Should deep neural nets have ears? The role of auditory features in deep learning approaches

Angel Mario Castro Martinez[1], *Niko Moritz*[1, 2], *Bernd T. Meyer*[1]

[1] Department für medizinische Physik und Akustik, Exzellenzcluster Hearing4all, Carl von Ossietzky Universität Oldenburg, Germany

[2] Fraunhofer IDMT - Hearing, Speech and Audio Technology, Oldenburg, Germany

angel.castro@uni-oldenburg.de, niko.moritz@idmt.fraunhofer.de, bernd.meyer@uni-oldenburg.de

Abstract

Features inspired by the auditory system have previously demonstrated improvement in automatic speech recognition (ASR). Similarly, the use of Deep Neural Networks (DNN) was found to outperform classic approaches to ASR in many conditions. Since DNNs have the potential to learn the task-relevant features from a conventional filter bank output, we investigate if the combination of auditory features and deep learning should be preferred over self-learned patterns. Specifically, noise-robust Gabor features and Amplitude Modulation Filter-Bank (AMFB) features, highly invariant against reverberation, are used as input to a state-of-the-art ASR system incorporating DNN processing. On the Aurora-4 task, both mel-frequency cepstral coefficients (MFCC) and filter bank (FBank) features are outperformed in many acoustic conditions through auditory processing, yielding average relative improvements of up to 69% over MFCC and 21% over the commonly used DNN-FBank setup. This highlights the mutual benefit of auditory signal processing and recent advances in machine learning.

Index Terms: deep neural network, deep learning, Gabor features, amplitude modulation filter-bank, speech recognition

1. Introduction

Voice controlled systems are becoming ubiquitous in our daily lives. Compared to human speech recognition (HSR) capabilities, however, automatic speech recognition (ASR) still has a long way to go, especially in noisy and reverberant environments. The approaches to increase performance of ASR systems can be divided into two categories that target either feature extraction or the classifier including the acoustic model. Since the human ear is still unmatched in its robustness [1] [2] [3], one branch in feature-related research focuses on copying the principles from the auditory system to machine listening, hence the healthy human ear often serves as a model to improve existing feature extraction methods for ASR.

Recently, improved techniques to train many-layered neural nets have gained much attention in speech research, since deep neural networks (DNNs) were found to outperform conventional combinations of Gaussian mixture models (GMMs) and hidden Markov models (HMMs) [4] [5] [6] [7] [8]. Because the general approach of deep learning is to supply to the classifier as much information as possible (e.g., by providing a simple representation such as a spectrogram), finding optimal feature representations becomes part of the training. Therefore, combining both approaches have rarely been combined so far, one noticeable exception being [9]. The main question of this study is to determine if auditory feature representations and deep learning can be successfully combined, i.e., if the knowledge about our auditory system and its signal processing strategies is helpful, or if a mainly data-driven approach based on self-learned features should be preferred in ASR.

For auditory processing, there is a vast number of studies focused on improving ASR robustness by incorporating psychoacoustic or physiological findings. For instance, Kim and Stern [10] suggested the use of power-normalized cepstral coefficients (PNCC) as an alternative to the conventional mel spectra, using the magnitudes of the squared spectra integrated via a Gammatone filterbank to better approximate the place-frequency mapping of the basilar membrane [11], and also incorporating a non-linear power function that mimics the dependency of the input sound level and the perceived loudness used to compress the output of the Gammatone filterbank. In this study, we focus on features that were developed earlier in our group, that incorporate explicit spectro-temporal processing and temporal modulation filtering:

The first feature type is motivated by physiologic and psychoacoustic studies that have shown the existence of neurons in the primary auditory cortex of mammals [12] [13], specifically tuned to spectro-temporal patterns of time-frequency representations (such as vowel transients in speech). To model such patterns, spectro-temporal 2-D Gabor filters suggested in [12] were subsequently used to exploit spectro-temporal information for robust ASR [14]. A challenge when designing filters for ASR is to determine a set of suitable parameters that result in a robust feature set. Kleinschmidt and Gelbart used a machine learning approach that started with a random set of filters that was iteratively optimized using a simple classifier (a linear neural net) with a small speech recognition task (the recognition of isolated digits). In more recent work, Gabor filters have been organized in a filterbank [15], which resulted in relative improvements of the word error rate (WER) by 30-45% compared to a MFCC baseline for ASR [15] [16], and 21% for speaker identification [17]. However, these features have so far only been combined with standard classifiers in speech research, while, to our knowledge, recent developments in classification have not been yet considered.

