Utterance-Wise Meeting Transcription System Using Asynchronous Distributed Microphones

Shota Horiguchi, Yusuke Fujita, Kenji Nagamatsu

Hitachi, Ltd.

{shota.horiguchi.wk, yusuke.fujita.su, kenji.nagamatsu.dm}@hitachi.com

Abstract

A novel framework for meeting transcription using asynchronous microphones is proposed in this paper. It consists of audio synchronization, speaker diarization, utterance-wise speech enhancement using guided source separation, automatic speech recognition, and duplication reduction. Doing speaker diarization before speech enhancement enables the system to deal with overlapped speech without considering sampling frequency mismatch between microphones. Evaluation on our real meeting datasets showed that our framework achieved a character error rate (CER) of $28.7\,\%$ by using 11 distributed microphones, while a monaural microphone placed on the center of the table had a CER of $38.2\,\%$. We also showed that our framework achieved CER of $21.8\,\%$, which is only 2.1 percentage points higher than the CER in headset microphone-based transcription.

Index Terms: meeting transcription, speech recognition, speaker diarization, asynchronous distributed microphones

1. Introduction

Meeting transcription is one practical use case of automatic speech recognition (ASR). Difficulties are i) that input audio signals suffer from reverberation and background noise because each utterance is recorded by distant microphones and ii) that they also suffer from speech overlap because each participant speaks at any time. To transcribe speech in such a wild condition, a powerful speech enhancement module is necessary. Most meeting transcription systems are therefore based on a microphone array [1, 2, 3, 4], sometimes one with an omnidirectional camera [5, 6] for face tracking. This means that the system requires special equipment to be introduced. If the microphone arrays can be replaced by more general devices, such as participants' smartphones or tablets, the usability of the system will be drastically improved. When such devices are distributed to transcribe a meeting, the problem is that they are asynchronous, and speech separation methods for synchronized signals cannot be simply applied.

Recently, some methods of meeting transcription using asynchronous distributed microphones have been proposed. One is the session-wise approach proposed by Araki *et al.* [7, 8]. They first synchronized multichannel observations by solving sampling frequency mismatch, then applied session-wise speech enhancement using the minimum variance distortionless response (MVDR) beamformer, then fed the enhanced signals into an ASR module to obtain the final transcription results. They showed that speech enhancement using asynchronous distributed microphones improved the ASR performance [7, 8]. The MVDR beamformer is a frequency-wise algorithm, however, so the well-known permutation problem of frequency-domain has to be solved. The common approach for multi-speaker cases is to prepare initial spatial correlation matrices from audio data with a fixed number of speakers and their positions [9]. Therefore, when the number of speakers in the inference audio is different from, especially larger than, that in the training set, we cannot provide initial spatial correlation matrices. If we cannot obtain such spatial correlation matrices beforehand, we have to solve the permutation problem as a post-processing [10, 11], but there are few reports that these methods perform well on real noisy and reverberant data.

Another is the block-wise approach proposed by Yoshioka *et al.* [12]. They synchronized input audio streams in a block-online manner and then applied block-wise speech separation. The separated audios are input into the ASR module, which is followed by speaker diarization. The benefit of this approach is that the effect of sampling frequency mismatch can be ignored within a block when the block is short enough because the scale of sampling frequency mismatch is about $100\,\mathrm{ppm}$ (parts per million) at most [13, 14]. However, their speech separation uses speech-vs-noise criteria and thus cannot deal with multiple speakers speaking simultaneously.

This paper investigates the utterance-wise approach, which is different from the session-wise or block-wise approaches described above. We first roughly synchronized audio signals recorded by distributed microphones and then applied speaker diarization. Speaker diarization is based on the clustering of features extracted from short segments, but we use features extracted from all the signals recorded by each microphone so that it can deal with overlapped speech. Then we applied guided source separation [15], which performed well for ASR in a dinner party scenario [16, 17]. This separation is conducted for each extracted utterance, which is short enough not to be suffered from sampling frequency mismatch between microphones. We applied ASR for each enhanced utterance, and finally, we conducted duplication reduction for ASR results to reduce the effect of errors on diarization or separation. Our approach can deal with speaker overlap without any methods to correct sampling frequency mismatch in the synchronization phase and solve the permutation problem in the speech enhancement phase. To evaluate our framework, we recorded eight sessions of real meetings using 11 distributed smartphones, each of which was equipped with a monaural microphone. The experimental results showed that our framework improved performance by using multiple microphones. We also showed that our framework could achieve performance comparable to that of headset microphone-based transcription if the oracle diarization results were known.

