A CHIME-3 CHALLENGE SYSTEM: LONG-TERM ACOUSTIC FEATURES FOR NOISE ROBUST AUTOMATIC SPEECH RECOGNITION

Niko Moritz[1,4], *Stephan Gerlach*[1,4], *Kamil Adiloglu*[2,4],
Jörn Anemüller[3,4], *Birger Kollmeier*[1,2,3,4], *Stefan Goetze*[1,4]

[1]Fraunhofer IDMT, Project Group for Hearing, Speech, and Audio Technology, Oldenburg, Germany
[2]Hörtech gGmbH, Oldenburg, Germany
[3]University of Oldenburg, Medical Physics Department, Oldenburg, Germany
[4]Cluster of Excellence Hearing4All, Oldenburg, Germany

ABSTRACT

The paper describes an automatic speech recognition (ASR) system for the 3rd CHiME challenge that addresses noisy acoustic scenes within public environments. The proposed system includes a multi-channel speech enhancement front-end including a microphone channel failure detection method that is based on cross-comparing the modulation spectra of speech to detect erroneous microphone recordings. The main focus of the submission is the investigation of the amplitude modulation filter bank (AMFB) as a method to extract long-term acoustic cues prior to a Gaussian mixture model (GMM) or deep neural network (DNN) based ASR classifier. It is shown that AMFB features outperform the commonly used frame splicing technique of filter bank features even on a performance optimized ASR challenge system. I.e., temporal analysis of speech by hand-crafted and auditory motivated AMFBs is shown to be more robust compared to a data-driven method based on extracting temporal dynamics with a DNN. Our final ASR system, which additionally includes adaptation of acoustic features to speaker characteristics, achieves an absolute word error rate reduction of approx. 21.53 % relative to the best CHiME-3 baseline system on the "real" test condition.

Index Terms**—** Amplitude modulation filter bank, frame splicing, deep neural network, temporal dynamics, chime challenge

1. INTRODUCTION

The 3rd CHiME speech separation and recognition challenge provides an evaluation platform to test different combinations of front- and back-end technologies for noise robust automatic speech recognition (ASR). The ASR user scenario is aimed at consumer electronics by using a tablet device that is equipped with multiple microphones [1]. Real and simulated audio recordings are generated in various noisy public environments including real-world speech and background noise recordings that pose high requirements to a noise robust ASR system.

Our proposed system is a combination of multi-channel speech enhancement (SE), extraction of long-term acoustic features, speaker adaptation, and deep neural network (DNN) based ASR. Moreover, due to occasional malfunctions of single microphones, a channel failure detector is used to exclude corrupted channels prior to SE and ASR.

The SE algorithm is taken from the CHiME-3 baseline [1], which is a time-varying minimum variance distortionless response (MVDR) beamformer that suppresses sound sources not arriving from the direction of the target speaker [2,3]. The direction of arrival (DOA) is estimated by the steered response power (SRP) source localization method, which is computed from the phase transform (PHAT) [4]. We investigated several modifications of the baseline SE, i.e., replaced the SRP-PHAT with the multiple signal classification (MUSIC) localization algorithm [5], added two different postfilters, i.e., the McCowan postfilter [6] and the single channel enhancement scheme from [7], and even replaced the CHiME-3 baseline SE by a rank 1 speech distortion weighted multi-channel Wiener filter [8]. However, using a “simple” ASR system trained on noisy data only and without adjusting SE parameters to the CHiME data, none of these SE modifications could achieve a significant word error rate (WER) reduction. Thus, mentioned modifications of the baseline SE are not further discussed in the present study.

Long-term acoustic features are extracted based on analyzing temporal amplitude fluctuations in spectral sub-bands by employing the amplitude modulation filter bank (AMFB) [9]. Amplitude modulation (AM) frequencies between 2 and 16 Hz contain the most important acoustic cues for speech understanding and recognition [10,11]. Analysis of low AM frequencies implies a comparatively long temporal context of approx. 300 ms [12,13,9] that is almost three times longer than the commonly used delta and double delta filters [14,9]. While Gaussian mixture model (GMM) and hidden Markov model (HMM) based ASR systems predominantly use short-term features such as Mel-frequency cepstral

