ADAPTIVE BEAMFORMING AND ADAPTIVE TRAINING OF DNN ACOUSTIC MODELS FOR ENHANCED MULTICHANNEL NOISY SPEECH RECOGNITION

[1,3]*Alexey Prudnikov,* [2,3]*Maxim Korenevsky,* [3]*Sergei Aleinik*

[1]Speech Technology Center Inc., [2]STC-Innovations Inc., [3]ITMO University, Saint-Petersburg, Russia
{korenevsky,prudnikov,aleynik}@speechpro.com

ABSTRACT

This paper describes our contribution to the development of an ASR system for the CHiME 2015 Challenge. We applied a new adaptive beamforming method of multichannel alignment for enhancing speech recorded with six microphones. Then we trained an effective CD-DNN-HMM acoustic model using CMVN for noise robustness as well as fMLLR and i-vectors for speaker and environment adaptation. As a result, our system provides 7.33% WER on the development set and 14.34% WER on the test set (58% WER reduction compared to the baseline system).

Index Terms— speech enhancement, beamforming, multichannel alignment, adaptation, i-vectors

1. INTRODUCTION

Despite the significant progress made in speech recognition in recent years, recognition under the conditions of complex transient interference remains an important task. The purpose of the CHiME Challenge is to develop new approaches to speech recognition in noisy multisource environments which are to facilitate greater use of speech recognition in everyday life. The CHiME 2015 Challenge [1] focuses on the recognition of speech recorded under real-world noisy conditions by several spaced microphones. This naturally suggests using various approaches developed for denoising and enhancing speech recorded with microphone arrays [2]. These approaches make it possible to select and amplify the speech signal from the look direction while suppressing interference and noise from the other directions. The speech signal obtained from a microphone array can be further processed to increase SNR and improve intelligibility.

The CHiME 2015 organizers provided a speech corpus recorded synchronously using 6 microphones mounted around a tablet PC. Five of them were oriented towards the user's side and the last one (middle upper in horizontal placement) in the opposite direction [1]. A baseline speech enhancement system was also provided. It is based on speaker localization followed by MVDR beamforming [3]. A modern baseline acoustic modeling system based on DNN-HMM was provided as a Kaldi [4] recipe. A MATLAB script for generating simulated recordings with similar properties which can be used for extending the training dataset was provided as well.

This paper describes our contribution to the development of an improved ASR system using the provided acoustic data. We used the baseline speaker localization algorithm but employed a new adaptive beamforming approach [5] which brings a significant gain in recognition accuracy. Another contribution is a more thorough acoustic modeling which uses adaptation and normalization methods extensively. Firstly, to train the DNN we used MFCC features adapted to a speaker and a channel, while the baseline recipe uses fBanks (log mel-scaled filter-bank energies) with no normalization or adaptation. Secondly, when training the DNN we additionally used i-vectors [6], which comprise information about both the speaker and the acoustic environment. This approach proposed in [7] reduces model dependence on the aforementioned acoustic variability factors and thereby increases recognition accuracy. To extract i-vectors from the recordings we employed the recently developed approach [8] which uses DNN trained for speech recognition.

The paper is organized as follows. Section 2 briefly describes the speech recording conditions and Challenge data preparation. Section 3 deals with the processing of microphone array signals and describes the new adaptive beamforming algorithm we propose to use. In Section 4 we discuss in detail our improvements in acoustic modeling and present the results of the main experiments. Conclusions and possible directions of future work are outlined in Section 5.

2. DATA FOR EXPERIMENTS

According to the scenario developed by CHiME 2015 organizers, several speakers were asked to read sentences taken from the well-known speech corpus WSJ0 [9], displayed on the tablet screen. Their speech was recorded in four real-life conditions, namely in a cafe, bus, street junction and pedestrian area. Speech was recorded synchronously by both the close-talking microphone and 6 distant microphones mounted around the tablet screen (3 microphones per long side). One of the microphones (the second, middle upper) was oriented

to the opposite side from the speaker. The placement and orientation of the tablet were arbitrary, so the location of the speaker with respect to the microphones could vary from one recording to another and even during a single recording.