2. Method

We assume that a meeting is recorded by M asynchronous distributed microphones and transcription is based on the known number of speakers K in an offline manner. An overview of our method is shown in Figure 1. Given M audio signals, we

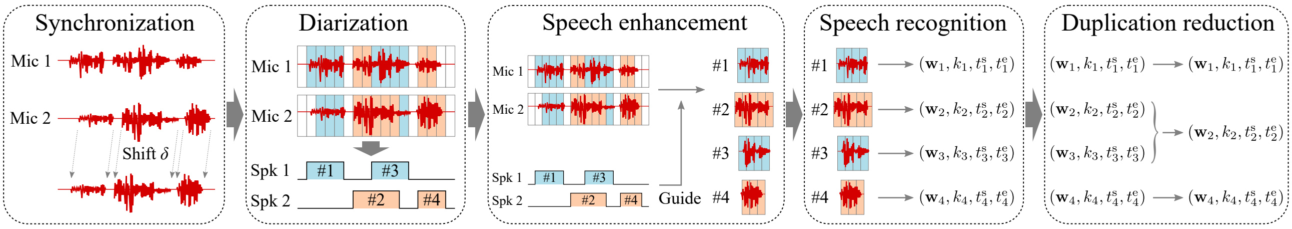
Synchronization
Mic 1
Mic 2
Shift δ
Diarization
Mic 1
Mic 2
Spk 1
#1
#3
Spk 2
#2
#4
Speech enhancement
Mic 1
Mic 2
Spk 1
#1
#3
Spk 2
#2
#4
Guide
#1
#2
#3
#4
Speech recognition
#1
#2
#3
#4
$(\mathbf{w}_1, k_1, t_1^s, t_1^e)$
$(\mathbf{w}_2, k_2, t_2^s, t_2^e)$
$(\mathbf{w}_3, k_3, t_3^s, t_3^e)$
$(\mathbf{w}_4, k_4, t_4^s, t_4^e)$
Duplication reduction
$(\mathbf{w}_1, k_1, t_1^s, t_1^e)$
$(\mathbf{w}_2, k_2, t_2^s, t_2^e)$
$(\mathbf{w}_3, k_3, t_3^s, t_3^e)$
$(\mathbf{w}_4, k_4, t_4^s, t_4^e)$
$(\mathbf{w}_1, k_1, t_1^s, t_1^e)$
$(\mathbf{w}_2, k_2, t_2^s, t_2^e)$
$(\mathbf{w}_4, k_4, t_4^s, t_4^e)$

Figure 1: *Overview of our meeting transcription system using asynchronous distributed microphones.*

first synchronize them by maximizing their correlation. The correction of sampling frequency mismatch between signals is not conducted in the synchronization part. With the synchronized signals, we conduct clustering-based diarization to obtain utterances for each speaker. After that, we perform speech enhancement for each utterance by using the diarization results as guides to avoid the permutation problem. The enhanced utterances are fed into the ASR module to obtain ASR results. Finally, to reduce errors caused by diarization or separation, we apply duplication reduction for the ASR results. In this section, we explain the details of each module of the system.

2.1. Blind synchronization

In this part, we conduct a correlation-based synchronization to correct start or end point differences of input signals. This rough synchronization can be operated under the existence of the sampling frequency mismatch. Assume that the observation of the m-th microphone ($m \in \{1, \dots, M\}$) is defined as $\hat{\mathbf{x}}_m := [\hat{x}_{m,n}]_{n=1}^{N_m}$. We select an anchor m_a from the M microphones and calculate the shift δ_m between signals of the anchor m_a and each microphone $m \in \{1, \dots, M\}$ as follows:

$$\delta_m = \begin{cases} \arg\max_{\delta \in \mathbb{Z}} \sum_\nu x_{m_a,\nu} x_{m,\nu+\delta} & (m \neq m_a) \\ 0 & (m = m_a), \end{cases} \tag{1}$$

$$x_{m,\nu} = \begin{cases} \hat{x}_{m,\nu} & (\nu \in \{1, \dots, N_m\}) \\ 0 & (\text{otherwise}). \end{cases} \tag{2}$$

