CHiME 2018 Workshop on Speech Processing in Everyday Environments
07 September 2018, Hyderabad, India

Channel Selection using Neural Network Posterior Probability for Speech Recognition with Distributed Microphone Arrays in Everyday Environments

Feifei Xiong[1], *Jisi Zhang*[1], *Bernd T. Meyer*[2], *Heidi Christensen*[1], *Jon Barker*[1]

[1]Speech and Hearing Group (SPandH), University of Sheffield, Sheffield, UK
[2]Medical Physics and Cluster of Excellence Hearing4All, University of Oldenburg, Germany

Abstract

This paper presents an automatic speech recognition (ASR) system for the 5th CHiME Speech Separation and Recognition Challenge for transcribing continuous conversations recorded in everyday environments with distributed microphone arrays. The main contribution of the proposed system is the investigation of an effective real-time channel selection scheme to pick up reliable microphones/array for target speakers. It is shown that the proposed channel selection method produces better ASR performance than the reference channel provided by the baseline system, as well as comparable results to a reference-required selection approach based on speech intelligibility test as the oracle case. Instead of including all available data for training data augmentation, channel selection can be also applied to the training data selection to minimize the training-test-mismatch. Further, complementary knowledge can be obtained when the best two channels are selected specially in periods of overlapping speech. The final ASR system, which additionally incorporates improved acoustic modeling and system combination, achieves absolute word error rate reductions of 9.2% and 7.0% with development and evaluation test set, respectively, compared to the baseline in the context of multiple-array track.

1. Introduction

Systems that capture multiple audio streams using distributed microphones are becoming increasingly common [1], e.g., different microphone arrays in mobile or TV devices, distributed microphone arrays inside smart homes, etc. While distributed microphones enable technologies like beamforming to exploit spatial knowledge for speech enhancement, adding more channels is not *guaranteed* to improve automatic speech recognition (ASR) system performance. For instance, recent studies with distributed microphones (cf. [2]) show that ASR performance in reverberant environments may even degrade when some single streams that have high reverberation and low signal-to-noise ratio (SNR) are included. Further, distributed microphones are required to be synchronized for proper usage by microphone array or blind source separation techniques, and the exact synchronization is itself quite challenging to achieve [3]. Therefore, strategies for the reliable microphones/array selection are of great interest.

Several channel selection (CS) approaches for application to ASR with distributed microphones have been proposed recently on the basis of the common underlying principle: The best channel should lead to the best performance, i.e., lowest word error rate (WER) for ASR. However, since WER is unknown during recognition in practice, an alternative measure to select proper channels, therefore, needs to be as correlated as possible with the ASR performance. Such a measure can be achieved at different stages of the ASR system. In the signal domain, signal-to-noise ratio (SNR) estimation is possibly the most widely used as the indication of signal quality for CS [4], which however, requires accurate voice activity detection that often fails in reverberant or non-stationary noisy environments. To consider the reverberant conditions with distant microphones, [5] introduced an envelope-variance measure based on the observation that reverberation smooths the time sequence of speech energy values, so that the variance of the compressed filter bank energies could reflect the degree of distortion (by reverberation), i.e., the highest energy for all subbands as the least distorted channel. Further, these *signal-based* approaches are specially meaningful for CS with distributed microphones to boost the computational efficiency of beamforming [6]. However, as they do not consider any knowledge of the ASR system, intuitively, one does not expect a high correlation with WER.

When incorporating information of the ASR decoding process (*decoder-based* measure), one straightforward method has been proposed in [7], where the channel with the highest acoustic likelihood is considered as the best. However, due to the non-normalized probability of the observation vector, likelihood, by itself, is not a good indicator of the signal quality if signals are coming from different channels. Alternatively, [8] presented a feature normalization method that compares the ASR hypothesis of original and normalized feature vectors for each channel, provided that normalization could compensate the distortion caused by adverse acoustic conditions and the channel with the smallest difference between the recognized word sequences from the original and the compensated version is supposed to be the best. Further, [9] introduced another CS method based on a class separability measure, attempting to search for the channel where the class separability measure is maximized. However, the choice of the class units, e.g., speech features (signal-based), phonemes or tri-phones (decoder-based) is not trivial. Moreover, CS has also been achieved from the perspective of confidence measures [10] or ASR quality estimation [11]. These decoder-based measures use some information from the ASR decoder, which, in principle, should be more correlated with the WER than the signal-based ones. On the other hand, an evident drawback of decoder-based measures is their significant computational complexity.

