone portable devices due to its convenience and flexibility [1]. Several limited tasks were initially investigated, including the TIdigits corpus [2], the TIMIT [3], the Wall Street Journal (WSJ) [4] and the LibriSpeech [5] corpora. However, these tasks do not take into account noisy or reverberant conditions. The CHiME (1-4) [6, 7, 8] series were launched to investigate the impact of background noises in far-field scenarios, which aimed to address ASR challenges in real-world applications. To enhance ASR robustness, a common approach is to use multi-channel speech enhancement as the front-end system. This category includes representative algorithms such as multi-channel Wiener filtering [9], blind source separation methods [10, 11, 12, 13], and beamforming methods [14, 15, 16]. Beamforming has gained popularity in the CHiME-3 Challenge [17]. In the CHiME-4 Challenge, the best system introduced a novel approach that combines conventional multi-channel speech enhancement with deep learning methods [18] to improve multi-channel speech recognition.

The CHiME-5 [19] and CHiME-6 [20] Challenge have recently provided the first large-scale corpus of real multi-talker conversational speech recorded via commercially available microphone arrays in multiple realistic homes [19]. And the CHiME-6 challenge revisits the previous CHiME-5 challenge and further considers the problem of distant multi-microphone conversational speech diarization and recognition in everyday home environments. The latest CHiME-7 [21] task involves joint ASR and diarization in far-field settings using multiple recording devices, which may be heterogeneous. Unlike previous challenges, the evaluation of systems includes 3 diverse scenarios (CHiME-6, DiPCo, and Mixer 6). The objective is for participants to develop a single system that can adapt to different array geometries and use cases without any prior information. The official baseline report shows a diarization attributed word error ratio (DA-WER) of 55.30%, highlighting the difficulty of the CHiME-7 ASR task.

Previous methods usually optimize the front-end modules independently, such as using the diarization module and then using the GSS module to get the segment single-channel speech, ignoring the interaction of each module information and the use of long-term information. In this paper, we propose a novel method called spatial long-term iterative mask estimation (SLT-IME) to enhance speaker diarization performance in various real-world acoustic scenarios. Our approach leverages the complementary information from a spatial long-term (SLT)-based spatial mixture model (SMM) to iteratively improve diarization accuracy. While the neural speaker diarization using memory-aware multi-speaker embedding (NSD-MA-MSE)-based diarization approach has shown powerful performance for multi-talker environments, it often struggles with accurate speaker estimation in low signal-to-noise ratio (SNR) or same gender cases due to limited spatial information. To improve the performance of the diarization system, we first segment the multi-channel data and NSD-MA-MSE-based diarization results into overlapping long-term blocks. This ensures that each block contains sufficient speech from all speakers. Next, we utilize cACGMM to estimate each speaker mask for each block and get the long masks with overlap-add. The long masks are converted to each speaker activity by threshold. Finally, we use the obtained speaker activity as initial values for official cACGMM and then fed the corresponding beamformed speech to the recognition system. The estimated speaker activity can also serve as the initial input for the diarization system, resulting in improved ASR performance. When tested on CHiME-7 Challenge track 1 tasks (multiple-array

*corresponding author

speech recognition), our proposed SLT-IME approach further enhances both diarization and ASR performance. Additionally, this approach plays a crucial role in our ensemble system which achieves top performance in CHiME-7 Challenge main track tasks.

2. THE WHOLE FRAMEWORK

The Fig. 1 illustrates the entire framework, which primarily focuses on addressing the issue of multi-speaker diarization and speech recognition in various far-field scenarios where the array location and speaker information are unknown. The framework includes channel selection, diarization, spatial long-term iterative mask estimation (SLT-IME), and automatic speech recognition (ASR). To begin with, envelope variance measure (EV) is employed for channel selection to eliminate faulty channels (further details can be found in [22]). We initially employ the NSD-MA-MSE-based diarization approach, which is a neural speaker diarization using memory-aware multi-speaker embedding. This approach demonstrates strong performance and utilizes a framework similar to target-speaker voice activity detection (TS-VAD). The key distinction lies in the use of dynamic speaker embeddings obtained by weighting the original speaker embeddings in the memory block. For further details, refer to [23]. Next, the proposed SLT-IME approach is used to reduce the speaker error caused by the initial diarization system by leveraging long-term multi-channel signal processing. The spatial information contained in the multi-channel data not only makes the clustering more stable, but also makes the recognition performance better. The outputs of SLT-IME can be utilized for diarization system to extract more reliable speaker features, which can enhance the performance of NSD-MA-MSE-based diarization system through an iterative process. Finally, we employ speaker adaptive automatic speech recognition (SA-ASR) [24] as our recognition system to tackle the challenges in multi-speaker recognition task.

