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

Channel-selection for distant-speech recognition on CHiME-5 dataset

Hannes Unterholzner[1], *Lukas Pfeifenberger*[1], *Franz Pernkopf*[1]
Marco Matassoni[2], *Alessio Brutti*[2], *Daniele Falavigna*[2]

[1]Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
[2]Fondazione Bruno Kessler, Center for Information and Communication Technology, Trento, Italy

hunterholzner@fbk.eu

Abstract

The 5th CHiME Speech Separation and Recognition Challenge represents a realistic scenario to validate the variety of techniques required to properly handle conversational multi-party speech acquired with distant microphones. We address the problem of channel selection using a DNN-based channel classifier that predicts good channels according to the oracle results. In combination with ROVER as a final combination step, we can improve the performance with respect to the baseline system.

1. Introduction

The paper discusses the scenarios associated to the multiple-array track of the challenge [1], considering all the channels available from the six Microsoft Kinect devices. We investigate the applicability of a channel-selection approach based on purely acoustic features (i.e. features that capture spatial information about the desired speech source) in order to identify a subset of candidate channels to combine after the decoding stage. Furthermore we discuss the approach of acoustic model adaptation [2]. We adopted the three baselines for array synchronization, enhancement (BeamformIt [3]), and conventional ASR based on a time-delayed neural network (TDNN) using lattice-free maximum mutual information (LF-MMI) [4].

2. Components

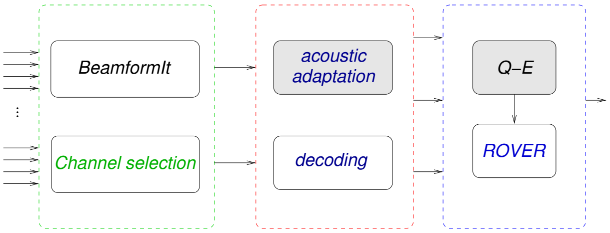


Figure 1: *The architecture of the proposed CHiME-5 automatic transcription system: signal enhancement based on BeamformIt; DNN-based channel classifier; the multiple input is decoded using a DNN-based adapted AM that exploits preliminary automatic transcription of the speech at hand; the hypotheses are finally combined to build the final output. Boxes colored in grey do not contribute in terms of improvement.*

2.1. Channel selection

The results on the a posteriori best channel selection (Sec 3.1) show that there is margin for impressive gain, if one is able to predict the best performing channel for each utterance. We define the oracle channel as the best channel, providing lowest word error rate (WER) for a given utterance. However, the oracle channel seems not to be related to the speaker position or to other spatial features. In this sense, it is extremely surprising that often two very close channels provides substantially different results. One attractive approach is employing a neural network which receives as input signal based features (i.e. filter bank features) and predicts the oracle channel. Since multiple oracle channels are available, this is a multi-label multi-class problem. We attack this problem using a DNN, trained on a subset of the training set using the binary cross-entropy loss and a sigmoid activation at the output of the last layer. Then, one can either select the best channels taking the maximum score, or can provide a channel ranking for the successive ROVER stage.

2.2. Acoustic model adaptation

It is known that adapting all the parameters of a DNN trained on a large corpus using a small adaptation set can generate overfitting. The solution adopted here is based on the principle of transfer learning where an already trained net is used to learn another task with additional examples; in this case we use *weight transfer* i.e. the last layer of the DNN is trained with a higher learning rate.

2.3. Hypotheses combination

The combination of multiple ASR hypotheses usually leads to significant improvement compared to the output of each individual system. ROVER, the most popular ASR system combination approach, performs hypotheses fusion by first building a word confusion network (CN) from the *1*-best hypotheses of the ASR systems entering the combination and then by selecting the best word in each CN bin via majority voting [5].

The hypotheses combination process considers the first input candidate as a “skeleton” to align the other hypotheses in a greedy manner. For this reason, depending on the order in which the hypotheses are considered when feeding the algorithm, the resulting combination can show large variations in quality. In the past we developed a system [6, 7] for optimally ranking the ASR hypotheses that feed ROVER. However, due to time constraints, this system has not been applied yet, and ASR hypotheses produced for this challenge are ranked with the approximate method described in Section 2.1. The application of the optimal ranking approach on CHIME-5 evaluation sets will be done in future.