Recently with the predominant use of deep neural networks (DNNs) in ASR back-end systems (cf. [12]), the analysis of DNN posterior probability provides potential for new measures that correlate with ASR performance. Based on the observation that with a good match to the DNN model, the distribution of phone posteriors will typically be dominated by clear phone classes, the entropy of DNN phone posterior distributions can reflect the degree of mismatch between the test and the training data, resulting in a strong correlation with the final WER. This has been adopted for acoustic confidence mea-

10.21437/CHiME.2018-5

sure [13] and for combination strategies in multi-stream ASR framework [14]. A statistical analysis of phoneme posteriors between the training and the test data has been conducted in [15], where large divergence between these two statistics indicates possible degradation of the classifier performance. A mean temporal distance measure of phoneme posteriors was further proposed in [16], based on the intuition that distant clean posterior vectors will be rather different (since they are likely to belong to different phoneme classes), while this difference should be smaller for noisy vectors. Alternatively, autoencoders have been employed in [17] to learn characteristics of the training data, and the reconstruction error obtained with test data was used to monitor the performance of each channel. These *posterior-based* approaches show the advantage of low computational complexity over *decoder-based* measures, meanwhile, the correlation with the ASR performance increases compared to *signal-based* ones, which also motivates this paper to apply DNN posterior probability analysis to achieve CS in the context of the 5th CHiME Challenge [18].

More specifically, the CHiME-5 Challenge considers the problem of conversational speech recognition in a dinner party scenario with four talkers and with six Microsoft Kinect devices (each with a linear array of four microphones) distributed in various positions in the kitchen, living and dinning room. Although all 24 microphones could be employed, synchronization between arrays itself is a challenging problem, meaning that it is more convenient to treat each device independently. In effect, recordings from Kinects which were not located in the room where the speakers were talking exhibit high reverberation and low SNRs, leading to a significant recognition degradation. It is therefore of importance to dynamically select the proper Kinect/channel for each active speaker to achieve the optimal recognition result.

To this end, we exploit entropy analysis [13, 14] of the DNN posterior probability for CS, which can be processed in a frame-wise mode resulting in potentially real-time applications. Usually a good reference DNN model is necessary for a distinguishable rank among channels: On one hand, this model should reflect a strong correlation to the final ASR performance — the best one will be chosen as the input for the ASR recognizer. On the other hand, discrimination in terms of DNN posterior probabilities among channels should be as noticeable as possible to avoid ambiguous selection, which could occur when applying a DNN model with multi-condition training that partially plays a role of equalization across channels with different conditions. Motivated by findings from pilot experiments with CHiME-5 data that recorded speech signals from binaural microphones worn by the party participants always yield the best recognition results and that these recordings can be considered as homogeneous (i.e., under similar acoustic condition), we therefore train the reference DNN model for CS only using the binaural speech data from Challenge training session. In fact, using these signals as a reference, CS can be achieved by a straightforward signal-based comparison, e.g., via a speech intelligibility test [19] where the highest scoring channel is presumed likely to yield the best recognition score [20]. However, when testing in real scenarios, it is not practical to obtain such reference signals. Furthermore, in order to mitigate the problem of overlapping speech which is not covered by the CS DNN model (without speaker identification), the scheme will select the best two channels/devices, motivated by the fact that in the CHiME recordings there are usually two Kinects located close to the active speakers, and their spatial diversity can provide complementary evidence, particularly during speaker overlap.