The second auditory feature type used in the present study was designed based on the observation that the processing of amplitude fluctuations plays a central role for speech intelligibility and understanding. Conventional feature extraction methods, such as the delta and double delta features, however, only perform limited temporal integration as compared to processing stages found in the auditory cortex of mammals [13]. It has been shown that the human auditory system decomposes an audio signal not only into its acoustic frequencies but also into its amplitude modulation frequency

components. Langner and Schreiner [18], for example, observed a periodotopic arrangement of neurons tuned to certain modulation frequencies in the inferior Colliculus (IC), which were found to be almost orthogonal to the tonotopic arrangement of neurons tuned to certain acoustic frequencies. These finding led to the idea and development of the amplitude modulation filterbank (AMFB) [19]. The AMFB uses complex exponential functions modulated by a Hann window to derive in a set of amplitude modulation filters for feature extraction. The bandwidth and center frequency settings of the AMFB are inspired by a psychoacoustical model proposed in [20]. In conjunction with a conventional hidden Markov model (HMM) classifier, the AMFB features have demonstrated improved robustness against additive noise and room reverberation, as recently shown in the REVERB challenge [21].

These features are used in this work to train and run a DNN-based recognizer; the results are compared to a conventional MFCC-based system as well as a state-of-the-art approach which uses a time-frequency representation from a filterbank as input to a DNN, in order to compare self-learned with engineered feature extraction. The remainder of this paper is structured as follows: Section 2 presents the details of the auditory features, i.e., Gabor and amplitude-modulated spectrogram features, along with the baseline features and the setup of the deep neural network. Results are presented and discussed in Section 3; the paper concludes in Section 4.

2. Methods

2.1 Gabor Filter-bank features

Gabor features are calculated by processing a spectro-temporal representation of the input signal by a number of 2D modulation filters. Filtering is performed by calculating the 2D convolution of the filter and a log-mel spectrogram with 31 frequency channels; Mel spectrograms were chosen because they incorporate several properties of the auditory system (i.e., non-linear frequency scaling and logarithmic compression of amplitude values).

Gabor filters are defined as the product of a complex sinusoidal function $s(n,k)$ (with n and k denoting the time and frequency index, respectively) and a Gaussian envelope function. In this work, the envelope is replaced by the Hann function $h(n,k)$, which was reported earlier to slightly improve results for ASR [22] due to improved filter characteristics compared to a Gaussian envelope with limited extent. In this notation, the complex sinusoid is defined as

$$s(n,k) = \exp[\, i\omega_n(n-n_0) + i\omega_k(k-k_0)\,] \qquad (1)$$

The Hann envelope is given by

$$h(n,k) = 0.5 - 0.5 \cdot \cos\left(\frac{2\pi(n-n_0)}{W_n+1}\right) \cdot \cos\left(\frac{2\pi(k-k_0)}{W_k+1}\right) \qquad (2)$$

The sinusoidal function $s(n,k)$ with window lengths W_n and W_k.

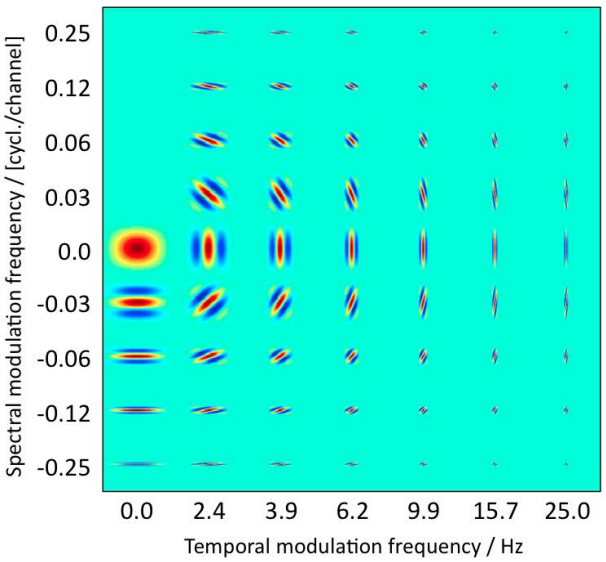

Figure 1: *Real components of Gabor filters used for the filter bank, arranged by temporal and spectral modulation frequencies.*

The periodicity of the carrier function is defined by the radian frequencies ω_n and ω_k, which allow the Gabor function to be tuned to particular directions of spectro-temporal modulation, including diagonal modulations. For purely temporal or spectral filters, this definition results in an infinite support function; in these cases, the support is limited to 69 frequency channels or 99 time frames, which corresponds to the maximum size of the other filters in the respective dimension.