3. SIGNAL MODEL

In the short-time-Fourier-transform (STFT) domain [25], the signal model can be expressed as [26]:

$$\boldsymbol{Y}(l,f) = \boldsymbol{g}(k)S(l,f) + \boldsymbol{N}(l,f) = \boldsymbol{X}(l,f) + \boldsymbol{N}(l,f), \quad (1)$$

where f is the frequency bin index, l is the frame index, $S(l,f)$ is the STFT of clean speech. $\boldsymbol{X}(l,f)$ and $\boldsymbol{N}(l,f)$ are M-dimensional complex vectors that represent the STFT-domain representations of clean speech received by microphone arrays and noise signal, respectively. $\boldsymbol{g}(k)$ is the signal propagation vector, which is in the same form as the so-called steering vector in the literature of array beamforming [27]. We assume that the analysis window is longer than all the channel impulse responses and that $\boldsymbol{N}(l,f)$ is relatively stationary.

For the diarization system, we can get the K vectors of 2-class outputs, denoted as $\boldsymbol{D}$, as following:

$$\boldsymbol{D} = \mathrm{dia}(\boldsymbol{Y}_m; \boldsymbol{E}), \quad (2)$$

where $\boldsymbol{Y}_m$ is the m^{th} microphone data and K is the speaker number. $\boldsymbol{E} = [\boldsymbol{E}_1, ..., \boldsymbol{E}_K]$ is the speaker embeddings, and $\mathrm{dia}(\cdot)$ is the NSD-MA-MSE-based diarization algorithm. $\boldsymbol{D} = [\boldsymbol{D}_1, ..., \boldsymbol{D}_K]$, where $\boldsymbol{D}_k = [d(1,k), ..., d(L,k)]$ represents the probabilities of speech and silence of speaker k in each frame l. More detailed can refer to [23].

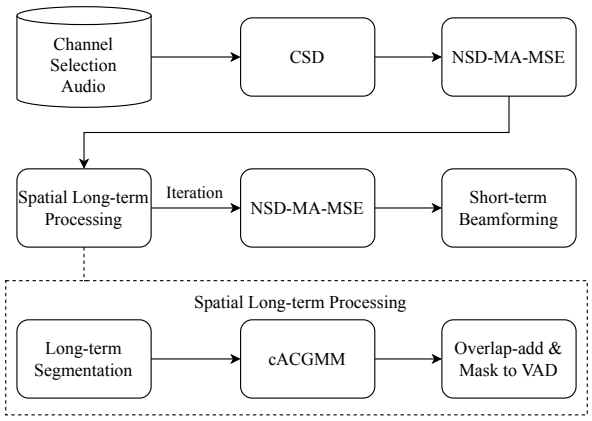

Fig. 1. The whole framework and proposed spatial long-term iterative mask estimation method.

4. PROPOSED SLT-IME METHOD

Compared to standard reading speech, conversational and spontaneous speech pose a greater challenge for speech recognition systems in dialogue scenarios. Casual pronunciation and frequent overlapping of speech significantly reduce the accuracy of acoustic models. The most significant hurdle is accurately identifying the speaker at any given moment. [23] introduces a novel diarization system that addresses this problem by utilizing speaker features and a deep learning model. However, diarization systems may experience reduced performance in cases where speech is heavily affected by background noises or when two speakers have similar acoustic properties (e.g., same gender). In order to improve robustness of the diarization system in multi-talker situations, a method based on spatial long-term iterative mask estimation (SLT-IME) is proposed.