In the reminder of this paper, we first briefly introduce the system chain for CHiME-5 Challenge in Section 2 aiming at reducing the final WER with the technologies ranging from channel selection, over improved acoustic modeling to system combination. The proposed channel selection approach is then described in detail in Section 3. Overall experimental evaluation with final submission results to CHiME-5 Challenge is presented in Section 4, and Section 5 concludes the paper.

2. System Overview

As illustrated in Figure 1, the improvements compared to the baseline conventional ASR [18] stem from the proposed channel selection approach, improved acoustic model (AM), and combination strategy with complementary input. Note that baseline BeamformIt [21] is applied as the speech enhancement (SE) algorithm since other microphone-array methods we tested (e.g., blind source separation [22]) could not achieve further WER reduction.

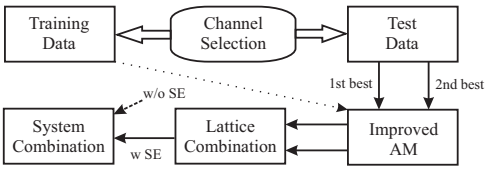

Figure 1: *The proposed system chain for CHiME-5 Challenge.*

2.1. Improved Acoustic Modeling

Besides the training data selection/augmentation (cf. Section 3) and SE, acoustic model for conventional ASR systems can be further improved in terms of robust feature extraction and advanced DNN architecture. Due to the conversational scenarios with overlap speaker, it would be beneficial for ASR to integrate speaker-related knowledge into the feature extraction, which is also verified by the performance improvement when 100-dimensional i-vectors are appended to the conventional 40-dimensional mel-frequency cepstral coefficients (MFCCs) in the baseline CHiME-5 Challenge system, as seen in Table 1. In order to further capture speaker characteristics in the feature domain, additional 3-dimensional pitch features [23] are employed including voicing probability, log-pitch and delta-pitch values, which provide a small but consistent WER reduction with the development test set (Dev).

Table 1: *Performance with robust feature extraction including i-vectors and pitch features appended to 40-dimensional MFCCs in the context of baseline settings for Dev.*

i-vector (#100)	pitch (#3)	WER
✗	✗	84.09
✓	✗	80.62
✓	✓	80.36

Further, motivated by the finding that it is beneficial to include long temporal contexts for DNN-HMM based AMs, i.e., to full exploit long-range correlations in the speech signals, advances in recurrent neural networks (RNNs) could be integrated into the baseline neural network architecture, which applied a chain model with time-delayed neural network (TDNN) [24]. Long short-term memory (LSTM) projected RNN [25] is employed to further capture the temporal dynamics

from the input frames, resulting in TDNN hidden layers as $\{0;(-1,0,1);0;\mathrm{LSTM};(-1,0,1);0;(-3,0,3);\mathrm{LSTM};(-3,0,3);(-6,-3,0);\mathrm{LSTM}\}$, where neuron dimension in each layer is 512 and the recurrent projection dimension in LSTM layer is 128. Compared to the baseline using TDNN, WER can be reduced by nearly 2.6% when LSTM is incorporated, as summarized in Table 2.

Table 2: *Performance with LSTM projected RNN integrated into baseline TDNN architecture for Dev.*

chain model	WER
TDNN	80.62
LSTM-TDNN	78.04

2.2. Combination Strategy

Combination is an effective strategy in the ASR back-end for achieving optimal transcription from different decoding streams that carry complementary knowledge. In general, combination can be processed at the lattice level via score interpolation (cf. [26]) when sharing the same AM (referred to as *lattice combination*), or at the transcript level when results come from different AMs (*system combination*) which can be achieved using minimum Bayes risk decoding [27].

Under the system chain in Figure 1, lattice combination is applied to combine the selected channels from the proposed CS approach (in the following section), provided that the first two best-ranking channels contains complementary information in terms of speakers in the signal domain. In contrast, system combination takes account of different AMs trained with different speech data from scratch, and it is required to properly choose these different but complementary systems in terms of ASR front-ends, e.g., different SE algorithms, and/or back-ends such as GMM incorporating DNN. According to the experimental results with different microphone channels and the combined version using BeamformIt in Table 3, two modules (with and without SE) are selected for system combination, assuming that the system with the second microphone (CH2) of the selected Kinect (without SE) could provide complementary recognized results for system with SE.