Experiments presented in this paper are based on a spectro-temporal filterbank proposed in [15]. The filterbank contains a set of temporal, spectral and spectro-temporal filters that were chosen to cover a wide range of modulation frequencies. The specific modulation frequencies were chosen so that the transfer functions of the filters exhibit a constant overlap in the modulation frequency domain. While the lowest temporal modulation frequency previously employed was 6 Hz, we use a modified version in which the lowest modulation frequency is 2 Hz, which was included to cover modulations arising from the syllable structure in spoken language. This parameterization results in 59 pairs of spectral and temporal modulation frequencies; the resulting filters are depicted in Fig. 1.

With 59 spectro-temporal filters and 31 frequency channels, the resulting feature vectors are rather high-dimensional with 1829 components. However, since filters with a large spectral extent result in relatively small changes in the feature values when shifted by one frequency channel, the redundancy of the filter output can be reduced by selection of specific feature channels. Hence, for each modulation filter, the center frequency channel (corresponding to a frequency of 1 kHz) is selected; additionally, the channels are included in the final vector for which the overlap of neighboring Gabor filters is 3/4. With this critical sampling, the number of selected channels lies between 1 (for $\omega_k = 0$ cycl. / oct.) and 31 ($\omega_k = +0.25$ cycl./oct.), and the feature dimension is reduced to 657.

2.2 Amplitude Modulated Filter-bank features

The AMFB is used as an ASR feature extraction method that analyzes the temporal dynamics of the speech [23] [24]. The computation of AMFB features is based on the log-mel-spectrogram (for frames of length 25 ms and with 10 ms frame shift). A subsequent discrete cosine transform along the spectral axis leads to the cepstrogram. See Fig.2. The amplitude modulation filters $\boldsymbol{q}$ are complex exponential functions that are modulated by a Hann-envelope, as described in Eq. (3) by the Hadamard product of $\boldsymbol{s}_{carr}$ (cf. Eq. (4)) and $\boldsymbol{h}_{\boldsymbol{env}}$ (cf. Eq. (5)), in which i is the imaginary unit, ℓ_c denotes the frame index of the cepstrogram, and W_{ℓ_c} is the Hann-envelope window length with the center index ℓ_{c0}. Note that beyond the length W_{ℓ_c}, the coefficients of Hann-envelope are set to zeroes in Equations (2) and (5).

$$\boldsymbol{q}[\ell_c] = \boldsymbol{s}_{carr}[\ell_c] \odot \boldsymbol{h}_{\boldsymbol{env}}[\ell_c] \tag{3}$$

$$\boldsymbol{s}_{carr}[\ell_c] = \exp[i\omega(\ell_c - \ell_{c0})] \tag{4}$$

$$\boldsymbol{h}_{\boldsymbol{env}}[\ell_c] = 0.5 - 0.5 \cdot \cos\left(\frac{2\pi(\ell_c - \ell_{c0})}{W_{\ell_c}+1}\right) \tag{5}$$

The periodicity of the sinusoidal-carrier function is defined by the radian frequency ω. By varying ω and W_{ℓ_c}, the AMFB can be tuned to cover different temporal amplitude modulation frequencies with different bandwidths. For the AMFB feature extraction we selected five AM filters, whose center frequencies and bandwidth settings are chosen according to the psycho-physically motivated AM filter-bank proposed in [20], which has a constant bandwidth of 5 Hz for AM frequencies up to 10 Hz and a constant-Q relationship with value of 2 for higher modulation frequencies; producing a 117 dimensional feature vectors.

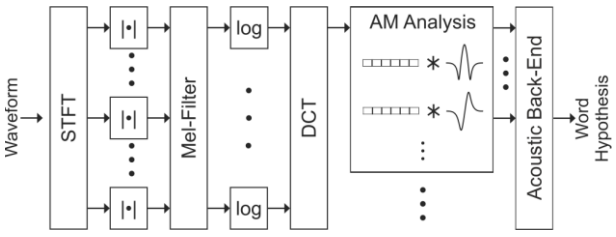

Figure 2: *Block diagram of signal processing steps for the computation of AMFB features.*