Synchronized signals $\mathbf{x}_m$ ($m = 1, \dots, M$) are defined in the time interval recorded by all the microphones as follows:

$$\mathbf{x}_m = [\hat{x}_{m,n}]_{n=n_{\text{begin}}+\delta_m}^{n_{\text{end}}+\delta_m}, \tag{3}$$

$$n_{\text{begin}} = \max_{m' \in \{1,\dots,M\}} (1 - \delta_{m'}), \tag{4}$$

$$n_{\text{end}} = \min_{m' \in \{1,\dots,M\}} (N_{m'} - \delta_{m'}). \tag{5}$$

In this study we assume that all the utterances to be transcribed are within the time interval of $\mathbf{x}_m$.

2.2. Speaker diarization

In this paper, we conduct speaker diarization by clustering vectors. One drawback of the conventional clustering-based diarization using a monaural recording is that it cannot deal with speaker overlap because each timeslot is assigned to one speaker. On the other hand, in our scenario, each meeting has been recorded by distributed microphones. Therefore, even when two speakers spoke simultaneously, it is expected that one microphone could have captured the one speaker's utterance at a sufficient signal-to-noise ratio (SNR) and another microphone could have captured the other speaker's utterance at a sufficient SNR. In this study, we extract features from all the signals from all the microphones and conduct clustering for the extracted features all together to deal with speaker overlap.

We first split the synchronized observations $\{\mathbf{x}_m\}_m$ into short segments $\{\mathbf{x}_{m,t}\}_{m,t}$ with 1.5 s of window size and 0.75 s of window shift, where $t = 1, \dots, T$ denotes the timeslot index. We apply power-based speech activity detection for each segment; as a result, each segment is classified as either speech or non-speech. From each speech segment, we extract features to be used for clustering. In this study, we concatenate two kinds of features: speaker characteristics based features and power ratio based features.

For features to represent speaker characteristics, we use x-vectors [18], which are used in the state-of-the-art diarization systems [19, 20]. We extract x-vectors from the audio of each microphone so that we can obtain different speaker characteristics from the same timeslot; thus we can deal with speaker overlap. Before we use the vectors for clustering, we subtract a mean vector within a session from each x-vector and normalized it to have unit norm. As a result, we obtain microphone and timeslot-wise D-dimensional features $\mathbf{c}_{m,t} \in \mathbb{R}^D$.

Although x-vectors from distributed microphones are potentially beneficial to diarize overlapped speech, it becomes a problem that an utterance from the same speaker could be judged as one from multiple speakers because x-vectors suffer from speaker-microphone distance and noisy environments. Thus, we introduce power-based timeslot-wise features $\mathbf{p}_t := [p_{1,t}, \dots, p_{M,t}]^\mathsf{T}$, where $p_{m,t}$ is the average power at $\mathbf{x}_{m,t}$. This speaker diarization part is a session-level one, so we avoid using phase-based features like GCC-PHAT [21] because they suffer from the sampling frequency mismatch.

Final $(D + M)$-dimensional features to be clustered are

$$\mathbf{v}_{m,t} = \begin{bmatrix} \mathbf{c}_{m,t} \\ \lambda \mathbf{p}_t / \|\mathbf{p}_t\| \end{bmatrix}, \tag{6}$$

where λ is the scaling factor to balance the effect of $\mathbf{c}_{m,t}$ and $\mathbf{p}_t$. We apply agglomerative hierarchical clustering for the features to divide the speech segments into K clusters. As a result, each feature from a speech segment belongs to one of the clusters $\mathcal{C}_1, \dots, \mathcal{C}_K$, where $\mathcal{C}_k$ corresponds to the speech cluster of k-th speaker. We also define the additional noise cluster $\mathcal{C}_{K+1} := \{\mathbf{v}_{m,t}\}_{m,t}$. The diarization results including noise $Y = \{y_t^{(k)}\} \in \{0, 1\}^{(K+1)\times T}$ are calculated as

$$y_t^{(k)} = \begin{cases} 1 & (\exists m \in \{1, \dots, M\},\ \mathbf{v}_{m,t} \in \mathcal{C}_k) \\ 0 & (\text{otherwise}). \end{cases} \tag{7}$$

In the diarization results, utterances are sometimes divided into some short fragments due to the existence of backchannels, noises, *etc*. In this study, we treat silence of 1.5 s or less between speech fragments from the same speaker as a speech by applying two iterations of binary closing along the time axis.