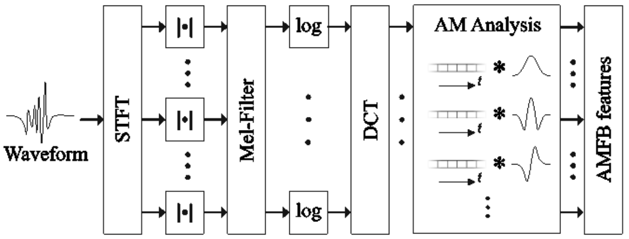

Figure I Signal processing scheme to extract amplitude modulation filter bank features. STFT denotes a short-term Fourier transform.

coefficients (MFCC) plus their temporal derivatives, long-term acoustic features have shown to be advantageous on numerous corpora [9].

In addition to the difficulty in finding a suitable feature transformation to extract the essential AM frequency cues of speech, there are two more reasons for the enduring use of short-term features in GMM-based systems. Firstly, the decomposition of short-term features into AM frequency components increases the feature dimensionality, whereby more attention is required to adjust the acoustic model (AM) or language model (LM) scaling factor and beamwidth, which are used by the Viterbi search [15]. Secondly, high dimensional feature vectors easily show correlations, which is disadvantageous in GMM-based systems with diagonal covariance matrices, and thus may require decorrelation techniques. The recent use of DNNs has overcome these difficulties. The extraction of temporal information is achieved by splicing adjacent feature frames using a context window and feeding them to a DNN feature extractor [16,17]. Thereby, the DNN learns weight matrices that can be interpreted in the first layer as spectro-temporal AM frequency filters.

By comparing GMM and DNN-based ASR systems with different feature inputs, it has been demonstrated that the benefit of DNNs can largely be attributed to an improved temporal processing [18]. Our results in the present contribution show that AMFB features significantly improve ASR performance in both GMM and DNN-based systems. Consequently, the AMFB with its pre-selected and fixed temporal weight pattern for extraction of AM information outperforms a comparable data-driven temporal processing method. In addition, features are adapted to speaker characteristics using the feature space maximum likelihood linear regression (fMLLR) transform [19,20].

2. METHODS

The proposed ASR system consists of three main modules that are SE, AMFB feature extraction, and acoustic modeling. The SE algorithm is based on the CHiME-3 baseline system, cf. Section 3.1. In addition, due to occasional microphone failures in the multi-channel CHiME-3 data, a detector for identification and removal of corrupted microphone channels has been developed.

2.1. Amplitude modulation filter bank feature extraction

The acoustic feature extraction scheme employs the amplitude modulation filter bank (AMFB) to decompose short-term spectral features into AM frequency components [9]. Signal processing steps are depicted in Figure I. The short-term spectral representation $Y_k(l)$ for block l is calculated by applying a discrete Fourier transform (DFT) on speech segments of 25 ms length with a hop size of 10 ms. Speech segments are windowed by the Hann function $w_b(n)$ to minimize the spectral leakage effect.

$$Y_k(l)=\sum_{n=-\infty}^{\infty} y(n)\cdot w_b(n-l)\cdot e^{-\frac{j2\pi kn}{N}},\quad 0\le k\le N-1 \qquad (1)$$

$$w_b(n)=\begin{cases}0.5-0.5\cdot\cos\left(\frac{2\pi n}{b}\right), & 0\le n\le b\\ 0 & \text{otherwise}\end{cases} \qquad (2)$$

In (1) and (2), n, k, b, and N represent the discrete time and frequency indices, the analysis window length, and the DFT length, respectively.

The magnitude of the complex valued spectrum $Y_k(l)$ is passed to the triangular-shaped Mel filters $F_{k,m}$ that integrate DFT bins into M critical spectral bands. Mel-spectral energies are compressed using a logarithmic function, whereby the log-Mel-spectrogram $\hat{Y}_m(l)$ is derived for each Mel band m.

$$\hat{Y}_m(l)=\log\left(\sum_{k=0}^{N-1}\left|Y_k(l)\right|\cdot F_{k,m}\right),\quad 0\le m\le M-1 \qquad (3)$$

Log-Mel-spectral energies are analyzed by a discrete cosine transform (DCT), which leads to the cepstrogram $\tilde{Y}_c(l)$ with C being the DCT length.