To train the acoustic models, the baseline Kaldi recipe uses both real-life recordings and data simulated by the special MATLAB utility. Real-set consists of about 5.6 hrs audio data per channel. Simu-set was generated on the basis of WSJ0 data and real-life noise recordings and consists of about 18 hrs per channel. The trainset and the devset were created from the joint sample of real and simu (hereafter *multi*) and their sizes ratio is about three-to-one. The trainset is intended to be used for acoustic model training while the devset for testing and tuning recognition parameters. The testset was provided additionally for the final testing of the best ASR system configuration.

Let us note that the Challenge participants were allowed to use a script for new simulated data generation to extend the training set. However, our experiments showed that real and simu data have rather different acoustic features, so we disregarded this possibility to avoid bias of the system to better recognition of simulated data. In fairness, however, we should note that training with only real data provides worse results compared to training with the whole multi-trainset while testing on both real and simu data.

3. SPEECH ENHANCEMENT

3.1. Baseline system

As noted above, the position of the speaker in relation to the microphone array could vary during recording. However, in order to apply methods of beamforming, the exact position of the speech source during the recording must be known, which requires the use of speaker localization methods. Baseline speech enhancement provided by the organizers uses the SPR-PHAT method [10] supplemented by the Viterbi algorithm to find the most probable trajectory of the speakers mouth changing its location. After speaker localization, the baseline speech enhancement system applies MVDR beamforming with diagonal loading [11] and calculates noise spatial covariances.They are estimated over a short segment of background signal preceding speech

Two GMM-HMM acoustic models were trained according to the first part of the baseline Kaldi recipe. One of them was trained on the noisy recordings from only one (5th) microphone and another on the data obtained from the baseline speech enhancement. The word error rates (WERs) observed for recognizing the corresponding parts of the devset are presented in the table 1. The upper and lower parts of the table correspond to using the 5th microphone recordings (noisy) and speech enhancement results, respectively, in both training and recognition. The abbreviations BUS, CAF, PED, STR and AVG stand for bus, cafe, pedestrian area and street junction conditions and average over all of them respectively.

Table 1. *Development set WER,% of the baseline system on GMM-HMM models*

Test data	BUS	CAF	PED	STR	AVG
train multi noisy					
real noisy	25.86	17.89	12.86	17.93	18.64
simu noisy	18.57	21.70	15.03	17.35	18.16
average	22.22	19.80	13.95	17.64	18.40
train multi enhanced					
real enh.	23.43	18.86	17.41	20.34	20.01
simu enh.	8.51	11.99	8.67	10.81	10.00
average	15.96	15.43	13.04	15.58	15.01

The table shows that using the baseline speech enhancement decreases the average WER over real and simu. However, while WER on simu data decreases almost by factor of two, WER on real data increases. The CHiME 2015 organizers attribute this difference to some inadequacy of simu data generation. However, our results indicate that the average WER can be reduced much more without a large gap between the real and simu results.

3.2. Beamforming methods

In our speech enhancement experiments we used the provided speaker localization algorithm and several most widespread algorithms of microphone array signal processing. It should be noted that prior to applying any beamforming algorithms we have equalized a signal power in every microphone channel because large differences in signal level may cause these algorithms to perform poorly.

First of all we used the widely-known Delay&Sum algorithm [2], which can be described in frequency domain by the expression

$$\begin{aligned} Y_{D\&S}(f,t) &= \frac{1}{M}\sum_{k=1}^{M} X_k(f,t)e^{-j\Delta\varphi_k(f,t)} = \\ &= \frac{1}{M}\sum_{k=1}^{M} Y_k(f,t), \end{aligned} \quad (1)$$

where $X_k(f,t)$ is the complex spectrum of the k-th microphone signal in the frequency band $f = 1,\ldots,N$ on the t-th frame, $Y_k(f,t)$ is the corresponding "pre-steered" spectrum, $Y_{D\&S}(f,t)$ is the complex spectrum of the resulting signal and

$$\Delta\varphi_k(f,t) = 2\pi F_s \frac{f-1}{N}\tau_k(t)$$

is the phase shift of the k-th microphone in the frequency band f, F_s is the sampling rate, $\tau_k(t)$ is the delay time for the k-th microphone on the t-th frame for aligning the phases.