2.3 Baseline MFCC and Filter-bank features

The Aurora-4 task baseline is GMM/HMM system based on mel-frequency cepstral coefficients (MFCC) as described in [25], plus their delta and double-delta coefficients combined with cepstral mean variance normalization (CMVN). Even though MFCC plus double deltas are established as quasi-standard in ASR, an additional input type was analyzed to investigate the representations automatically learned by a deep neural network (DNN). As a highly general representation, 40-dimensional mel-scale filter bank features, spliced across 9 adjacent frames [26] (to give temporal context of +/- 4 frames) followed by an LDA to reduce the dimension to the original 40 coefficients, were used to train a DNN as a second baseline [4]. Owing to the spectrogram resemblance of filterbank representations, a DNN is less constrained to create any structure of the input data; thus we were able to contrast "hand-crafted" features (which were designed based on physiological evidence as outlined in the introduction) with DNN-learnt features and assess the role of AMFB and Gabor in deep learning.

2.4 DNN Recipe

The deep neural network (DNN) system implemented is based on the one described in [27]. Owing to the capability of layer-wise pre-training using Restricted Boltzmann Machines (RBM) [28] and optimization via stochastic gradient descent (SGD) on the graphics processing unit (GPU), we opted for this particular Kaldi ASR toolkit DNN implementation [29] which can be summarized in two phases (excluding the so called sequence-discriminative training): Pre-training and Cross entropy tuning. On the former phase, a 7-layer deep belief network (DBN) is trained one layer at the time as a stack of RBM using Contrastive Divergence algorithm; prior to pre-training, an auxiliary GMM triphone system is trained to provide the alignment of context dependent states to frames. On the latter phase the DBN is fine-tuned to classify frames into triphone-states using back-propagation via SGD, which results in a 7-layer DNN. Every hidden layer in our DNN has 2048 sigmoid neurons. A soft-max output layer to provide posterior probabilities of each context-dependent HMM state. DNN trained was performed by the GPU NVIDIA Tesla K20C using the multi-condition 16 kHz sampling frequency set of the Aurora-4 corpus, which took approximately 4 days for each feature type. To avoid information loss due to feature dimensionality, experiments were conducted without speaker adaptation transformation (e.g. fMLLR).

3. Results and Discussion

The results for different combinations of feature extraction methods and ASR back-end settings are presented in Table 1. Baseline word error rates (WER) from [30] are compared to our system combining conventional short-term features or auditory features and deep learning. The rows represent the 14 different Aurora-4 test conditions, with SNRs ranging from 10-20 dB for the noisy environments. "Mic 1" refers to a high-quality, close-talk Sennheiser microphone, while "Mic 2" represents one of 18 secondary microphones.

Filterbank features are used as a reference for DNN processing. All DNN-based systems clearly outperform the MFCC recognizer based on a GMM/HMM architecture throughout all conditions. In three conditions, filterbank features processed with the DNN (referred to as DNN (Filterbank)) exhibit the lowest WER among all experimental conditions. However, on average both auditory feature types result in ASR scores superior to the Filterbank features. On average, the lowest WER is achieved by AMFB features combined with a DNN, which is referred to as DNN(AMFB). Both DNN(Gabor) and DNN(AMFB) are significantly more robust than the baseline or the DNN(Filterbank) reference in the presence of channel distortions, i.e. if training and test transfer functions differ due to different microphone

characteristics (rows 8-14), which are likely to occur in real-life scenarios. This finding can also be seen in Table 2, in which the results of the 14 different test sets are summarized in four groups: A) and B) correspond to the clean and noisy conditions for "Mic 1", respectively, and likewise C) and D) for "Mic 2".

Table 1: *Word error rate for the Aurora-4 task for the baseline system, as well for conventional filter bank and auditory features processed with deep neural networks.*

		MFCC baseline	DNN(Filter Bank)	DNN(Gabor)	DNN(AMFB)
Mic 1 (close-talk)	Clean	19.1	**3.9**	**3.9**	4.4
	Car	23.4	**5.0**	6.2	5.4
	Babble	31.7	8.0	8.3	**7.5**
	Restaurant	35.5	10.7	10.3	**9.6**
	Street	35.3	10.1	9.0	**8.7**
	Airport	33.1	**7.4**	7.9	7.4
	Train Station	36.4	10.3	**9.1**	9.1
Mic 2	Clean	40.9	14.4	8.8	**7.9**
	Car	47.4	19.3	13.1	**10.6**
	Babble	50.3	23.9	20.3	**19.8**
	Restaurant	48.9	27.4	22.8	**21.5**
	Street	54.7	27.4	**19.3**	19.5
	Airport	49.3	23.4	19.8	**19.3**
	Train Station	51.8	27.9	**19.9**	21.5
	Average	39.8	15.7	12.8	**12.3**