$$\tilde{Y}_c(l)=\sum_{m=0}^{M-1}\hat{Y}_m(l)\cdot\cos\left(\tfrac{\pi}{M}\left(m+\tfrac{1}{2}\right)c\right),\quad 0\le c\le C-1 \qquad (4)$$

Temporal dynamics of the cepstrogram are analyzed using the AMFB. The AMFB consists of I complex exponential functions, which are denoted by $q_i(l_0)$, that are windowed by the zero-phase Hann envelope $W_i(l_0)$.

$$q_i(l_0)=e^{-j\Omega_i l_0\cdot T}\cdot W_i(l_0),\quad 0\le i\le I-1 \qquad (5)$$

$$W_i(l_0)=\begin{cases}0.5+0.5\cos\left(\frac{2\pi l_0}{B_i}\right), & -\left\lceil\frac{B_i-1}{2}\right\rceil< l_0<\left\lceil\frac{B_i-1}{2}\right\rceil\\ 0 & \text{otherwise}\end{cases} \qquad (6)$$

$$B_i=\frac{9.06}{2\pi\cdot\beta_i\cdot T} \qquad (7)$$

B_i determines the AM filter length with the sampling period

Table I Center frequency (CF) and bandwidth (BW) parameters of the amplitude modulation filter bank.

i	0	1	2	3	4
CF [Hz]	0	5.5	10.15	15.91	27.03
BW [Hz]	8.25	5.5	6.13	8.27	19.52

T. Ω_i and β_i are the angular AM frequency and the -3 dB AM filter bandwidth, respectively. Convolution of $q_i\,(l_0)$ and $\tilde{Y}_c\,(l_0)$ yields the AM frequency decomposition of the cepstrum.

$$Q_{c,\mathrm{i}}(l) = \left(\tilde{Y}_c * q_i\right)(l) \tag{8}$$

Center frequency (CF) and bandwidth (BW) settings of the employed AM filters are presented in Table I, which are derived by an ASR study on finding optimal AMFB parameters using different ASR corpora. The last step of AMFB feature extraction is the concatenation of real and imaginary AM filter outputs to form a feature vector. Note that the imaginary part of the DC filter is zero, and thus is not taken into account.

2.2. Detection of microphone failures

The multi-channel CHiME-3 data occasionally show dropouts of single microphones that can last from a few milliseconds up to seconds. Short dropouts are perceived as crackling noise. Corrupted recordings of individual channels affect the ASR performance, and thus should be excluded from the succeeding processing modules.

The detection of an erroneous microphone channel is achieved by cross comparing the signals of each channel, whereby outliers are identified. Therefore, we compute the correlation coefficient between channels based on a spectro-temporal signal representation. In order to increase robustness against noise, non-speech AM frequency components are suppressed by using three filters of the AMFB whose CFs are 5.5 Hz, 10.15 Hz, and 15.91 Hz (cf. Table I). Subsequently, AM filter outputs are combined by computing the mean. After estimating all cross-correlation coefficients, the average correlation of each channel compared to all other channels is computed and normalized by the maximum value. If the normalized average correlation of a channel is lower than a threshold of 0.9, then an outlier and error is assumed, respectively. However, this method can result in many false-positive recognitions, e.g., if the signal-to-noise ratio (SNR) considerably varies between channels. Therefore, a pseudo-SNR is estimated for each channel by comparing low and high AM band energies using filters of the AMFB. If the pseudo-SNR of a supposedly malfunctioning channel is amongst the best values, then the corresponding microphone signal is probably not corrupted but contains a lower amount of noise relative to the other channels.

Moreover, the pseudo-SNR is used to select a different channel, if a demanded microphone channel is erroneous.

2.3. Acoustic model training

Two different acoustic model (AM) architectures are employed for ASR. The first is based on GMM-HMMs for estimating the fMLLR transform, which is used to equalize speaker characteristics in feature space, as well as to derive initial state labels to train the second ASR back-end, which is a hybrid DNN-HMM system [21].

2.3.1. GMM-HMM training

Context-dependent acoustic models are learned using maximum likelihood (ML) training of GMM-HMM models with a total maximum number of 30K Gaussian mixture components. Triphone states are clustered based on a decision-tree with 3000 leaves. Speaker independent models are further trained with speaker normalized features that are derived by estimating and applying an fMLLR transform [20]. Finally, GMM-HMMs are discriminatively trained using the minimum phoneme error (MPE) criterion [22].