The Delay&Sum method forms a fixed directivity pattern of the microphone array, which does not depend on received

signals. Adaptive beamforming methods continuously update their parameters based on the received signals. One of the most widespread adaptive methods is the MVDR (Minimum Variance Distortionless Response) beamforming [12] implemented in the baseline speech enhancement system. In this implementation the spatial covariance matrix of noise is estimated over a short segment of the background signal preceding speech (and is not changed further), and diagonal loading [11] is used.

Another well-known adaptive method is the GSC (General Sidelobe Canceller) beamformer [13]. In algorithms based on the GSC idea microphone signals are fed into a "Blocking matrix" which is intended to form target speech free signals. Then these signals are used to adaptively form the error signal subtracted from the "quiescent" signal (usually $Y_{D\&S}(f,t)$) for better noise suppression. The adaptive weights of error signals may be determined based on different optimization criteria. According to the published results, very good performance is observed when maximizing non-Guassianity of the resulting signal [14]. Since the Gaussian distribution is specific to the noise spectrum these approaches make resulting signal different from noise as much as possible.

Various postfiltering algorithms are often applied on top of beamforming to to suppress the residual noise. The most widespread approach is the Zelinski postfilter [15], [16] and its modification proposed by Simmer et al. [17]. A weakness of the Zelinski postfilter is that better noise suppression often comes with speech distortions which can be detrimental for subsequent recognition.

We applied most of these beamforming algorithms to the raw CHiME 2015 data (both trainset and devset) and trained acoustic models using the baseline recipe up to the stage of triphone fMLLR-SAT GMM-HMM models. The results are presented in the upper part of the table 2. It turns out that the simplest Delay&Sum beamformer provides the best recognition quality on real data (as well as on the average over real and simu). Moreover, when using Delay&Sum the gap in WER between real and simu results becomes insignificant. Different implementations of GSC showed results comparable to MVDR or worse, which is possibly explained by implementation drawbacks. The Zelinski postfilter applied on top of different beamformers also deteriorated recognition. The Simmer et al. postfilter applied on top of the Delay&Sum showed comparable results to those obtained without a postfilter, but WER on real data increased.

3.3. The influence of the second microphone

As pointed above, the second microphone was oriented to the side opposite to that of five others, so the target speech on it is attenuated and the signal is dominated by noise. It was probably done to facilitate using this microphone's signal as a reference for adaptive noise subtraction. However, noise signal of second microphone turns out to be weakly coherent to those of other microphones, so adaptive noise subtraction fails. On the other hand, the permanent rejection of the second microphone provides significantly better results on real data with all the beamforming algorithms we used. This is demonstrated by the results from the lower part of the table 2.

Table 2. *Development set WER,% of GMM-HMM models for various beamforming algorithms. "D&S" stands for Delay&Sum, "+Z" and "+S" stand for application of the Zelinski [16] and Simmer et al. [17] postfilters respectively*

Data	MVDR	D&S	D&S+Z	D&S+S	MCA
Six microphones					
real	20.01	13.90	18.51	14.59	13.06
simu	10.00	13.58	14.83	13.01	10.42
aver	15.01	13.74	16.67	13.80	11.74
Without the second microphone					
real	18.20	12.43	14.29	12.75	10.72
simu	10.78	14.52	15.25	14.14	12.50
aver	14.49	13.48	14.77	13.45	11.61

3.4. Multi-Channel Alignment (MCA) method

A new method of adaptive beamforming was recently proposed in [5]. It is very simple to implement and demonstrates good results in practice. The Multi-Channel Alignment (MCA) method works as follows:

1. Complex spectra of microphone signals are phase-compensated and averaged (just as in Delay&Sum method) to form the signal $Y_{D\&S}(f,t)$ (1) of fixed beamforming .

2. Then the transfer function is calculated for every channel:
$$H_k(f,t) = \frac{\langle \Phi_{Y_k Y_{D\&S}}(f,t) \rangle}{\langle \Phi_{Y_k Y_k}(f,t) \rangle}, \quad (2)$$
where Y_k are defined in (1) and $\langle \cdot \rangle$ means temporal exponential smoothing of the cross-spectrum (in nominator) or power spectrum (in denominator):
$$\langle \Phi_{XY}(f,t) \rangle = \alpha \langle \Phi_{XY}(f,t-1) \rangle + \\ + (1-\alpha) X(f,t) Y^*(f,t), \quad (3)$$
and $(\cdot)^*$ stands for complex conjugate.