Here each timeslot in the diarization results corresponds to 0.75 s, which is inconsistent with the signals used in speech enhancement in the next section. Thus, we upsample the diarization results so that each timeslot corresponds to 16 ms. Hereafter, $Y = \{y_t^{(k)}\}$ denotes the upsampled diarization results.

2.3. Speech enhancement

In this study, we conducted speech enhancement for each utterance by using guided source separation (GSS) [15]. While the original GSS utilized oracle speech activities, we instead use estimated diarization results described in the previous section.

We first apply Weighted Prediction Error [22] to the input multichannel signals in a short-time Fourier transform (STFT) domain for dereverberation. The frame length and the frame shift for the STFT were set to 64 ms and 16 ms, respectively. After that, speech separation by GSS [15] using a complex Angular Central Gaussian Mixture Model (cACGMM) [23] is applied. Given M-channel observations in the STFT domain $\mathbf{X}_{t,f} \in \mathbb{C}^M$, the probability density function of the cACGMM for the signals is defined as

$$p\left(\hat{\mathbf{X}}_{t,f}; \{\alpha_f^{(k)}, B_f^{(k)}\}_k\right) = \sum_k \alpha_f^{(k)} \mathcal{A}\left(\hat{\mathbf{X}}_{t,f}; B_f^{(k)}\right), \quad (8)$$

where $\hat{\mathbf{X}}_{t,f} = \mathbf{X}_{t,f} / \|\mathbf{X}_{t,f}\|$ and $\alpha_f^{(k)}$ is the mixture weight for the k-th source of the frequency bin f. $\mathcal{A}(\hat{\mathbf{X}}; B)$ is a complex Angular Central Gaussian distribution [24] parameterized by $B \in \mathbb{C}^{M \times M}$. The cACGMM is optimized by the EM algorithm. At the E-step we calculate posteriors $\gamma_{t,f}^{(k)}$ for each speaker at time-frequency bin as follows:

$$\gamma_{t,f}^{(k)} \leftarrow \frac{\alpha_f^{(k)} y_t^{(k)} \frac{1}{\det\left(B_f^{(k)}\right)} \frac{1}{\left[\hat{\mathbf{X}}_{t,f}^{\mathsf{H}} \left(B_f^{(k)}\right)^{-1} \hat{\mathbf{X}}_{t,f}\right]^M}}{\sum_{k'} \alpha_f^{(k')} y_t^{(k')} \frac{1}{\det\left(B_f^{(k')}\right)} \frac{1}{\left[\hat{\mathbf{X}}_{t,f}^{\mathsf{H}} \left(B_f^{(k')}\right)^{-1} \hat{\mathbf{X}}_{t,f}\right]^M}}. \quad (9)$$

Here the diarization result $y_t^{(k)}$ works as a guide at this E-step to force the posterior probability to be zero when the speaker k does not speak at time t. At the M-step the parameters $\alpha_f^{(k)}$ and $B_f^{(k)}$ are updated as follows:

$$\alpha_f^{(k)} \leftarrow \frac{1}{T} \sum_t \gamma_{t,f}^{(k)}, \quad B_f^{(k)} \leftarrow M \frac{\sum_t \gamma_{t,f}^{(k)} \frac{\hat{\mathbf{X}}_{t,f} \hat{\mathbf{X}}_{t,f}^{\mathsf{H}}}{\hat{\mathbf{X}}_{t,f}^{\mathsf{H}} \left(B_f^{(k)}\right)^{-1} \hat{\mathbf{X}}_{t,f}}}{\sum_t \gamma_{t,f}^{(k)}}, \quad (10)$$

where $(\cdot)^{\mathsf{H}}$ denotes Hermitian transpose. To solve the permutation problem, 15 s of audios before and after each segment are used as "context". After 10 iterations of optimization, we calculate the spatial covariance matrices for speech and noise as follows:

$$R_f^{\text{speech}} = \frac{1}{T} \sum_t \gamma_{t,f}^{(k_{\text{target}})} \mathbf{X}_{t,f} \mathbf{X}_{t,f}^{\mathsf{H}} \in \mathbb{C}^{M \times M}, \quad (11)$$

$$R_f^{\text{noise}} = \frac{1}{T} \sum_t \left(1 - \gamma_{t,f}^{(k_{\text{target}})}\right) \mathbf{X}_{t,f} \mathbf{X}_{t,f}^{\mathsf{H}} \in \mathbb{C}^{M \times M}. \quad (12)$$