2.3.2. DNN-HMM training

The training of a 7-layer hybrid DNN system is initialized by stacked restricted Boltzmann machines (RBMs) that are pre-trained in a greedy layer-wise manner [23]. Each DNN-layer has 2047 neurons and sigmoid hidden units except for the final hidden layer, which employs a softmax activation function instead. Mini-batch stochastic gradient descent (SGD) with the cross-entropy objective function is utilized to train the DNN at frame-level [24]. The cross-entropy trained acoustic model is then used to generate lattices and state alignments for a succeeding sequence-discriminative DNN training using the state-level minimum Bayes risk (sMBR) criterion [24]. Note that sMBR training is sequentially performed twice.

3. EXPERIMENTS

All shown word error rates (WERs) are obtained following the CHiME-3 challenge rules and are based on official training and test sets [1]. Section 3.1. provides a description of the CHiME-3 datasets and ASR baseline results. Section 3.2. presents recognition scores of the proposed ASR system.

3.1. CHiME-3: Datasets and baselines

The CHiME-3 scenario is using a multi-channel microphone array for ASR in various noisy environments, i.e., public transport (BUS), café (CAF), street junction (STR), and pedestrian area (PED) [1]. The microphone array has six-channels, which are distributed along the frame of a tablet device. For each of the four noise conditions, real and simulated recordings are generated. The simulated data is built by mixing clean recordings of the Wall Street Journal corpus (WSJ0) [25] with real background noise recordings of

Table II Word error rates of the CHiME-3 baseline ASR systems. Results are obtained with multi condition training and the pruned trigram (tgpr) LM. When indicated, SE is used on both training and test data.

AM	SE	DEV [%]		EVAL [%]		Avg.
		Real	Simu	Real	Simu	[%]
GMM	×	18.70	18.71	33.23	21.59	22.59
GMM	✓	20.55	9.79	37.36	10.59	19.10
DNN	×	16.13	14.30	33.43	21.51	20.68
DNN	✓	17.72	8.17	33.76	11.19	17.20

the different environments. The real data is recorded in each environment by four different native English speakers reading sentences from the WSJ0 corpus. A development (DEV) and evaluation (EVAL) test set is provided for real and simulated recordings consisting of 410 and 330 utterances per environment, respectively. Note that average WERs presented in this paper take the number of test files into account. The training set contains 7138 simulated utterances taken from WSJ0 plus additional 1600 real recordings [1], what is referred to as the "multi-noisy" training set.

Moreover, two baseline ASR systems are provided to train a GMM-HMM and DNN-HMM system using the Kaldi toolkit [21,1]. The GMM-HMM system employs MFCC features with 13 cepstral coefficients but without deltas and double deltas. Raw MFCCs are spliced using a context window of ± 3 frames prior to a linear discriminant analysis (LDA) to reduce the feature dimensionality from 91 to 40. Features are further processed by a maximum likelihood linear transform (MLLT) and speaker adapted training is performed based on the fMLLR method. Training of the baseline DNN-HMM system is similar to the procedure described in Section 2.3.2. 40 Mel-spectral features with a splicing context of 11 continuous frames are used as an input for the DNN. WERs of the two CHiME-3 baseline systems are shown in Table II. The baseline SE algorithm is a MVDR beamformer that uses 400 to 800 ms context immediately before the utterance to estimate the multi-channel covariance matrix of the noise [1]. The steering vector is derived by a SRP-PHAT based localization method, whereby movements of the source are tracked over time using the Viterbi algorithm. Note that the baseline SE algorithm already incorporates a simple microphone failure detection that rejects channels of low signal energy.

Two standard 5k words closed-vocabulary WSJ0 language models (LMs) are provided, i.e., a full trigram model (tg) and a pruned trigram model (tgpr). Baseline results, which are shown in Table II, are obtained using the tgpr LM.