3. Transfer functions are multiplied on the corresponding complex spectra and the results are averaged over all channels:
$$\begin{aligned} Y_{MCA}(f,t) &= \frac{1}{M} \sum_{k=1}^{M} Y'_k(f,t) = \\ &= \frac{1}{M} \sum_{k=1}^{M} H_k(f,t) Y_k(f,t). \quad (4) \end{aligned}$$

The important feature of this approach is a double (in decibels) suppression of sidelobes level (i.e. the suppression coefficient is squared!) compared to Delay&Sum method. Besides, when the expression (2) is used the width of the mainlobe gets more than twice less compared to Delay&Sum method. However, this can sometimes lead to undesirable effects like appearance of grating lobes on high frequencies. In such cases the magnitude of the transfer functions (2) can be used instead of complex expressions.

We applied MCA to the CHiME 2015 recordings and obtained the best result among all the tested beamforming algorithms. These results are presented in the rightmost column of the table 2. Note that an average WER for MCA is comparable for both using all microphones and all but the second one, however result on the real data is much better when the second microphone is rejected. That is why the variant without the second microphone was chosen as a basic in the subsequent experiments. The results presented in table 2 for MCA algorithm correspond to the spectral smoothing (3) with $\alpha = 0.7$ which provides best result on real devset.

All the subsequent results relate to this variant of speech enhancement algorithm (MCA with $\alpha = 0.7$ for spectral smoothing and without second microphone) unless otherwise stated.

4. ACOUSTIC MODELING ENHANCEMENT

A baseline acoustic modeling system provided by the Challenge organizers as a Kaldi recipe, consists of the following main stages:

1. Training of speaker-independent (SI) triphone GMM-HMM model on MFCC with CMN;
2. Splicing feature vectors from 7 consecutive frames with a current frame in the middle, computation of LDA-MLLT transform and re-training triphone GMM-HMM model on the transformed features;
3. Training fMLLR-SAT triphone GMM-HMM model on top of previous features;
4. Splicing 40-dimensional vectors of fBank features from 11 consecutive frames with a current frame in the middle. Training of DNN using cross-entropy criterion, hereafter CE. Preparation of the CD-DNN-HMM model;
5. Additional sequence discriminative training of DNN using sMBR criterion according to [18].

In the experiments described in Section 3 we used only three first stages of this recipe. To improve recognition accuracy we changed and substantially extended two last stages. We paid considerable attention to more thorough and effective adaptation of the acoustic model to the speaker and environment, as well as obtaining acoustic features with much better discriminative abilities. The devset recognition results obtained on different stages of the baseline Kaldi recipe (3, 4 and 5) are presented in the table 3 for both baseline and proposed speech enhancement algorithms.

Table 3. *Development set WER,% on different stages of the baseline recipe*

Data	GMM-HMM	DNN (x-ent)	DNN (sMBR)
Baseline enhancement			
real	20.55	20.00	17.72
simu	9.79	9.34	8.17
aver	15.17	14.67	12.95
MCA beamforming			
real	10.72	11.14	10.41
simu	12.50	12.63	11.57
aver	11.61	11.88	10.99

Table 3 shows that step from GMM-HMM to DNN-HMM in the baseline recipe results in only slight WER improvement. That is why we focused on the improving the procedure of training DNN-based acoustic model.

We combined the main steps of our experiments into the general scheme shown in Figure 1. In this scheme:

- the blocks where acoustic model is built are depicted in green;
- the abbreviation "concat" means the concatenation of input vectors;
- the term "Splice $m/\pm k$' stands for splicing feature vectors of $(2m + 1)$ frames separated by k frames, i.e. the concatenation of vectors $\mathbf{o}_{t-km}$, ..., $\mathbf{o}_{t-k}$, $\mathbf{o}_t$, $\mathbf{o}_{t+k}$, ..., $\mathbf{o}_{t+km}$;
- "CE DNN" and "sMBR DNN" stand for DNN training with CE and sMBR minimization criteria respectively.

Other blocks are described in more detail below. Original Kaldi tools were used for GMM-HMM training and DNN sequence training with sMBR criterion. Nesterov accelerated gradient (NAG) algorithm [19] with momentum 0.7 was used for cross-entropy DNN training.