Figure 2 shows the combination results compared to the single system performance as well as the baseline, and both combination strategies are effective to further reduce WER by $2-3\%$. It seems that 0.5 and 0.6 are good choices for the weight of the first contribution system, i.e., 1st best channel and system with SE, for lattice and system combination, respectively.

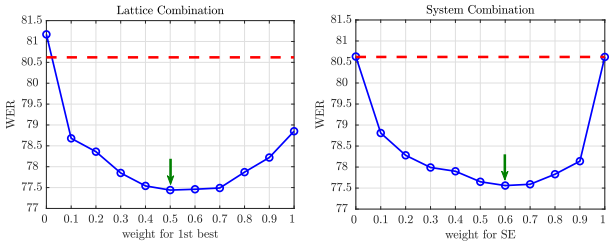

Figure 2: *Performance of lattice and system combination schemes with two contributing inputs for Dev. Dashed lines denote the baseline performance. Combination weights are varied from* 1 *(equals to the* 1*st contributing system performance) to* 0 *(equals to the* 2*nd) with step size of* 0.1.

Table 3: *WERs for baseline system with Dev speech data from worn microphone (left one) and reference Kinect array (*4 *microphones and SE when BeamformIt is used).*

worn (L)	SE	CH1	CH2	CH3	CH4
47.22	**80.62**	80.89	**80.63**	80.94	80.97

3. Channel Selection

The available speech signals from the binaural microphone worn by the party participants in the training set are used to generate the DNN model for the proposed channel selection scheme, as illustrated in Figure 3. First, a hybrid DNN-HMM AM is trained using only data from binaural microphones (16 sessions, 32 speakers, 149456 utterances) under the baseline DNN settings (cf. Section 2.1). The posterior probability $P(s,t)$ with s,t as state and frame index, respectively, is then calculated by a DNN forward-pass, and the entropy $\mathcal{E}(t)$ of $P(s,t)$ is determined by

$$\mathcal{E}(t) = -\sum_{s}(P(s,t)\cdot\log_2 P(s,t))\,. \qquad (1)$$

It has been observed that high noise levels in $P(s,t)$ (mismatched input to the reference DNN model) often increase the entropy of DNN posterior distributions [13, 14]. Therefore, channel selection can be achieved by ranking the entropy among all candidate channels, i.e., $\arg\min_c \mathcal{E}_c(t)$ with c as channel index, and the channel with lowest entropy is supposed to match the most to the binarual speech signals, which will yield the lowest WER. This is illustrated in Figure 3 where the posteriors with the binaural test sample exhibit the lowest entropy value.

Usually single-frame decisions from $\mathcal{E}(t)$ are expected to be noisy, and a temporal average/smoothing could result in a more accurate entropy comparison. Although Figure 3 shows that entropy could converge to saturation only within about 5 frames (150 ms) due to the inherent temporal context window used in the reference model (with TDNN), a temporal averaging over all frames of the test utterance (*utterance-based* processing) is used. For comparison, a longer temporal averaging window with 180 s (per speaker) is tested, which will work will if the test session is fairly static so that the best channel for each speaker does not change rapidly. Additionally, *oracle* channel selection (with available binarual speech signals in Dev as reference, but non-practical during testing in real scenarios) is achieved by a straightforward signal comparison via an objective speech intelligibility test (STOI algorithm) [19] where the highest scoring channel is presumed likely to yield the best recognition score.