3.2. Results

Table III summarizes the effect of different changes that are incorporated into the proposed ASR system. The starting point is a standard ML trained GMM-HMM system that uses raw MFCC features including deltas and double deltas as input. Note that our GMM-HMM system is trained with twice as many Gaussian mixture components and approx. 25 % more senone targets in comparison to the CHiME-3 baseline GMM-HMM systems, whose WERs are shown in Table II. The ML trained and fMLLR speaker adapted MFCC system of Table III already achieves 2.75 % lower WERs on average compared to the GMM-HMM baseline system without SE. Discriminative training with the MPE criterion further enhances ASR scores by approximately 2 %. Application of the SE algorithm reduces the average WER by another 3.74 %, which is largely due to lower WERs on the simulated test data, whereas results for the real data remain almost unchanged. The channel failure detection method improves WERs by approx. 0.84 %, which can primarily be attributed to improvements with real test data, while results of the simulated test conditions are degraded.

Table III Word error rates of GMM-HMM based ASR systems. MFCCs include 13 cepstral coefficients, which are computed from 31 Mel-channels, as well as delta and double delta features. AMFB features use the same MFCC computation as a basis (without deltas and double deltas) for the succeeding AM sub-band decomposition. The asterisk indicates that 20 cepstral coefficients are employed instead of 13. "Fail Det." denotes usage of the proposed microphone failure detection method. SE and fMLLR indicate the application of speech enhancement and speaker adapted training, respectively. R.I. denotes the relative improvement and "#feats" gives the feature dimensionality.

Feat. Type	#feats	GMM Tr.	fMLLR	SE	Fail Det.	LM	DEV [%]		EVAL [%]		Avg.	R.I. [%]
							Real	Simu.	Real	Simu	[%]	
MFCC	39	ML	×	×	×	tgpr	22.40	22.43	34.00	28.67	26.40	-
MFCC	39	ML	✓	×	×	tgpr	16.35	17.27	25.94	21.27	19.84	25.04
MFCC	39	MPE	✓	×	×	tgpr	14.19	15.87	23.33	19.25	17.82	32.84
MFCC	39	MPE	✓	✓	×	tgpr	14.75	9.02	22.76	10.86	14.08	46.93
MFCC	39	MPE	✓	✓	✓	tgpr	12.80	10.10	18.81	12.12	13.24	49.67
MFCC	39	MPE	✓	✓	✓	tg	12.12	9.48	17.93	11.43	12.53	52.42 / -
AMFB	117	MPE	✓	✓	✓	tg	11.23	8.45	16.04	10.03	11.27	56.99 / 10.01
AMFB*	180	MPE	✓	✓	✓	tg	10.70	8.06	15.26	9.77	10.78	58.82 / 13.75

Table IV Word error rates of DNN-HMM based ASR systems. Employed MFCC features include deltas and double deltas. † and ‡ denote the use of 40 (out of 40) and 25 (out of 31) cepstral coefficients, respectively. "#feats" indicates the number of features that are fed to the DNN.

feature type	fMLLR	Splicing	#feats	DEV [%]		EVAL [%]		Avg. [%]	R.I. [%]
				Real	Simu	Real	Simu		
MFCC†	×	±7	1800	12.66	8.90	20.54	11.36	13.08	-
AMFB†	×	±1	1080	11.56	8.40	18.89	10.13	12.00	8.36
MFCC‡	✓	±7	1125	9.97	6.57	15.21	8.17	9.79	25.28
AMFB‡	✓	±1	675	8.81	6.12	12.98	6.90	8.57	34.34

Note that we do not use speech enhanced training data but the channel failure detection is applied. A closer investigation as to why WERs for simulated test conditions are degraded has identified application of the channel failure detection on the training data as the culprit. A possible reason is the use of varying channel characteristics during training that may be disadvantageous for recognition of simulated data. Nevertheless, the failure detection continues to be used for training and testing in subsequent experiments.

Replacing the pruned trigram by a full LM effects a WER reduction of about 0.71 %. Although the current ASR system is already highly optimized and shows a relative WER improvement of 52.42 % compared to the initial system, AMFB features can further decrease absolute WERs by 1.26 %, corresponding to a relative improvement of 10.01 %. Increasing the number of cepstral coefficients from 13 to 20 lowers absolute WERs of AMFBs by an additional 0.49 %. Finally, all changes applied to the GMM-HMM based ASR system help to increase WERs by approx. 58.82 % relative to the initial standard system.