It should be noted that we used the same MFCC input features (transformed with LDA-MLLT and fMLLR) for training both GMM-HMM and DNN-HMM models. This differs from the baseline recipe where fBank features are used to train DNNs. The blocks sequence 1–8 corresponds to the training steps of the baseline recipe used in the experiments of Section 3 (the only difference is using CMVN instead of CMN in block 2).

4.1. Adaptation using fMLLR and i-vectors

The importance of adapting acoustic models to the speaker and environment characteristics for improving recognition ac-

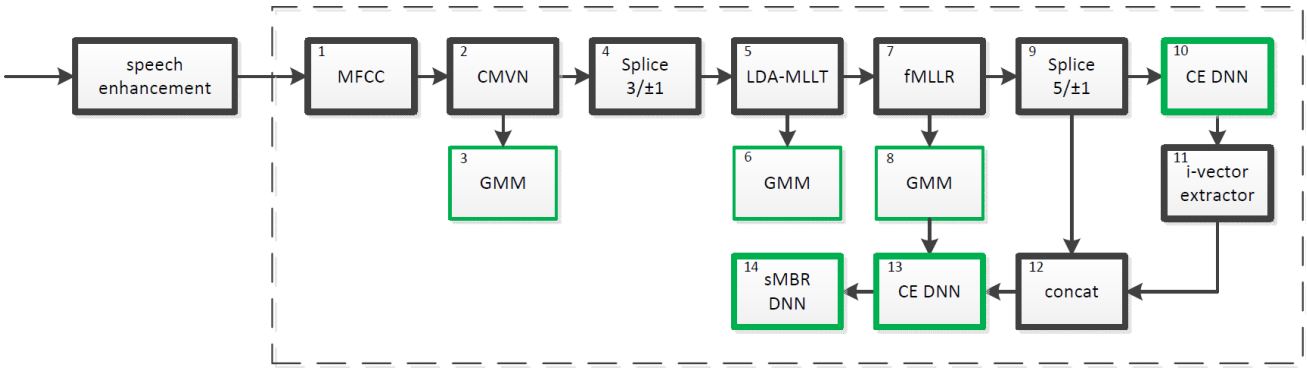

Fig. 1. The scheme of acoustic model training stages

curacy is widely known [20]. The adaptation methods for GMM have been evolving for the last 30 years. Feature space MLLR (fMLLR) [21] is one of the most popular ways of speaker adaptation because it operates in the feature space and thus is well-suited for speaker adaptive training. The development of DNN adaptation methods has started relatively recently and many of them require some modification of original unadapted model, which can be complex and laborious. In [7] the i-vectors which are widely used for the speaker recognition tasks were proposed to use for SAT-training of DNNs, and it was noted that fMLLR and i-vectors are complementary. The i-vector extracted from a recording (or several recordings, pertaining to the same speaker) is appended to the spliced feature vector and the extended vector is used as a DNN input. We have already used this approach in our experiments on Russian spontaneous speech recognition [22] and it provided considerable WER reduction, thus we decided to use it for CHiME 2015 as well.

The extension of DNN input layer with i-vectors was proposed in [23]. Applying a special kind of regularization which penalizes the deviation of the learnt DNN weights from those of the basic DNN (without i-vectors) was also proposed. It should be noted that on the CHiME 2015 data fMLLR and i-vectors based adapation seem to be complementary which as in [7]. To extract i-vectors in this work we used an idea proposed in [8], [24]. This is a very convenient approach because it uses just previously trained speech recognition DNN to extract i-vectors, but not some special models.

Using the combination of i-vectors and fMLLR made it possible to decrease WER of CE DNN-models by 0.4% on average compared to fMLLR only system. Additional sMBR-training further reduces WER by more than 1% on average (see the results in the table 4).

4.2. Analysis of the detailed recognition results

In this section we present the detailed recognition results on the devset and testset obtained using both baseline and proposed algorithms of speech enhancement and acoustic modeling. The comparison of these results shows what are the contributions of the proposed algorithms separately and in combination.