10.21437/CHiME.2018-20

3. Experimental evaluation

3.1. Oracle results

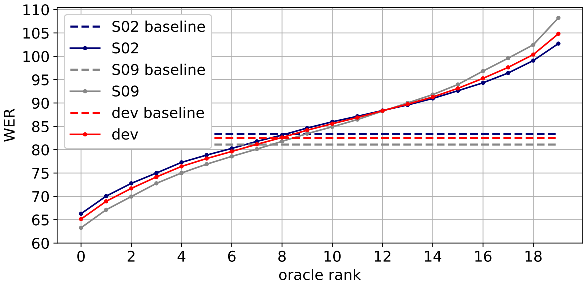
The oracle provides an upper performance bound by selecting the best hypothesis according to the WER among a set of decoded channels on utterance-level. Several oracle experiments were conducted with different sets of hypotheses in order to investigate the possible performance gain. Table 1 shows the oracle scores for the development set and its individual sessions that serve as an indicator to which extend both the best selection and the set of channels impact the final performance.

Table 1: *WER (%) results of the oracle for different sets of decoded channels. U indicates a set of 20 single array channels while U_ref are the four channels from the reference array provided by the baseline system. The parenthesized number states the number of available channels in this specific setting. BfIt stands for the BeamformIt beamformer enhancement.*

Channels	**Dev**		
	S02	S09	Overall
Baseline: U_ref + BfIt (1)	83.4	81.1	82.5
U_ref (4)	76.1	72.8	74.8
U + BfIt (5)	70.8	68.2	69.3
U (20)	66.3	63.3	65.1
U + BfIt, U (25)	65.5	62.3	64.3
U_ref, U + BfIt, U (29)	64.6	62.2	63.6

Most of the experiment are conducted without using array five, since it is partially missing in the development set. However, information coming from this array is introduced when U_ref is part of the channel set. Using all the available hypotheses from 29 channels the oracle scores to a WER of 63.6% on the development set, which is a total of 18.9% in absolute word error rate reduction compared to the baseline system. Moreover importantly, remarkable results are also obtained when solely selecting among the 20 single array channels, without using any enhanced signal or information of the reference array. However, as illustrated in Figure 2, a good selection is crucial since the performance drastically decreases with increasing oracle rank.

Figure 2: *WER (%) results for the development set on the per utterance oracle informed channel ranks, considering the 20 array channels.*



3.2. Channel selection

Table 2 presents the WER of the best selected channel for the full development set and separately for its two sessions. We investigated different signal based features for either directly ranking the channels or as input for the classifier network. Apart from the frame based mel-filter-bank features, all other features are computed on utterance level (i.e. one feature vector per utterance). For the envelope variance (EV) [8] features we calculate 12 sub-bands for each channel without computing a weighted sum over the normalized sub-band variances, since this is expected to be learned from the network. For the entropy features we take the average of the DNN posterior entropy over the full utterance. The network itself is an LSTM (1 recurrent layer followed by 2 dense layers) when the frame based features serve as input, otherwise it is composed by two fully connected layers, trained on session S03, S04 S07 and S13; training on all the sessions did not lead to an improvement. Only a subset of the utterances was selected for training, removing those for which all channels perform equally. Note that since device U05 is not available for session S09 of the development set, we considered only 20 channels. At this point taking always the best scored channel provided by the classifier does not lead to a gain in performance. However, we observed that the network learns the training data by heart which indicates the need for more informative features and proper regularization techniques to allow proper generalization.

Table 2: *WER (%) results for both direct and classifier based channel selection for different feature types.*

Method	**Channels**	**Feature**	**Dev**		
			S02	S09	Overall
direct	U+BfIt (5)	Energy	81.2	81.6	81.3
		GCC-PATH	81.1	81.7	81.4
	U (20)	Energy	82.2	82.2	82.2
classifier	U (20)	Energy	82.2	82.7	82.8
		EV	83.7	82.6	82.7
		Fbank	83.8	83.5	83.7
		Entropy	81.7	82.8	82.1

3.3. Hypotheses combination

We combined the channels with respect to the ranking obtained from the classifier trained on the mel-filter-bank features. As illustrated in Figure 3 we gain the most in case when we train on four sessions and combine the top 10 channels. For comparison hypotheses fusion was also conducted on a random and the oracle informed ranking. The overall results on the mulitple-array track are listed in Table 3. With the channel selection approach described in 3.2 and ROVER we gain 3% in absolute WER compared to the baseline system.