Here we assume that the target speaker is $k_{\text{target}} \in \{1, \ldots, K\}$. The MVDR beamformer $\mathbf{w}_f$ is calculated using the spatial covariance matrices as

$$\mathbf{w}_f = \frac{{R_f^{\text{noise}}}^{-1} R_f^{\text{speech}} \mathbf{r}}{\operatorname{tr}\left({R_f^{\text{noise}}}^{-1} R_f^{\text{speech}}\right)}, \quad (13)$$

where $\mathbf{r}$ is an one-hot vector that corresponds to the reference microphone. Finally, Blind Analytic Normalization (BAN) postfilter [25] is applied for $\mathbf{w}_f$ to obtain the final beamformer, which is used for speech enhancement. The enhanced utterance in the STFT domain is calculated as

$$z_{t,f} = \mathbf{w}_f^{\mathsf{H}} \mathbf{X}_{t,f}. \quad (14)$$

2.4. Speech recognition

For each enhanced utterance, we apply ASR consisting of a CNN-TDNN-LSTM acoustic model (AM) [26] followed by 4-gram-based and recurrent neural network-based language models (LMs) [27]. The AM takes 40-dimensional log-scaled Mel-filterbank and 40-dimensional Mel-frequency cepstral coefficients as input audio features. 100-dimensional i-vectors are also fed into the AM for online adaptation for speaker and environment [28]. It was trained by 1700 hours of Japanese speech corpus using the lattice-free maximum mutual information criterion [29]. The LMs were trained by transcriptions of the corpus used for AM training and the Wikipedia corpus.

2.5. Duplication reduction

The diarization and speech enhancement is not perfect, so the same transcription is sometimes included in multiple estimated utterances. Therefore, we apply duplication reduction for the ASR results. Widely used ensemble techniques such as ROVER [30] and confusion network combination [31] are for the different ASR results obtained from the same utterance; thus, they cannot be used in this situation where the utterances to be merged have different start and end points. To overcome this issue, we propose a combination technique for such utterances which have different time intervals. We first find which pairs of utterances should be merged. Given the set of U ASR results $\mathcal{W} = \{(\mathbf{w}_u, k_u, t_u^{\text{s}}, t_u^{\text{e}})\}_{u=1}^U$, where $\mathbf{w}_u$, k_u, t_u^{s}, and t_u^{e} denote the sequence of words, speaker, start time, and end time of u-th result, respectively, we calculate an adjacency matrix $A = \{a_{i,j}\}_{i,j} \in \{0,1\}^{U \times U}$ as follows:

$$a_{i,j} = \begin{cases} 1 & (\max(t_i^{\text{s}}, t_j^{\text{s}}) < \min(t_i^{\text{e}}, t_j^{\text{e}}) \wedge \\ & \quad s(\mathbf{w}_i, \mathbf{w}_j) > \tau \wedge k_i \neq k_j) \\ 0 & (\text{otherwise}), \end{cases} \quad (15)$$

where $\tau \in [0,1]$ is the threshold value. Here $s(\mathbf{w}_i, \mathbf{w}_j)$ is the similarity between $\mathbf{w}_i$ and $\mathbf{w}_j$ defined as follows:

$$s(\mathbf{w}_i, \mathbf{w}_j) := \frac{\max(|\mathbf{w}_i|, |\mathbf{w}_j|) - d(\mathbf{w}_i, \mathbf{w}_j)}{\min(|\mathbf{w}_i|, |\mathbf{w}_j|)}, \quad (16)$$

where $d(\mathbf{w}_i, \mathbf{w}_j)$ is the Levenshtein distance between $\mathbf{w}_i$ and $\mathbf{w}_j$, and $|\mathbf{w}|$ denotes the number of words in $\mathbf{w}$. With this adjacency matrix, all the elements in $\mathcal{W}$ can be clustered into C clusters. We denote the clustering result as $\mathcal{C} = \{c_u\}_{u=1}^U \in \{1, \ldots, C\}^U$, which fulfill $c_i = c_j$ if a path between i-th and j-th elements exists in A and $c_i \neq c_j$ otherwise. Assuming that $\mathcal{W}_{k,c} \subseteq \mathcal{W}$ is the set of ASR results which belong to the cluster c and are uttered by speaker k, we obtain the representative speaker k^c of the cluster c by