Table 4. *Development set WER,% of DNN-HMM models with and without using i-vectors*

Features, criterion	**real**	**simu**	**aver**
fMLLR, CE	8.75	10.59	9.67
fMLLR+i-vectors, CE	8.32	10.21	9.27
fMLLR+i-vectors, sMBR	7.33	9.12	8.22

Table 5 contains recognition results of different acoustic models on both dev and test datasets for several speech enhancement algorithms used. We can make several conclusions from these results:

- The difference between test and dev real data is significant. And despite applied normalization and adaptation techniques we still could not make the WER difference sufficiently small. To mitigate the influence of acoustic difference we tried to use other robust acoustic features such as PNCC [25], ETSI AFE [26] and MVA [27] but MFCC+CMVN provided the best results.

- Almost all the applied steps of acoustic modeling have similar effect (in terms of relative WER reduction) when training on data from both proposed and baseline speech enhancement.

- The baseline speech enhancement provides the best results on the simulated data and the worst results on the real data. In contrast, the proposed speech enhancement provides much better results on the real data (the WER reduction is 36–46% depending on dataset) and worse results on the simulated data.

- The proposed speech enhancement provides much smaller difference of recognition results on real and simulated data than the baseline algorithm does.

Table 5. *WER,% of different system configurations*

Model	Features	Baseline enhancement				Proposed enhancement			
		dev		test		dev		test	
		real	simu	real	simu	real	simu	real	simu
GMM-HMM	fMLLR	18.22	10.65	29.74	11.49	10.90	12.90	18.15	18.51
DNN-HMM, CE	fMLLR	14.27	8.07	24.49	9.47	8.75	10.59	15.47	15.25
DNN-HMM, CE	fMLLR+i-vectors	13.94	7.69	24.25	8.86	8.32	10.21	15.18	14.89
DNN-HMM, sMBR	fMLLR+i-vectors	12.34	6.47	21.28	7.16	7.33	9.12	14.34	13.84

The results for all combinations of the baseline and proposed acoustic modeling and speech enhancement are given in the table 6.

Table 6. *Recognition results (WER,%) on different combinations of the baseline/proposed systems*

Acoustic modeling	Speech enhancement	dev		test	
		real	simu	real	simu
baseline	baseline	17.72	8.17	33.76	11.19
baseline	proposed	10.71	11.35	22.63	23.59
proposed	baseline	12.34	6.47	21.28	7.16
proposed	proposed	7.33	9.12	14.34	13.84

From the table 6 we can conclude that

- the proposed acoustic modeling provides the relative WER reduction of 21-33% compared to baseline training recipe;
- the proposed speech enhancement provides the relative WER reduction of 36–46% on real data compared to the baseline speech enhancement; however, the WER on simulated data increases to 40–98% relative depending on dataset;
- a combination of the both proposed approaches provides both the relative WER reduction up to 58% on real data and slight WER reduction on simulated data.

The detailed recognition results of the ASR system which provides the best results on the real datasets is given in the table 7.

Table 7. *Detailed recognition results of the best configuration*

Environment	dev WER,%		test WER,%	
	real	simu	real	simu
BUS	9.31	7.67	17.37	9.43
CAF	7.20	11.15	11.47	14.16
PED	5.25	7.68	18.01	14.16
STR	7.57	9.96	10.52	17.61
AVG	7.33	9.12	14.34	13.84

5. CONCLUSIONS AND FUTURE WORK

We described our contribution to the development of ASR system for recognizing speech recorded with microphone array in real-life conditions. The new adaptive multichannel alignment beamforming method we used provides an effective way to enhance speech and suppress noises recorded on different microphones. Another improvement relates to the acoustic modeling. We applied two adaptation approaches, namely fMLLR and using i-vectors to effectively adapt ASR system to both the speaker identity and the environment. As a result, our CD-DNN-HMM model trained with sMBR criterion provides large WER reduction on real data compared to provided baseline system.

We believe that our results can be improved further by means of using more sophisticated speech enhancement and acoustic modeling techniques. So, our preliminary experiments on GMM-HMM models show that we can reduce devset WER by 0.5-1% absolute when using decision-directed SNR estimation [28] followed by MMSE Log-Spectral Amplitude (LSA) [29] estimation on top of the proposed beamforming. Comparable improvements can be obtained by application beamforming on different microphone subsets, training several acousic models on beamforming results and applying model combination techniques like ROVER [30]. Another promicing direction of the entire system improvement is further reduction of WER difference between devset and testset results.