Table 4 summarizes the performance from the proposed CS scheme with utterance-based and 180 s-based temporal averaging. SE (BeamformIt) is applied to the available 6 Kinects/arrays, resulting in 6 candidate channels for selection, and *ref mic* denotes the reference microphone array offered by the baseline (according to the positions of the used Kinect devices in each Session and Location during Challenge data recording). It shows that *ref mic* is not the best channel for speech recognition of active speakers, and there is room for an improvement of around 3.5% WER reduction in comparison to the *oracle* selection, which emphasizes the importance of channel selection in this distributed microphones/arrays scenario. The proposed CS strategy outperforms the baseline by around 2%, and the utterance-based CS performs better on average than the one with a longer temporal averaging window

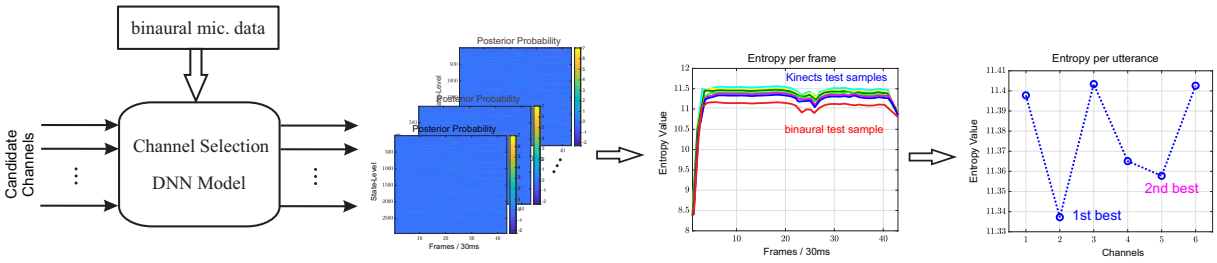

Figure 3: *The processing chain of the proposed channel selection (CS) approach, i.e., training the DNN model, forward-run to obtain the posterior probabilities, posterior entropy calculation per frame, and temporal averaging across frames.*

Table 4: *WERs with channel selection on Dev data (2 Sessions and 3 Locations) in the context of baseline acoustic model. Six Kinects/arrays with SE are used as candidate channels for selection.*

CS	Session	Kitchen	Dining	Living	Overall
Baseline (ref mic)	S02	86.50	78.89	78.64	80.62
	S09	81.39	79.60	76.65	
Oracle (STOI)	S02	76.50	78.31	72.82	76.18
	S09	78.58	77.19	76.60	
180 s -based	S02	84.80	79.26	76.42	79.51
	S09	80.27	79.29	76.60	
Utterance -based	S02	83.16	**79.21**	75.21	78.85
	S09	79.55	78.65	**77.96**	

(180 s), indicating that the best channel w.r.t. active speaker is not static enough in the test scenarios on average and CS with low latency is preferred.

Additionally, Figure 4 shows the selection decisions for each utterance in Dev from different CS methods. It can be clearly observed that compared to the oracle results by STOI algorithm, one ref mic per session or location is not sufficient in the Dev test scenario due to multiple active speakers and the dynamic/natural talking style, e.g., body and head moving, pointing directions, etc. The proposed CS method seems to be capable of efficiently and effectively selecting the reliable channel which yields (close to) the best ASR performance.

On the other hand, the proposed CS does not always guarantee an improvement for all considered sessions, e.g., the two specific sessions *'S02 Dining'* and *'S09 Living'* in Table 4 (bold text). It seems that the problem in session *'S09 Living'* can be solved by using a longer temporal averaging window for entropy ranking, as shown by the improved performance achieved using 180 s based CS method. It is also of interest to note that there exists no significant recognition improvement in this session, i.e., only 0.05% WER reduction can be achieved even by the oracle algorithm, mainly because session *'S09 Living'* is relatively static without position changes of the active speakers.