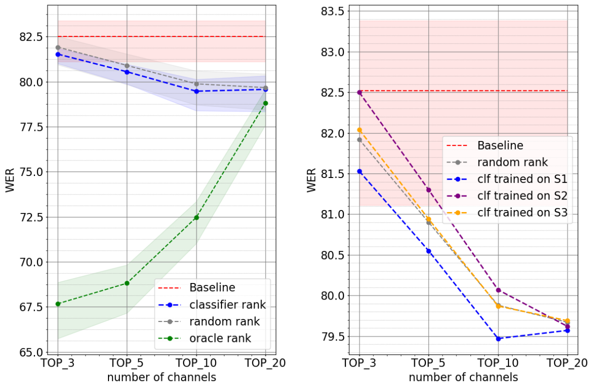
Table 3: *Results for the best system. WER (%) per session and location together with the overall WER.*

Track	Session	Kitchen	Dining	Living	Overall
Multiple	S02	83.6	79.5	77.3	79.5
	S09	78.4	78.8	79.5	

3.4. Remarks

The approach of acoustic model adaptation is implemented updating the output layer of the DNN with small adaptation sets. We conducted a first experiment by adapting the network with a selected subset of utterances from one session of the development set while testing on the other one. Exploiting the information from the decoded channels, the adaptation set consists

Figure 3: *WER (%) results after applying ROVER on the top N channels. Transparency colored regions states the performance deviation among the two development sessions. Classifier trained on 4 sessions (S1), 6 sessions (S2) and 10 sessions (S3).*



of utterances with WERs below a certain threshold (i.e. 60%). However, this approach was not successful, probably due to the available bad automatic supervision.

4. Conclusions

The proposed system is built upon the modules developed for the previous challenge [9]; nevertheless, the CHiME-5 is characterized by an extremely challenging scenario and many beneficial techniques proved to be effective in earlier work need specific customization to this task. As a consequence, resulting WERs are still unsatisfactory and additional work is required. In particular, it seems promising to address channel selection since, from the oracle results presented in Section 3.1, a large gain is expected.

5. References

[1] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, “The fifth CHiME Speech Separation and Recognition Challenge: Dataset, task and baselines,” in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018)*, Hyderabad, India, 2018.

[2] M. Matassoni, R. Gretter, D. Falavigna, and D. Giuliani, “Non-native children speech recognition through transfer learning,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Calgary, Canada, April 2018.

[3] 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. 2011–2021, September 2007.

[4] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, Y. Wang, X. Na, and S. Khudanpur, “Purely sequence-trained neural networks for asr based on lattice-free mmi,” in *Interspeech*, 2016.

[5] J. G. Fiscus, “A Post-processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER),” in *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. Santa Barbara, CA, USA: IEEE, 1997, pp. 347–354.

[6] S. Jalalvand, M. Negri, M. Turchi, J. G. de Souza, D. Falavigna, and M. R. Qwaider, “Transcrater: a tool for automatic speech recognition quality estimation,” in *Proceedings of ACL-2016 System Demonstrations. Berlin, Germany: Association for Computational Linguistics*, 2016, pp. 43–48.

[7] S. Jalalvand, M. Negri, D. Falavigna, M. Matassoni, and M. Turchi, “Automatic quality estimation for ASR system combination,” *Computer Speech and Language*, April 2017.

[8] M. Wolf and C. Nadeu, “Channel selection measures for multi-microphone speech recognition,” *Speech Commun.*, vol. 57, pp. 170–180, Feb. 2014. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2013.09.015

[9] M. Matassoni, M. Ravanelli, S. Jalalvand, A. Brutti, and D. Falavigna, “The fbk system for the chime-4 challenge,” in *4th International Workshop on Speech Processing in Everyday Environments*, San Francisco, US, September 2016.