The SLT-IME method consists of four stages: segmentation, long-term block processing, overlap-add, short-term beamforming. In the segmentation stage, sequential multi-channel audios and diarization results are segmented into long overlapped blocks. Then, each block is passed to cACGMM to estimate the presence probability of each speaker and noise, denoted as speaker mask. Next, the speaker masks in all blocks are concatenated into long sequential masks with overlap-add method and VAD for each speakers are generated by the long sequential masks. Finally, the estimated speaker activity can also serve as the initial input for the diarization system, resulting in improved ASR performance.

4.1. Long-term segmentation

The sequential inputs $\boldsymbol{Y}$ and $\boldsymbol{D}$ are split into blocks of length B (B is pre-defined) and hop size $W = B/2$, with 50% overlap. The length of last block is $L - \lfloor L/W \rfloor \times W$, where $\lfloor \cdot \rfloor$ is round down function. The inputs in each blocks are generated, denoted as $\boldsymbol{Y}_s$ and $\boldsymbol{D}_s, s = 1, ..., S$, where $S = \lceil L/W \rceil$ is the number of blocks, where $\lceil \cdot \rceil$ is round up function. To make each block containing enough speech for all speakers, the block size should be long enough.

4.2. Long-term block processing

The segmentation outputs $\boldsymbol{Y}_s$ and $\boldsymbol{D}_s$ are then passed to the spatial mixture model, namely the complex Angular Central Gaussian

Mixture Model (cACGMM) [28]:

$$p(\mathbf{z}(l,f)) = \sum_{k=1}^{K+1} \pi(l,f)\mathcal{A}(\mathbf{z}(l,f);\mathbf{B}(l,f)), \tag{3}$$

where $\pi(l,f)$ is a time-varying a priori probability and $\mathbf{z}(l,f) = \mathbf{Y}(l,f)/||\mathbf{Y}(l,f)||$. The number of classes $K+1$ is set to the number of speakers K plus one for the noise. The cACGMM parameters are trained using the EM algorithm, which involves alternating E- and M-steps. The update equations for the cACGMM are as follows:

$$\gamma(l,f,k) = \frac{\pi(l,f)\det(\mathbf{B}(l,f))^{-1}(\mathbf{z}(l,f)^{\mathrm{H}}\mathbf{B}(l,f)^{-1}\mathbf{z}(l,f))^{-N}}{\sum_{k=1}^{K+1}\pi(l,f)\det(\mathbf{B}(l,f))^{-1}(\mathbf{z}(l,f)^{\mathrm{H}}\mathbf{B}(l,f)^{-1}\mathbf{z}(l,f))^{-N}} \tag{4}$$

$$\mathbf{B}(l,f) = \frac{N}{\sum_{l=1}^{L}\gamma(l,f,k)}\sum_{l=1}^{L}\gamma(l,f,k)\frac{\mathbf{z}(l,f)^{\mathrm{H}}\mathbf{z}(l,f)}{\mathbf{z}(l,f)^{\mathrm{H}}\mathbf{B}(l,f)^{-1}\mathbf{z}(l,f)} \tag{5}$$

where $\gamma(l,f,k)$ is the a posterior probability that the class k is active at a time-frequency (l,f) bin. Based on the description of cACGMM algorithm, the EM algorithm converges to a local optimum of the objective function and is thus susceptible to its initialization. It is important to choose a appropriate initial values for the posterior $\gamma(l,f,k)$.

For block s, the posterior is initialled with the diarization results $d_s(l,k)$ as

$$\gamma_s^{\mathrm{init}}(l,f,k) = \frac{d_s(l,f,k)}{\sum_{k=1}^{K+1} d_s(l,f,k)}, \tag{6}$$

where $d_s(l,f,k)$ is a copy of $d_s(l,k)$ in the frequency dimension. Suppose that noise is present at any time, so $d_s(t,K+1) = 1, t = 1,...,T$. For current novelty diarization algorithm based on deep learning, there are few researches to explore how to utilize the spatial information for multi-array cases. Because it is hard to construct a training dataset containing kinds of array shapes and distributions for model training. cACGMM is an algorithm adaptive to the test signal and it can utilize the spatial information for every time steps. If one speaker are recognized as the other speakers or missed detection in a long-term block, the cACGMM model can correct the wrong speaker activity information while iterating. Therefore, it can further reduces the speaker error significantly in low SNR cases or the same gender of speakers.