$$k^c = \mathop{\arg\max}_{k \in \{1, \ldots, K\}} f(\mathcal{W}_{k,c}), \quad (17)$$

where $f(\cdot)$ is the selection function. In this study, we select the speaker with the longest utterance(s), *i.e.*, $f(\mathcal{W}_{k,c}) = \sum_{(\mathbf{w},k,t^{\text{s}},t^{\text{e}}) \in \mathcal{W}_{k,c}} |\mathbf{w}|$. The set of de-duplicated ASR results $\mathcal{W}'$ can be obtained as follows:

$$\mathcal{W}' = \bigcup_{c \in \{1, \ldots, C\}} \mathcal{W}_{k^c, c}. \quad (18)$$

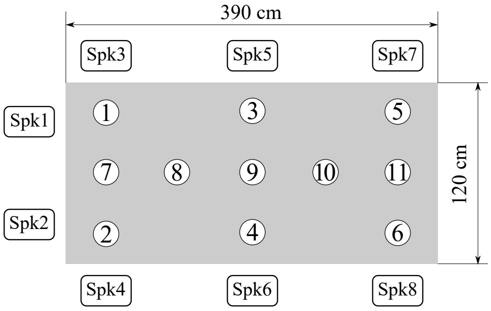

Figure 2: *Recording environment.* ①-⑪ *denote smartphones, each of which is equipped with a monaural microphone.*

Table 1: *Statistics of the recorded meetings.*

Session ID	#Speakers	Duration	#Utterances	Overlap ratio (%)
I	7	19:49	160	6.9
II	8	14:27	150	14.0
III	5	13:13	198	16.6
IV	7	12:08	184	19.9
V	6	12:37	80	5.5
VI	6	16:50	256	14.7
VII	7	16:25	223	11.3
VIII	7	11:25	185	19.9
Avg.	-	116:54	1436	13.2

3. Experiments

3.1. Data

To evaluate the performance of our method, we collected eight sessions of real meeting data. The recording environment is shown in Figure 2. Each session had at most eight participants and was recorded by 11 smartphones distributed on the table. Each smartphone was equipped with a monaural microphone to record meetings at 16 kHz / 16 bit. Each participant wore a headset microphone, and the groundtruth transcriptions were based on the headset recordings. The statistics of collected data are shown in Table 1. The recordings correspond to about two hours of meetings with an average overlap ratio of 13.2 %.

3.2. Results

We investigated various combinations of asynchronous distributed microphones: 2 microphones (⑧&⑩ in Figure 2), 3 microphones (⑦&⑨&⑪), 6 microphones (①-⑥), and 11 microphones (①-⑪). For comparison, we also evaluated the performance of one monaural microphone (⑨) and of headset microphones that the participants wore during each session.

The character error rates (CERs) obtained using various microphone combinations in each session are shown in Table 2. In these experiments, the weighting parameter λ in Equation 6 was set to 1.0. By using multiple microphones, we could have reduced CERs, especially by using a large number of microphones. Note that in two-, three-, and six-microphone settings, using more microphones not always resulted in better CERs. This is because the sets of microphones in these settings are disjoint and the CERs highly depended on the positions of microphones and speakers. On the other hand, we observed the best CERs in almost every session by using all the 11 microphones. This result indicated that adding microphones has almost no negative effect on CERs. In Table 2, we also showed CERs with 11 microphones in the case when oracle diarization was used for GSS. It achieved the CER of 21.8 %, which is only 2.1 percentage points worse than the CER of 19.7 % obtained using headset microphones. It can be said that our method can potentially achieve nearly headset-level CERs when it is used with a more powerful diarization method [32, 33, 34].