Table IV presents results of several variants of DNN-based ASR systems that all apply the tg LM, speech enhancement, and microphone failure detection. Two different feature types are employed as input to the DNN. The first are MFCCs including deltas and double-deltas, and the second are AMFB features. MFCCs are spliced with a context window of ±7 frames. AMFB features are spliced using three adjacent frames, which is a favorable setting for AMFBs. If no feature space speaker adaptation is used, MFCC and AMFB features utilize 40 cepstral coefficients, which is a quasi-standard for hybrid DNN systems [17,16]. If the fMLLR transform is used, which is estimated by a preceding decoding pass with a GMM-HMM system, the number cepstral coefficients is set to 25 instead.

The total average WER of the two MFCC-based DNN-HMM systems amount to 13.08 % without and 9.79 % with speaker adaptation, respectively. Thus, the fMLLR speaker adaptation effects a relative WER improvement of approx. 25.28 %, when used in combination with a hybrid DNN back-end. The AMFB as a temporal pre-processing concept prior to a DNN feature extractor reduces absolute WERs on average by more than 1 % compared to both MFCC-based DNN systems with frame splicing. Thereby, the relative improvement amounts to 8.36 % without and to 12.1 % with speaker adapted features, respectively.

Table V Word error rates of the final best performing ASR system sperately for each test condition using the standard CHiME-3 "multi-noisy" training set.

	DEV [%]		EVAL [%]		Avg. [%]
	real	simu	real	simu	
BUS	10.55	5.59	14.83	5.29	8.96
CAF	8.17	7.74	14.83	7.64	9.42
PED	8.13	5.38	11.75	7.19	7.97
STR	8.39	5.77	10.52	7.49	7.94
Avg.	8.81	6.12	12.98	6.90	8.57

Table VI Word error rates of the final ASR system using an extended version of the original "multi-noisy" training set by using each channel (except channel 2) plus a speech enhanced version of the "multi-noisy" training data.

	DEV [%]		EVAL [%]		Avg. [%]
	real	simu	real	simu	
BUS	9.07	4.65	13.50	4.82	7.89
CAF	7.48	7.11	13.52	7.45	8.72
PED	6.90	4.73	10.56	6.37	7.00
STR	7.74	5.49	9.23	7.13	7.31
Avg.	7.80	5.50	11.70	6.44	7.73

Detailed WERs of the final best performing ASR system, i.e., the hybrid DNN with fMLLR transformed AMFB features, are shown in Table V. Results are further enhanced by increasing the amount of training examples. Table VI presents WERs with an extended training set that is build using channel 1, 3, 4, 5, and 6 as well as a SE processed version of the "multi-noisy" training data.

4. DISCUSSION AND CONCLUSIONS

The paper presents an investigation of the amplitude modulation filter bank (AMFB) used to extract long-term information prior to GMM and DNN-based ASR systems. It is demonstrated that the AMFB, which comprises a fixed weight pattern to extract temporal information, can outperform a data-driven method, i.e., the commonly used frame splicing approach, in conjunction with a hybrid DNN system. Since the AMFB design is inspired by psychophysical findings [9], its application in DNN-based system can be motivated in a related way as the use of filter bank features, which mimic acoustic frequency resolution of the human inner ear [26]. Recent investigations on learning neural net (NN) based ASR front-ends directly on raw time signals indicate that the first NN-layer will learn similar weights compared to an auditory-inspired filter bank front-end [27,28]. This provides evidence that auditory-motivated signal processing concepts are well-matched to ASR, which endorses their future use. In addition, learning of an NN-based front-end directly on raw waveforms requires a comparably large amount of training data in order to obtain comparable results to filter bank features [27]. Therefore, the AMFB might be a good candidate to incorporate further auditory-inspired processing concepts to DNN-based ASR systems, by which learning efforts, computational costs, and WERs can be reduced.

The proposed CHiME-3 challenge system, which involves multi-channel SE, AMFB features, fMLLR speaker adaptation, and a DNN-HMM ASR back-end, achieves an average WER reduction of 8.63 % compared to the best baseline system. Thereby, the WER of the real evaluation test condition is reduced by 20.25 % using the original unmodified “multi-noisy” training set. Increasing the training data by considering the official training set in different versions provides a further WER reduction of 1.28 %.