4.3. Overlap-add

After the convergence of the EM algorithm initialized with diarization results in each block, the probability of speech presence at (l,f) bin is detected through the learned class posterior probabilities, denoted as $\gamma_s^{\mathrm{dia}}(l,f,k)$.

$$\gamma_s^{\mathrm{re}}(l,f,k) = \begin{cases} \gamma_s^{\mathrm{dia}}(l,f,k), & s=1 \\ [\gamma_{s-1}^{\mathrm{dia}}(l,f,k) + \gamma_s^{\mathrm{dia}}(l,f,k)]/2, & s>1, \end{cases} \tag{7}$$

where $(s-1)W \le l < sW$. All blocks are then concatenated together to form a 3-D tensor $\boldsymbol{\gamma}^{\mathrm{re}} = [\boldsymbol{\gamma}_1^{\mathrm{re}},...,\boldsymbol{\gamma}_S^{\mathrm{re}}] \in \mathcal{R}^{L\times F\times K}$. The final speaker activity at frame level, denoted as $M_{\mathrm{vad}}^{\mathrm{re}}(l,k)$, which also provides the diarization information of which speaker is active and when can be obtained as follow:

$$M_{\mathrm{vad}}^{\mathrm{re}}(l,k) = \begin{cases} 1, & \text{if } \beta(l,k) > 0.2, l = l-6,...,l \\ 0, & else, \end{cases} \tag{8}$$

where $\beta(l,k) = (\sum_{f=1}^{F}\gamma^{\mathrm{re}}(l,f,k))/F$ can provide the probability of speaker presence at each frame l. We also can repeat the whole process using the refined speaker activity $\boldsymbol{M}_{\mathrm{vad}}^{\mathrm{re}}$.

4.4. Short-term beamforming

To generate the speech for speech recognition, we also utilize the short-term beamforming, which is the official GSS process in CHiME-7. In this process, the whole sequential inputs $\boldsymbol{Y}$ is split into many short-term segments according to the diarization results. The beamformed speech at each segment, denoted as $\hat{\boldsymbol{X}}_{\mathrm{seg}}$, can be obtained as follow:

$$\hat{\boldsymbol{X}}_{\mathrm{seg}} = \mathrm{bf}(\boldsymbol{Y}_{\mathrm{seg}};\boldsymbol{M}_{\mathrm{seg}}), \tag{9}$$

where $\mathrm{bf}(\cdot)$ denotes the beamforming algorithm. $\boldsymbol{Y}_{\mathrm{seg}}$ and $\boldsymbol{M}_{\mathrm{seg}}$ denote the segment of $\boldsymbol{Y}$ and $\boldsymbol{M}_{\mathrm{vad}}^{\mathrm{re}}$.

5. EXPERIMENTS

5.1. Data corpus

The recently CHiME-7 Challenge provides the large-scale corpus of real multi-talker conversational speech recorded via commercially available microphone arrays in multiple realistic homes and meeting rooms [21]. This challenge contains three datasets: CHiME-6 [29], Dipco [30] and Mixer6 [31]. The three corpus essentially congregates a large number of acoustic problems that may exist in real life, which poses a great challenge to existing ASR systems, especially for the front-end processing in the case of noise, reverberation, overlapping speech. There are four main challenging of the corpus: 1) different array topologies: linear (CHiME-6, 6 Kinect array devices with 4 microphones each for a total of 24 microphones), circular (DiPCo, 5 far-field devices each with a 7-mic circular array (six plus one microphone at the center)), or heterogeneous (Mixer 6, 14 microphones of varying styles, placed in various locations); 2) variable numbers of speakers in each session; 3) linguistic differences between dinner party scenarios (CHiME-6 and DiPCo) versus interviews (Mixer 6); and (4) diverse acoustic conditions.

5.2. Implementation details

For front-end configurations, speech waveform is sampled at 16 kHz, and the corresponding frame length is set to 1024 samples (64 msec) with a frame shift of 256 samples. The STFT analysis is used to compute the DFT of each overlapping windowed frame. The diarization system is mainly based on neural speaker diarization using memory-aware multi-speaker embedding (NSD-MA-MSE). For the proposed SLT-IME system, the long-term block size is 7500, and the hop size is 3750. For beamforming, we stack all arrays into one big array according to [32]. The channel selection [33] and beamforming [34] are also adopted. For the back-end configurations, The speaker adaptive automatic speech recognition (SA-ASR) is just used as black-box for evaluating the performance. More details can refer to [24].