Table 2: *Summary of the WERs presented in Table 1.*

	MFCC baseline	DNN(Filter Bank)	DNN(Gabor)	DNN(AMFB)
A	19.1	**3.9**	**3.9**	4.4
B	32.6	8.6	8.5	**8.0**
C	40.9	14.4	8.8	**7.9**
D	49.0	23.4	17.7	**17.1**

Finally in Table 3, the relative improvements of DNN(Gabor) and DNN(AMFB) for all configurations are presented, first against the MFCC (GMM-based) baseline and secondly against the DNN(Filterbank) system. While the average improvement for DNN(Filterbank) features over the baseline system is already high (61% on average), further improvements are achieved by performing a feature extraction with Gabors (18.5% compared to DNN(Filterbank)).

Table 3: *Relative WER improvement over MFCC baseline (a) and filterbank features processed with a deep neural net (b).*

Improvement over		a) MFCC baseline			b) DNN(FB)	
		DNN(FB)	DNN(Gabor)	DNN(AMFB)	DNN(Gabor)	DNN(AMFB)
Mic 1 (close-talk)	Clean	79.4	**79.5**	77.0	**0.5**	-11.4
	Car	**78.8**	73.7	76.8	-23.9	-9.1
	Babble	74.8	73.9	**76.3**	-3.5	**5.9**
	Restaurant	69.9	70.9	**72.9**	3.5	**10.1**
	Street	71.4	74.5	**75.2**	10.9	**13.4**
	Airport	**77.7**	76.1	77.6	-7.1	-0.5
	Train Station	71.8	**75.1**	75.0	**11.7**	11.3
Mic 2	Clean	64.8	78.4	**80.8**	38.8	**45.5**
	Car	59.4	72.3	**77.7**	31.9	**45.2**
	Babble	52.4	59.7	**60.7**	15.3	**17.3**
	Restaurant	43.9	53.5	**56.0**	17.1	**21.5**
	Street	49.9	**64.8**	64.4	**29.7**	29.1
	Airport	52.6	59.8	**60.9**	15.2	**17.5**
	Train Station	46.1	**61.5**	58.6	**28.6**	23.2
	Average	60.7	68.0	**69.2**	18.5	**21.4**

4. Summary and Conclusions

This paper investigated the question if two important techniques in ASR can be effectively combined: Features that are inspired by the human auditory system on the one hand, and neural networks that make use of recent progress in the field of deep learning on the other. While neural nets have the capability to self-learn feature representations relevant for recognition, it was unclear if a specific aspect of the output signal can be perceived by the human auditory system. Two different feature types that were developed in our group were tested: Amplitude modulation filterbank (AMFB) features that explicitly encode temporal modulations that have been found to be relevant for speech recognition, and spectro-temporal Gabor features that take diagonal structures in a time-frequency representation of a signal into account. On the Aurora-4 ASR task, both feature types achieve substantial improvements over the MFCC baseline (with an average improvement of 68-69%) when combined with a deep neural net. Further, relative WER reductions of 19-21% are obtained compared to filterbank features with DNNs, which constitute a state-of-the art baseline. It therefore appears that the relation of auditory processing and deep learning is fruitful and should be further explored. One direction for future research is to exploit complementary information of auditory and classic features, by combining the local filtering in the time-frequency plane as supplied by Gabor filters and the longer, global signal analysis for temporal modulation filtering as performed by AMFB features.

5. Acknowledgements

This work was funded by the Cluster of Excellence Hearing4All (http://hearing4all.eu).

6. References

[1] R. Lippmann, "Speech recognition by machines and humans," *Speech Commun.,* vol. 22, no. 1, pp. 1-15, 1997.

[2] O. Scharenborg, "Reaching over the gap: A review of efforts to link human and automatic speech recognition research," *Speech Commun.,* pp. 336-347, 2007.

[3] B. Meyer and B. Kollmeier, "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition," *Speech Commun.,* vol. 53, pp. 753-767, 2011.

[4] A. R. Mohamed, G. Hinton and G. Penn, "Understanding how deep belief networks perform acoustic modelling," in *Proc. ICASSP*, 2012.

[5] G. Dahl, D. Yu, L. Deng and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. Audio, Speech, and Language Processing," *IEEE Transactions,* vol. 20, no. 1, pp. 30-42, 2012.