6. ACKNOWLEDGEMENTS

This work was partially financially supported by the Ministry of Education and Science of the Russian Federation, Contract 14.575.21.0033 (ID RFMEFI57514X0033).

7. REFERENCES

[1] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *Proceedings of 2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)*, 2015, submitted.

[2] J. Benesty, J. Chen, and Y. Huang, *Microphone array*

signal processing, Springer Science & Business Media, 2008.

[3] H.L. Van Trees, *Optimum Array Processing*, NY: Wiley-Interscience, New York, 2002.

[4] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *Proceedings of 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2011)*, 2011.

[5] M. Stolbov and S. Aleinik, "Improvement of microphone array characteristics for speech capturing," *Modern Applied Science*, vol. 9, no. 6, pp. 310–319, 2015.

[6] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proceedings of 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, 2009.

[7] G. Saon, H. Soltau, D. Nahamoo, and M Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proceedings of 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2013)*, 2013, pp. 55–59.

[8] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically aware deep neural network," in *Proceedings of 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, 2014, pp. 1695–1699.

[9] J.S. Garofolo, D. Graff, D. Paul, and D. Pallett, *CSR-I (WSJ0) Complete LDC93S6A*, Linguistic Data Consortium, Philadelphia, PA, 1993.

[10] H. Do and H.F. Silverman, "Srp-phat methods of locating simultaneous multiple talkers using a frame of microphone array data," in *Proceedings of 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, 2010, pp. 125–128.

[11] J. Li, P. Stoica, and Wang Z., "On robust capon beamforming and diagonal loading," *IEEE Transactions on Signal Processing*, vol. 51, no. 7, pp. 1702–1715, 2013.

[12] J. Capon, "High resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, pp. 1408–1418, 1969.

[13] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," in *Proceedings of 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1996)*, 1996, pp. 925–928.

[14] K. Kumatani, T. Arakawa, K. Yamamoto, J. McDonough, B. Raj, R. Singh, and Tashev I., "Microphone array processing for distant speech recognition: Towards real-world deployment," in *APSIPA Annual Summit and Conference*, 2012.

[15] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proceedings of 1988 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1988)*, 1988, pp. 2578–2581.

[16] H. W. Lollmann and P. Vary, "Post-filter design for superdirective beamformers with closely spaced microphones," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007, pp. 291–294.

[17] K.U. Simmer, Fischer S., and Wasiljeff A., "Suppression of coherent and incoherent noise using a microphone array," *Annals of telecommunications*, vol. 7/8, pp. 439–446, 1994.

[18] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence discriminative training of deep neural networks," in *Proceedings of 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, 2013, pp. 2345–2349.

[19] Yu. Nesterov, *Introductory Lectures on Convex Optimization. A Basic Course*, Kluwer Academic Publishers, 2004.

[20] D. Giuliani and R. De Mori, Eds., *Speaker adaptation*, chapter 11, pp. 363–404, Academic Press, 1998.

[21] M.J.F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," Tech. Rep., Cambridge University Engineering Department, May 1997.

[22] A. Prudnikov, I. Medennikov, V. Mendelev, M. Korenevsky, and Y. Khokhlov, "Improving acoustic models for russian spontaneous speech recognition," in *Proceedings of 17th International Conference on Speech and Computer (SPECOM 2015)*, 2015, accepted for publication.

[23] A. Senior and I. Lopez-Moreno, "Improving dnn speaker independence with i-vector inputs," in *Proceedings of 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, 2014, pp. 225–229.

[24] S. Novoselov, T. Pekhovsky, O. Kudashev, V. Mendelev, and A. Prudnikov, "Non-linear plda for i-vector speaker verification," in *Proceedings of 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015)*, 2015, accepted for publication.

[25] C. Kim and R. Stern, "Power-normalized cepstral coefficients (pncc) for robust speech recognition," in *Proceedings of 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, 2012, pp. 4101–4104.

[26] European Telecommunications Standards Institute, *Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithms*, es 202 050, rev. 1.1.5 edition, 2007.

[27] C.-P. Chen and J. Bilmes, "Mva processing of speech features," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 257–270, January 2009.

[28] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[29] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[30] J.G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Proceedings of 1997 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 1997)*, 1997, pp. 347–354.