Table 2: *CERs (%) obtained using various microphone combinations.*

#Mic	Session ID								Avg.
	I	II	III	IV	V	VI	VII	VIII	
1	31.2	30.1	37.1	37.6	28.2	48.4	50.4	52.5	38.2
2	22.9	25.3	30.5	37.0	21.8	41.7	36.8	45.7	31.4
3	26.8	24.4	35.9	37.2	23.2	43.1	41.9	46.6	33.7
6	22.3	22.2	36.0	32.1	21.0	38.1	35.1	44.3	30.2
11	21.2	21.1	32.5	30.9	19.6	37.6	34.0	41.0	28.7
11*	17.0	16.3	21.7	21.2	17.7	27.0	27.0	32.8	21.8
Headset	18.3	15.8	21.0	20.1	13.6	21.3	24.9	25.8	19.7

* The oracle diarization was used for speech enhancement.

Table 3: *CERs (%) obtained with various scaling factors λ in Equation 6.*

#Mic	λ						
	2^{-3}	2^{-2}	2^{-1}	2^{0}	2^{1}	2^{2}	2^{3}
2	33.7	31.7	32.2	**31.4**	31.8	33.3	35.5
3	34.2	34.3	34.0	33.7	**33.0**	34.8	35.2
6	33.5	34.1	33.4	**30.2**	31.4	31.9	32.4
11	33.5	32.5	31.1	28.7	28.9	**28.4**	28.9

Table 4: *Ablation study using 11 microphones.*

Method	CER (%)
Baseline (11 mics)	28.7
w/o binary closing	30.6
w/o speech enhancement	37.8
w/o duplication reduction	31.9

In Table 3 we show the average CERs over sessions with various weighting parameters λ in Equation 6. Combinations of speaker characteristics based features and power ratio based features improved transcription performance, especially when the number of microphones is smaller and the power ratio thus has less information about the directions of speakers.

Finally, we conducted ablation studies by removing binary closing in diarization, speech enhancement by using recordings of the reference microphone instead, and duplication reduction, respectively. Here we used 11 microphones with $\lambda = 1.0$. The results are shown in Table 4. We found 1.9, 9.1, and 3.2 percentage points degradation from the baseline by removing binary closing, speech enhancement, and duplication reduction, respectively. From there results, we concluded that these three components contributed to the improvement of the CER.

4. Conclusions

In this paper, we proposed a meeting transcription system based on utterance-wise processing using asynchronous distributed microphones. It consists of the following modules: blind synchronization, speaker diarization, speech enhancement, speech recognition, and duplication reduction. Evaluation on the real meeting data showed the effectiveness of our framework and its components, and also showed that it could perform comparably to the headset microphone-based transcription if the oracle diarization was given. The future perspective of this research is to operate this framework in an online manner.

5. References

[1] A. Stolcke, X. Anguera, K. Boakye, Ö. Çetin, A. Janin, M. Magimai-Doss, C. Wooters, and J. Zheng, "The SRI-ICSI spring 2007 meeting and lecture recognition system," in *Multimodal Technologies for Perception of Humans*. Springer, 2007, pp. 450–463.

[2] T. Hain, L. Burget, J. Dines, P. N. Garner, F. Grézl, A. El Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, "Transcribing meetings with the AMIDA systems," *IEEE TASLP*, vol. 20, no. 2, pp. 486–498, 2011.

[3] N. Ito, S. Araki, M. Delcroix, and T. Nakatani, "Probabilistic spatial dictionary based online adaptive beamforming for meeting recognition in noisy and reverberant environments," in *ICASSP*, 2017, pp. 681–685.

[4] T. Yoshioka, H. Erdogan, Z. Chen, X. Xiao, and F. Alleva, "Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks," in *INTERSPEECH*, 2018, pp. 3038–3042.

[5] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura, and Y. Junji, "Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera," *IEEE TASLP*, vol. 20, no. 2, pp. 499–513, 2011.

[6] T. Yoshioka, I. Abramovski, C. Aksoylar, Z. Chen, M. David, D. Dimitriadis, Y. Gong, I. Gurvich, X. Huang, Y. Huang, A. Hurvitz, L. Jiang, S. Koubi, E. Krupka, I. Leichter, C. Liu, P. Parthasarathy, A. Vinnikov, L. Wu, X. Xiao, W. Xiong, H. Wang, Z. Wang, J. Zhang, Y. Zhao, and T. Zhou, "Advances in online audio-visual meeting transcription," in *ASRU*, 2019, pp. 276–283.

[7] S. Araki, N. Ono, K. Konoshita, and M. Delcroix, "Meeting recognition with asynchronous distributed microphone array," in *ASRU*, 2017, pp. 32–39.