5.3. Results and analysis

Table 1 shows the Diarization results (false alarm (FA), missed (MISS), speaker error (SPKERR) and diarization error rate (DER)) on the DEV with the NSD-MA-MSE and the proposed SLT and SLT-IME based diarization refine strategies. Note that base on the speech analysis of the three datasets, the overall signal-to-noise ratio (SNR) of CHiME-6 and DiPCo is much lower than that of Mixer 6. And CHiME-6 has a more casual speaking style than other sets. First, the "NSD-MA-MSE", "SLT" and "SLT-IME" denote the diarization system [23], the proposed spatial long-term (SLT) and spatial long-term iterative mask estimation (SLT-IME) approach. For the SPKERR index in diarization, the proposed SLT can hugely reduce the SPKERR in development and evaluation sets, e.g., the SPKERR were significantly reduced from 1.92% to 1.45%

Table 1. Diarization results on the DEV with the baseline and the proposed SLT-based diarization refine.

Scenario	Method	DEV			
		FA	MISS	SPKERR	DER
CHiME-6	NSD-MA-MSE	1.64	25.44	2.85	29.93
	+SLT	13.32	16.32	2.58	32.22
	+SLT-IME	2.66	19.91	3.43	**26.00**
DiPCo	NSD-MA-MSE	1.80	14.50	1.92	18.22
	+SLT	16.65	6.64	1.45	24.74
	+SLT-IME	2.61	11.11	1.39	**15.12**
Mixer 6	NSD-MA-MSE	1.54	7.68	0.63	9.85
	+SLT	13.25	3.11	0.11	16.48
	+SLT-IME	1.48	7.57	0.36	**9.41**

Table 2. DA-WER on the DEV and EVAL with the baseline and the proposed SLT-based diarization refine.

Scenario	Baseline		NSD-MA-MSE		SLT-IME	
	DEV	EVAL	DEV	EVAL	DEV	EVAL
CHiME-6	62.4	77.4	33.6	35.8	32.8	33.2
DiPCo	56.6	54.7	35.1	33.7	32.0	31.6
Mixer 6	22.5	33.7	12.8	11.9	12.4	11.6
Macro	47.2	55.3	27.2	27.1	25.7	25.5

on development set of DiPCo, respectively. With better SPKERR initialization for NSD-MA-MSE, the SLT-IME provided huge diarization performance improvement on all three datasets, e.g., the SPKERR were significantly reduced from 29.93% to 26.00% on development set of ChiME-6.

Table 2 shows the Diarization Attributed Word Error Rate (DA-WER) on the DEV and EVAL with the official baseline, NSD-MA-MSE and the proposed SLT-IME. Comparing with NSD-MA-MSE, SLT-IME can further improve the performance of the recognition in terms of WER on both DEV and EVAL sets, e.g., the WER reduced from 35.8% to 33.2%, 33.7% to 31.6% and 11.9% to 11.6% on EVAL sets of CHiME-6, DiPCo and Mixer 6. By utilizing spatial information effectively, especially on overlapping regions, SLT-IME can serve as a better initialization system for NSD-MA-MSE.

Fig. 2 presents the oracle label and diarization label of 16s speech segments from selected utterance which belong to DiPCo dataset. Fig. 2 a) shows the spectrogram of speech segments from selected utterance which belong to DiPCo dataset. We can find that there are some overlap and background noise which is challenge for diarization. Fig. 2 b) shows the oracle labels. We can find that the official annotation labels two consecutive speech segments as one segment, causing the silent segment in the middle to also be labeled as a human voice, for example the first segment of speaker 4. Fig. 2 c) shows the diarization labels. The baseline diarization system can detect most speech segments, but it misses some segments belong to speaker 1 and speaker 2 due to the background noise and overlap. Fig. 2 d) and e) show the estimated masks and vad by SLT-IME method corresponding to speaker 1 and speaker 2. We can find that SLT-IME method is able to retrieve the speech segments missed by baseline diarization system due to its powerful spatial modeling capabilities. Fig. 2 g) shows the estimated mask and vad by SLT-IME method for speaker 4. The SLT-IME method can obtain more accurate VAD information for each speaker and the refined segments information is better for beamforming avoiding the