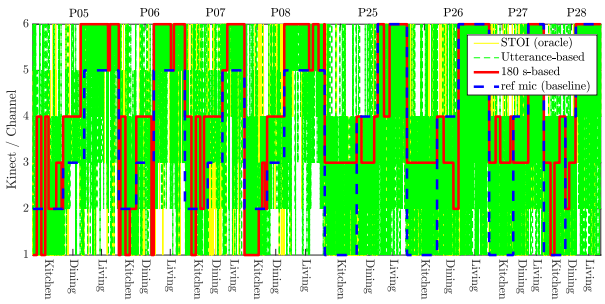

Figure 4: *The channel selection label for each utterance in Dev.*

However, this does not hold for *'S02 Dining'*, since no improvement can be observed when comparing utterance-based to 180 s based CS. It seems that many ambiguous selections (similar entropy values among channels) occur in session *'S02 Dining'* probably due to the non-diversely distributed Kinects in this session (cf. released floorplans), leading to potentially different best channel for different speaker. Note that the DNN model in the proposed CS method (cf. Figure 3) can not discriminate different speakers well when solely relying on the posterior knowledge. In order to circumvent this disadvantage, the 2nd best channel that is supposed to carry complementary knowledge w.r.t. different speakers is included to support the best channel through a lattice combination strategy (cf. Section 2.2). As shown in Table 5, the 1st best channel exhibits the best performance for the two male speakers, while the 2nd best channel is beneficial to female speakers. These respective advantages can be preserved by lattice combination, resulting in a further 2% WER reduction on average compared to single channel selection.

Table 5: *WERs with two selected best channels from utterance-based CS method for Session 'S02 Dining' in Table 4.*

CS	Speakers			
	P05(f)	P06(m)	P07(m)	P08(f)
Baseline	88.34	71.00	75.20	90.03
1st best	88.52	**70.16**	**75.20**	93.72
2nd best	**88.15**	71.07	75.80	**90.40**
Combine	86.00	66.09	73.57	90.04

Furthermore, besides the processing in the test stage, the proposed channel selection can be performed in the training stage to minimize the typical training-test-mismatch effect. Usually, data augmentation in training data improves the generalization during DNN training to further reduce WER. On the other hand, too much mismatched data introduced into the training data might degrade the final ASR performance. As shown in Table 6, it seems that it is a good compromise when integrating Kinects' data with 1st and 2nd best rankings chosen by the proposed CS method into binaural speech signals to construct the training set, rather than randomly choosing 100 K utterances from Kinects' signal as in the baseline system. Note that the proposed CS approach considers the input after SE, and the selection results are directly applied to the system without SE (i.e., CH2), due to their very similar ASR performance (potentially) resulting in similar entropy ranking in terms of CS DNN model.

Table 6: *WERs for CS (only) performing on training data selection from Kinects' signals in addition to binaural signals (L+R with* 149456 *utterances in total) from worn microphones. If SE is applied, it is applied for both training and Dev data.*

Training set	SE	CH2
Baseline (100 K: random)	80.62	80.63
CS (74728: 1st best)	79.92	80.35
CS (149456: 1st+2nd best)	**79.42**	**79.40**
CS (1598909: all Kinects)	80.83	-

Table 7 further summarizes the ASR performance when CS processing on both training and test data. As a result, 4.5% absolute WER reduction over baseline can be achieved by the proposed CS scheme.

Table 7: *WERs for CS performing on both training and Dev.*

Training	Test	SE	CH2
Baseline	ref mic	80.62	80.63
Baseline	Oracle (STOI)	76.18	-
Oracle (STOI)	Oracle (STOI)	74.82	-
CS 1st+2nd best	CS 1st best	77.40	78.03
	CS 2nd best	80.12	80.56
	Combine	76.17	76.79

4. Overall Evaluation

Table 8 summarizes the results for different subsystems tested on Dev. It shows that using channel selection for creating a new set of training data can bring 1.2% absolute improvement. A further 3.1% WER reduction can be achieved by integrating LSTM network as well as pitch features. System combination provides additional 2.7% WER reduction, resulting in a total 7% absolute ASR improvement for the single-array track. In the multiple-array track, channel selection can be applied to select the best two channels, and it is verified that the best channel outperforms the reference channel by nearly 2% when comparing *'M1'* to baseline. More importantly, the selected two channels do carry complementary information, as proved by approximately 2% improvement after lattice combination. After system combination, an overall 9.23% absolute WER reduction is achieved.