[8] S. Araki, N. Ono, K. Kinoshita, and M. Delcroix, "Meeting recognition with asynchronous distributed microphone array using block-wise refinement of mask-based MVDR beamformer," in *ICASSP*, 2018, pp. 5694–5698.

[9] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online MVDR beamformer based on complex gaussian mixture model with spatial prior for noise robust ASR," *IEEE/ACM TASLP*, vol. 25, no. 4, pp. 780–793, 2017.

[10] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE TASLP*, vol. 12, no. 5, pp. 530–538, 2004.

[11] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE TASLP*, vol. 19, no. 3, pp. 516–527, 2010.

[12] T. Yoshioka, D. Dimitriadis, A. Stolcke, W. Hinthorn, Z. Chen, M. Zeng, and X. Huang, "Meeting transcription using asynchronous distant microphones," in *INTERSPEECH*, 2019, pp. 2968–2972.

[13] S. Miyabe, N. Ono, and S. Makino, "Blind compensation of interchannel sampling frequency mismatch for ad-hoc microphone array based on maximum likelihood estimation," *Signal Processing*, vol. 107, pp. 185–196, 2015.

[14] S. Araki, N. Ono, K. Kinoshita, and M. Delcroix, "Estimation of sampling frequency mismatch between distributed asynchronous microphones under existence of source movements with stationary time periods detection," in *ICASSP*, 2019, pp. 785–789.

[15] C. Boeddeker, J. Heitkaemper, J. Schmalenstoeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in *CHiME-5*, 2018.

[16] N. Kanda, C. Boeddeker, J. Heitkaemper, Y. Fujita, S. Horiguchi, K. Nagamatsu, and R. Haeb-Umbach, "Guided source separation meets a strong ASR backend: Hitachi/Paderborn University joint investigation for dinner party scenario," in *INTERSPEECH*, 2019, pp. 1248–1252.

[17] C. Zorilă, C. Boeddeker, R. Doddipatla, and R. Haeb-Umbach, "An investigation into the effectiveness of enhancement in ASR training and test for CHiME-5 dinner party transcription," in *ASRU*, 2019, pp. 47–53.

[18] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *ICASSP*, 2018, pp. 5329–5333.

[19] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, and S. Khudanpur, "Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge," in *INTERSPEECH*, 2018, pp. 2808–2812.

[20] M. Diez, L. Burget, S. Wang, J. Rohdin, and J. Černockỳ, "Bayesian HMM based x-vector clustering for speaker diarization," in *INTERSPEECH*, 2019, pp. 346–350.

[21] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE TASSP*, vol. 24, no. 4, pp. 320–327, 1976.

[22] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE TASLP*, vol. 18, no. 7, pp. 1717–1731, 2010.

[23] N. Ito, S. Araki, and T. Nakatani, "Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *EUSIPCO*, 2016, pp. 1153–1157.

[24] J. T. Kent, "Data analysis for shapes and images," *Journal of statistical planning and inference*, vol. 57, no. 2, pp. 181–193, 1997.

[25] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE TASLP*, vol. 15, no. 5, pp. 1529–1539, 2007.

[26] N. Kanda, Y. Fujita, and K. Nagamatsu, "Lattice-free state-level minimum Bayes risk training of acoustic models," in *INTERSPEECH*, 2018, pp. 2923–2927.

[27] ——, "Investigation of lattice-free maximum mutual information-based acoustic models with sequence-level Kullback-Leibler divergence," in *ASRU*, 2017, pp. 69–76.

[28] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *ASRU*, 2013, pp. 55–59.

[29] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *INTERSPEECH*, 2016, pp. 2751–2755.

[30] J. G. Fiscus, "A post-processing system to yield reduced work error rates: Recognizer output voting error reduction (ROVER)," in *ASRU*, 1997, pp. 347–352.

[31] G. Evermann and P. C. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *NIST Speech Transcription Workshop*, vol. 27, 2000, pp. 78–81.

[32] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," in *ASRU*, 2019, pp. 296–303.

[33] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, and A. Romanenko, "The STC system for the CHiME-6 Challenge," in *CHiME-6*, 2020.

[34] S. Horiguchi, Y. Fujita, S. Wananabe, Y. Xue, and K. Nagamatsu, "End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors," in *INTERSPEECH*, 2020.