[6] A. R. Mohamed, T. N. Sainath, G. Dahl, B. Ramabhadran, G. E. Hinton and M. A. Picheny, "Deep belief networks using discriminative features for phone recognition," in *Proc. ICASSP*, 2011.

[7] T. N. Sainath, B. Kingsbury, B. Ramabhadran, P. Fousek, P. Novak and A. R. Mohamed, "Making deep belief networks effective for large vocabulary continuous speech recognition," in *IEEE Workshop ASRU*, 2011.

[8] H. Hermansky, D. Ellis and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. Interspeech*, 2000.

[9] H. Hermansky, D. Ellis and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. Interspeech*, 2000.

[10] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *Proc. Interspeech*, 2009.

[11] R. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang and M. Allerhand, "Complex sounds and auditory images," in *Auditory physiology and perception, Proc. 9th International Symposium on Hearing*, 1992.

[12] A. Qiu, C. Schreiner and M. Escabi, "Gabor analysis of auditory mid- brain receptive fields: spectro-temporal and binaural composition," *Journal of Neurophysiology,* vol. 90, pp. 456-476, 2003.

[13] N. Mesgarani, D. Stephen and S. Shamma, "Representation of phonemes in primary auditory cortex: how the brain analyzes speech," in *Proc. ICASSP*, 2007.

[14] M. Kleinschmidt and D. and Gelbart, "Improving word accuracy with Gabor feature extraction,," in *Proc. Interspeech*, 2002.

[15] M. R. Schädler, B. Kollmeier and B. T. Meyer, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *J. Acoust. Soc. Am.,* pp. 4134-4151, 2011.

[16] B. Meyer, C. Spille, B. Kollmeier and N. Morgan, "Hooking up spectro-temporal filters with auditory-inspired representations for robust automatic speech recognition," in *Proc. Interspeech*, 2012.

[17] H. Lei, B. Meyer and N. Mirghafori, "Spectro-temporal Gabor features for speaker recognition," in *Proc. ICASSP*, 2012.

[18] G. Langner and C. Schreiner, "Periodicity coding in the inferior Colliculus of the cat. I. Neuronal mechanisms," *J. of Neurophysiology,* vol. 60, pp. 1799-1822, 1988.

[19] B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," *J. Acoust. Soc. Am.,* vol. 95, no. 3, pp. 1593-1602, 1994.

[20] T. Dau, B. Kollmeier and A. Kohlrausch, "Modeling Auditory Processing of Amplitude Modulation. I. Detection and Masking with Narrow-Band Carriers," *J. Acoustic Soc. Am.,* vol. 102, no. 5, p. 2892–2905, 1997.

[21] F. Xiong, N. Moritz, R. Rehr, J. Anemüller, B. Meyer, T. Gerkmann, S. Doclo and S. Goetze, "Robust ASR in Reverberant Environments Using Temporal Cepstrum Smoothing for Speech Enhancement and an Amplitude Modulation Filterbank for Feature Extraction," in *Proc. REVERB Workshop*, 2014.

[22] B. Meyer, T. Brand and B. Kollmeier, "Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes," *J. Acoust. Soc. Am.,* vol. 129, pp. 388-403, 2011.

[23] N. Moritz, J. Anemüller and B. Kollmeier, "Amplitude Modulation Spectrogram based Features for Robust Speech Recognition in Noisy and Reverberant Environments," in *Proc. ICASSP*, 2011.

[24] N. Moritz, J. Anemüller and B. Kollmeier, "Amplitude Modulation Filters as Feature Sets for Robust ASR: Constant Absolute or Relative Bandwidth?," in *Proc. Interspeech*, Portland, USA, 2012.

[25] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 28, no. 4, pp. 357-366., 1980.

[26] S. Rath, D. Povey, K. Veselý and J. Cernocký, "Improved feature processing for deep neural networks," in *Interspeech*, 2013.

[27] K. Veselý, A. Ghoshal, L. Burget and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. Interspeech*, 2013.

[28] G. Hinton, "A practical guide to training restricted Boltzmann machines," *Momentum,* vol. 9, no. 1, p. 926, 2010.

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel und K. Veselý, „The Kaldi speech recognition toolkit,“ in *Proc. ASRU*, 2011.

[30] N. Parihar and J. Picone, *Aurora working group: DSR front end LVCSR evaluation AU/384/02,* Inst. for Signal and Information Process, Mississippi State University, Technical Report, 2002.