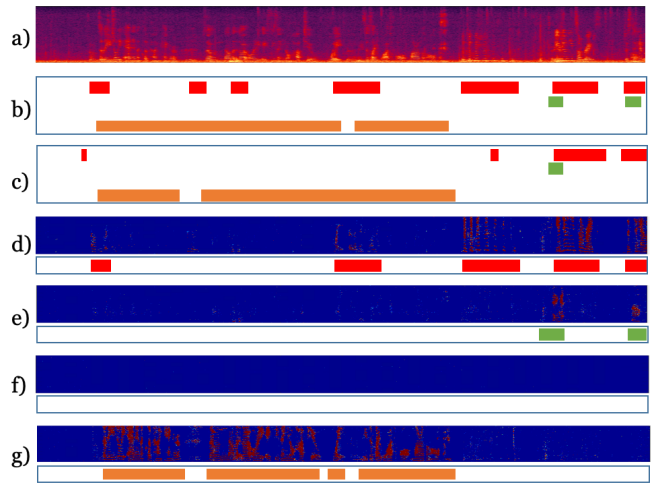

Fig. 2. An example of a long speech segment from the DiPCo development set: a) Spectrogram; b) Oracle speaker activity labels; c) The baseline diarization results; d) - g) The SLT-IME results for speaker $s \in \{1, 2, 3, 4\}$.

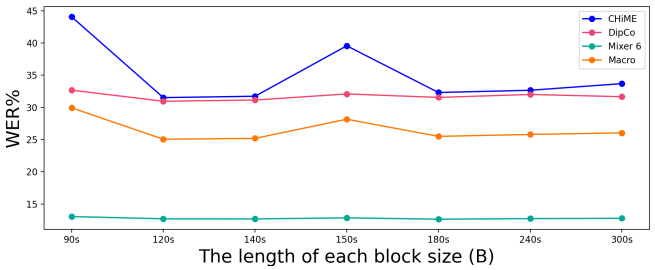

Fig. 3. Comparison of Word Error Ratio (WER%) between SLT-IME method under different lengths of long-term chunks configurations for multi-channel speech enhancement on the development set.

influence of noise segments. And this also explains the deterioration of the FA index in Table 1.

Fig. 3 shows the comparison of Word Error Ratio (WER%) between SLT-IME method under different lengths of long-term chunks configurations for multi-channel speech enhancement on the development set. We can find that the performance is worst when the length of the chunk is 90s. And SLT-IME gets the best performance when the length is 120s. As the length becomes longer, the performance gradually deteriorates.

6. CONCLUSION

In this paper, we propose a simple and effective method called SLT-IME to enhance the robustness of diarization systems iteratively, yielding a ASR performance improvement. By utilizing spatial long-term information, the SLT model can not only make full use of the space and speaker information but also distinguish different speakers from multi-channel noisy data. In the future, we can improve SLT-IME by leveraging upon better spatial beamforming approaches and more informative feedback from diarization system.

7. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China under Grant 62171427 and 62001446.

8. REFERENCES

[1] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: an overview," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 8599–8603.

[2] R. Leonard, "A database for speaker-independent digit recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 1984 IEEE International Conference on*, 1984, vol. 9, pp. 328–331.

[3] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett, "Getting started with the darpa timit cd-rom: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, vol. 107, pp. 16, 1988.

[4] D.B. Paul and J.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.

[5] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015, pp. 5206–5210.

[6] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, 2013.

[7] E. Vincent, J. Barker, S. Watanabe, Jonathan Le R., F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: datasets, tasks and baselines," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*. IEEE, 2013, pp. 126–130.

[8] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: dataset, task and baselines," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, 2015, pp. 504–511.

[9] B. Cornelis, M. Moonen, and J. Wouters, "Performance analysis of multichannel Wiener filter-based noise reduction in hearing aids under second order statistics estimation errors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1368–1381, 2011.

[10] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[11] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, 2011, pp. 189–192.