As listed in Table 9 with detailed WERs per session and location, consistent improvement (except for *'S09 Living'* with 0.9% degradation) can be observed in Dev by the proposed CS scheme from single- to multiple-array track, indicating that the proposed CS scheme is effective to exploit the potential diversity gain from the distributed Kinect devices, and the most significant improvement (nearly 3.5%) contributed by CS is attributed to sessions *'S02 Kitchen' and 'S09 Dining'*. When the final evaluation set (Eval) is tested, a similar improvement trend can be observed in single-array track in comparison to baseline, whereas some biased situations occur when CS is applied to multiple-array track. In particular, WER increases by more than 3.5% in sessions *'S21 Dining' and 'Living'*, leading to the overall WER increasing by 1% compared to single-track case. This could be partly explained by the observation from floorplans of Eval that there always exists only one nearest Kinect device during dinner party, which can not provide diversity gain for improvement in terms of channel selection in comparison to the given reference channel that could be already the best channel.

5. Conclusions

This paper presented an effective channel selection approach based on the entropy analysis of the neural network posterior probabilities, which was used to choose the reliable acoustic channel for ASR systems to yield the potentially best recognition results with distributed microphone arrays. Compared to the reference-required oracle channel selection by a speech intelligibility measure, the proposed algorithm has exhibited comparable reliability for channel selection with further advantages of reference-free and frame-based processing mode that are of great interest for real-time applications. Experimental results in the context of the development test set in the CHiME-5 Challenge showed that 3.5% word error rate reduction could be achieved by the proposed channel selection when applied to multiple-array track from single-array scenarios. However, this advantage is not fully transferable to the evaluation test set due to the lack of spatial diversity gain. More detailed analysis is required to characterize the situations in which the approach is likely to be beneficial.

6. Acknowledgments

The authors would like to thank Lin Wang for valuable discussions about blind source separation using distributed microphones. The authors also acknowledge the support of Toshiba Research Europe Limited for J. Zhang, the support of the Cluster of Excellence 1077/1 Hearing4all for B. T. Meyer, and the support of a Google Faculty Research Award for F. Xiong.

7. References

[1] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *IEEE Symposium on Communications and Vehicular Technology (SCVT)*, Ghent, Belgium, Nov. 2011, pp. 1–6.

[2] J. Dennis and T. H. Dat, "Single and multi-channel approaches for distant speech recognition under noisy reverberant conditions: I^2R'S system description for the ASpIRE challenge," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, AZ, USA, Dec. 2015, pp. 518–524.

[3] R. Lienhart, I. Kozintsev, S. Wehr, and M. Yeung, "On the importance of exact synchronization for distributed audio signal processing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, China, Apr. 2003, pp. 7423–7426.

[4] M. Wölfel, C. Fügen, S. Ikbal, and J. W. McDonough, "Multi-source far-distance microphone selection and combination for automatic transcription of lectures," in *Proceedings of Interspeech*, Pittsburgh, Pennsylvania, USA, Sep. 2006, pp. 361–364.

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

[6] K. Kumatani, J. McDonough, J. F. Lehman, and B. Raj, "Channel selection based on multichannel cross-correlation coefficients for distant speech recognition," in *Workshop on Hands-free Speech Communications and Microphone Arrays (HSCMA)*, Edinburgh, UK, May 2011, pp. 1–6.

[7] Y. Shimizu, S. Kajita, K. Takeda, and F. Itakura, "Speech recognition based on space diversity using distributed multi-microphone," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Istanbul, Turkey, Jun. 2000, pp. 1747–1750.