ACKNOWLEDGEMENTS

This work was partly supported by funding from the Federal Ministry of Education and Research (BMBF) and the European Commission (projects KNOTS, AAL-2013-6-144 and KANTATE, V4ARB029).

5. REFERENCES

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

[2] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone arrays*. Springer Berlin Heidelberg, 2001, ch. 2, pp. 19-38.

[3] X. Mestre and M. Á. Lagunas, "On diagonal loading for minimum variance beamformers," in *IEEE International Symposium on Signal Processing and Information Technology*, Darmstadt, 2003, pp. 459-462.

[4] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*. Springer Berlin Heidelberg, 2001, ch. 8, pp. 157-180.

[5] R. Schmidt, "Multiple emitter location and signal paramter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276-280, 1986.

[6] I. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 12, no. 6, pp. 709-716, 2003.

[7] B. Cauchi, et al., "Joint dereverberation on noise reduction using beamforming and a single-channel speech enhancement scheme," in *REVERB Challenge Workshop*, Florence, 2014.

[8] T. V. Bogaert, J. Wouters, S. Doclo, and M. Moonen, "Binaural cue preservation for hearing aids using an interaural transfer function multichannel Wiener filter," in *IEEE Internation Conference on Acoustics, Speech, and Signal Processing*, Honolulu, 2007, pp. IV-565-IV-568.

[9] N. Moritz, J. Anemüller, and B. Kollmeier, "An auditory inspired amplitude modulation filter bank for robust feature extraction in automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1926-1937, 2015.

[10] T. M. Elliot and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS Computer Biology*, vol. 5, no. 3, 2009.

[11] R. V. Shannon, F. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science 270*, pp. 303-304, 1995.

[12] P. Schwarz, "Phoneme recognition based on long temporal context," PhD Thesis, Brno University of Technology, 2009.

[13] N. Moritz, J. Anemüller, and B. Kollmeier, "Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments," in *International Conference on Acoustics, Speech, and Signal Processing*, Prague, 2011, pp. 5492-5495.

[14] S. Furui, "Speaker-independent isolated word recognition based on emphasized spectral dynamics," in *International Converence on Acoustics, Speech, and Signal Processing*, Tokyo, 1986, pp. 1991-1994.

[15] S. Young, et al., *The HTK book (for HTK version 3.4)*. Cambridge University: Cambridge University Engineering Department, 2009.

[16] G. Hinton, et al., "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.

[17] A.-R. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14-22, 2012.

[18] J. Pan, C. Liu, Z. Wang, Y. Hu, and H. Jiang, "Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition: Why DNN surpasses GMMs in acoustic modeling," in *International Symposium on Chinese Spoken Language Processing*, Hong Kong, 2012, pp. 301-305.

[19] Y. Li, H. Erdogan, Y. Gao, and E. Marcheret,

"Incremental on-line feature space MLLR adaptation for telephony speech recognition," in *International Conference on Spoken Language Processing*, Denver, 2002, pp. 1417-1420.

[20] M. J. F. Gales, "Maximum likelihood linear tranformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75-98, 1998.

[21] D. Povey, et al., "The Kaldi speech recognition toolkit," in *IEEE Automatic Speech Recognition and Understanding Workshop* , Hawaii, 2011.

[22] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *International Conference on Acoustics, Speech, and Signal Processing*, Orlando, 2002, pp. 105-108.

[23] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527-1554, 2006.

[24] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Interspeech*, Lyon, 2013, pp. 2345-2349.

[25] D. B. Paul and J. M. Baker, "The design of the wall street journal-based CSR corpus," in *Workshop on Speech and Natural Language*, New York, 1992, pp. 357-362.

[26] E. Zwicker, G. Flottorp, and S. S. Stevens, "Critical Band Width in Loudness Summation," *Journal of the Acoustical Society of America*, vol. 29, no. 548, pp. 548-557, 1957.

[27] Z. Tüske, P. Golik, R. Schlüter, and H. Ney, "Acoustic modeling with deep neural networks using raw time signals for LVCSR," in *Interspeech*, Singapore, 2014, pp. 980-894.

[28] Y. Hoshen, R. J. Weiss, and K. W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *International Conference on Acoustics, Speech, and Signal Processing*, Brisbane, 2015, pp. 4624-4628.