[12] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE transactions on speech and audio processing*, vol. 13, no. 1, pp. 120–134, 2005.

[13] L. Wang, H. Ding, and F. Yin, "A region-growing permutation alignment approach in frequency-domain blind source separation of speech mixtures," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 3, pp. 549–557, 2011.

[14] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[15] R. Talmon, I. Cohen, and S. Gannot, "Convolutive transfer function generalized sidelobe canceler," *IEEE transactions on audio, speech, and language processing*, vol. 17, no. 7, pp. 1420–1434, 2009.

[16] 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, 2010.

[17] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, "The third ¡®chime¡¯ speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Automat. Speech Recognition and Understanding Workshop.(ASRU)*, 2015.

[18] Yanhui Tu, Jun Du, Lei Sun, Feng Ma, Haikun Wang, Jingdong Chen, and Chinhui Lee, "An iterative mask estimation approach to deep learning based multi-channel speech recognition," *Speech Communication*, vol. 106, pp. 31–43, 2019.

[19] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: dataset, task and baselines," *arXiv preprint arXiv:1803.10609*, 2018.

[20] Shinji Watanabe, Michael Mandel, Jon Barker, and Emmanuel Vincent, "Overview of the 6th chime challenge," in *CHiME6 Workshop*, 2020.

[21] Samuele Cornell, Matthew Wiesner, Shinji Watanabe, Desh Raj, Xuankai Chang, Paola Garcia, Matthew Maciejewski, Yoshiki Masuyama, Zhong-Qiu Wang, Stefano Squartini, and Sanjeev Khudanpur, "The chime-7 dasr challenge: Distant meeting transcription with multiple devices in diverse scenarios," 2023.

[22] Martin Wolf and Climent Nadeu, "Channel selection measures for multi-microphone speech recognition," *Speech Communication*, vol. 57, pp. 170–180, 2014.

[23] Mao-Kui He, Jun Du, Qing-Feng Liu, and Chin-Hui Lee, "Ansd-mamse: Adaptive neural speaker diarization using memory-aware multi-speaker embedding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1561–1573, 2023.

[24] Ruoyu Wang, Maokui He, Jun Du, Hengshun Zhou, Shutong Niu, Hang Chen, Yanyan Yue, Gaobin Yang, Shilong Wu, Lei Sun, Yanhui Tu, Haitao Tang, Shuangqing Qian, Tian Gao, Mengzhi Wang, Genshun Wan, Jia Pan, Jianqing Gao, and Chin-Hui Lee, "The ustc-nercslip systems for the chime-7 dasr challenge," in *Interspeech 2023 Workshop*, 2023.

[25] R. J Mcaulay and T. F Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[26] C. Zhang, D. Florencio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Transactions on Signal Processing*, vol. 10, no. 3, pp. 538–548, 2008.

[27] B. D. Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE Signal Processing Magazine*, vol. 10, no. 3, pp. 4–24, 1988.

[28] Nobutaka Ito, Shoko Araki, and Tomohiro Nakatani, "Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *2016 24th European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1153–1157.

[29] Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal, "The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines," *INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association*, pp. 1561–1565, 2018.

[30] Maarten Van Segbroeck, Ahmed Zaid, Ksenia Kutsenko, Cirenia Huerta, Tinh Nguyen, Xuewen Luo, Björn Hoffmeister, Jan Trmal, Maurizio Omologo, and Roland Maas, "Dipco – dinner party corpus," 2019.

[31] Linda Brandschain, David Graff, Christopher Cieri, Kevin Walker, Chris Caruso, and A Neely, "The mixer 6 corpus: Resources for cross-channel and text independent speaker recognition," in *Proc. of LREC*, 2010.

[32] Naoyuki Kanda, Christoph Boeddeker, Jens Heitkaemper, Yusuke Fujita, Shota Horiguchi, and Reinhold Haebumbach, "Guided source separation meets a strong asr backend: Hitachi/paderborn university joint investigation for dinner party asr," *conference of the international speech communication association*, 2019.

[33] Kamil Wojcicki and Philipos C Loizou, "Channel selection in the modulation domain for improved speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 2904–2913, 2012.

[34] Takuya Higuchi, Nobutaka Ito, Takuya Yoshioka, and Tomohiro Nakatani, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.