Table 8: *Overall WER (%) for the systems tested on the development test set. Lattice combination is based on the two selected channels (their WERs are written as* 1*st best% /* 2*nd best%).*

Ranking A Track	System	SE	Channel Selection		Pitch	LSTM	Lattice Combination	System Combination	WER
			Training	Test					
Baseline		✗	✗	✗	✗	✗	✗	-	80.63
Single	1	✓	✓	✗	✗	✗	✗	-	79.42
	2	✓	✓	✗	✓	✗	✗	-	79.17
	3	✓	✓	✗	✓	✓	✗	-	76.30
	4	✗	✓	✗	✓	✓	✗	-	76.24
	5	-	-	-	-	-	-	3+4	73.53
Multiple	1	✓	✗	✓	✗	✗	✗	-	78.85 / 81.17
	2	✓	✓	✓	✓	✓	✗	-	74.49 / 76.95
	3	✓	✓	✓	✓	✓	✓	-	72.44
	4	✗	✓	✓	✓	✓	✓	-	73.75
	5	-	-	-	-	-	-	3+4	71.39

Table 9: *Results for the best system 'S5' and 'M5' in Table 8. WER (%) per session and location together with the overall WER for both Dev and Eval (with baseline* 73.29%*).*

Track	Session		Kitchen	Dining	Living	Overall
Single	Dev	S02	80.89	72.61	70.37	73.53
		S09	73.02	73.02	69.48	
	Eval	S01	74.40	58.86	75.69	65.25
		S21	68.89	57.64	62.02	
Multiple	Dev	S02	77.41	71.30	67.57	71.39
		S09	71.58	69.61	70.38	
	Eval	S01	75.64	58.18	75.64	66.27
		S21	68.38	**61.14**	**66.24**	

[8] Y. Obuchi, "Multiple-microphone robust speech recognition using decoder-based channel selection," in *ISCA Workshop on Statistical and Perceptual Audio Processing*, Jeju, Korea, Oct. 2004.

[9] M. Wölfel, "Channel selection by class separability measures for automatic transcriptions on distant microphones," in *Proceedings of Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 582–585.

[10] M. Wolf and C. Nadeu, "Channel selection using N-best hypothesis for multi-microphone ASR," in *Proceedings of Interspeech*, Lyon, France, Aug. 2013, pp. 3507–3511.

[11] S. Jalalvand, M. Negri, D. Falavigna, M. Matassoni, and M. Turchi, "Automatic quality estimation for ASR system combination," *Computer, Speech & Language*, vol. 47, pp. 214–239, 2018.

[12] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[13] J. Barker, G. Williams, and S. Renals, "Acoustic confidence measures for segmenting broadcast news," in *International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia, Dec. 1998.

[14] H. Misra, H. Bourlard, and V. Tyagi, "New entropy based combination rules in HMM/ANN multi-stream ASR," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, China, Apr. 2003, pp. 741–744.

[15] N. Mesgarani, S. Thomas, and H. Hermansky, "A multistream multiresolution framework for phoneme recognition," in *Proceedings of Interspeech*, Makuhari, Japan, Sep. 2010, pp. 318–321.

[16] H. Hermansky, E. Variani, and V. Peddinti, "Mean temporal distance: Predicting ASR error from temporal properties of speech signal," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 1520–6149.

[17] S. H. Mallidi, T. Ogawa, K. Veselý, P. S. Nidadavolu, and H. Hermansky, "Autoencoder based multi-stream combination for noise robust speech recognition," in *Proceedings of Interspeech*, Dresden, Germany, Sep. 2015, pp. 3551–3555.

[18] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Proceedings of Interspeech*, Hyderabad, India, Sep. 2018, pp. 1561–1565.

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[20] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Analysis and outcomes," *Computer Speech & Language*, vol. 46, pp. 605–626, 2017.

[21] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[22] 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.

[23] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 2494–2498.

[24] 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 *Proceedings of Interspeech*, San Francisco, USA, Sep. 2016, pp. 2751–2755.

[25] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proceedings of Interspeech*, Singapore, Sep. 2014, pp. 338–342.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Big Island, HI, USA, Jul. 2011.

[27] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech & Language*, vol. 25, no. 4, pp. 802–828, 2011.