



Investigating Modulation Spectrogram Features for Deep Neural Network-based Automatic Speech Recognition

Deepak Baby and Hugo Van hamme

Department ESAT, KU Leuven, Belgium

{Deepak.Baby, Hugo.Vanhamme}@esat.kuleuven.be

Abstract

Deep neural network (DNN) based acoustic modelling has been shown to yield significant improvements over Gaussian Mixture Models (GMM) for a variety of automatic speech recognition (ASR) tasks. In addition, it is also becoming popular to use rich speech representations, such as full-resolution spectrograms and perceptually motivated features, as input to the DNNs as they are less sensitive to the increase in the input dimensionality. In this work, we evaluate the performance of a DNN trained on the perceptually motivated modulation envelope spectrogram features that model the temporal amplitude modulations within sub-band speech signals. The proposed approach is shown to outperform DNNs trained on a variety of conventional features such as Mel, PLP and STFT features on both TIMIT phone recognition and the AURORA-4 word recognition tasks. It is also shown that the approach outperforms a sophisticated auditory model based on Gabor filter bank features on TIMIT and the channel matched conditions of the AURORA-4 database.

Index Terms: deep neural networks, automatic speech recognition, modulation envelopes

1. Introduction

Gaussian Mixture Model (GMM) -based hidden Markov models (HMMs) have traditionally been the state-of-the-art in the field of automatic speech recognition (ASR) technology. Recent advances in deep neural network based approaches have shown significant performance improvements over the GMM based approaches on a variety of ASR tasks [1–3], thanks to its multiple hidden layers learning rich multiple projections. It is also shown to be robust to various kinds of distortions when compared to the GMMs [4,5], sometimes improving the performance by large margins.

However, the DNN performance is still far from that of humans especially in noisy environments. Therefore, there is still a growing interest in feature-related research that focuses on applying our knowledge about human auditory processing into this framework. Traditionally, GMMs needed uncorrelated observations due to its diagonal covariance design and this forced most of these attempts to make use of a feature decorrelation step in the end. On the other hand, it is shown that DNNs are less sensitive to the increase in input dimensionality and correlation between features. In particular, Mel-filter bank outputs are shown to yield better performance than the conventional lower dimensional features such as MFCC or PLP coefficients [3,6].

This work has been funded with support from the European Commission under Contract FP7-PEOPLE-2011-290000 (INSPIRE).

The author would like to thank Bernd T. Meyer of University of Oldenburg for valuable discussions on the GBFB feature extraction.

This allows us to use richer, physiologically motivated features to train the DNNs and aim a better cross-fertilization between the human speech recognition (HSR) and the ASR communities.

There exist some studies that evaluate the performance of DNNs trained on physiologically motivated features. Most of these analyses take into account the poorer frequency resolution of the basilar membrane and the role of spectral and temporal modulations in human hearing. In [7], a comparison of various features such as Gammatone filter coefficients, damped oscillator coefficients etc. extracted from the time domain signal without explicitly going to the frequency domain are presented. Another approach is to extract the various spectro-temporal modulation patterns from the log-compressed Mel spectrogram. An investigation based on the Gabor filter analysis and amplitude modulation filter-banks are presented in [8]. Most of these features were found to yield better performance over the Mel-filter bank features.

In this work, we investigate the performance of an auditory model which relies on the amplitude modulations within frequency bands [9]. These are computationally modelled as modulation envelopes that capture the amplitude envelope of the half-wave rectified sub-band speech signals [10]. These features have been successfully used for noise robust speech recognition [11] and phone classification [12]. Since the human speech contains very low modulation frequencies of the order of 20-30 Hz [13], a low-pass filter with a cut-off frequency of around 30 Hz is employed to capture the speech information.

The low-pass filtering used to obtain the modulation envelopes has two benefits; One, it helps in getting rid of the added noise containing higher modulation frequencies. Two, it yields a compact representation of speech in the spectral domain. Therefore, the spectrograms of these envelopes are taken and are truncated to the lowest few significant bins that fall below the 3 dB cut-off frequency of the low-pass filter used. This representation of modulation envelopes in the spectral domain are referred to as modulation spectrogram (MS) features. In our previous works, these features have been successfully used for exemplar-based speech enhancement as a front-end for DNN-based ASR [14].

In this work, the MS features are used to train and evaluate a DNN-based recognizer and the results are compared with the traditional Mel, STFT and PLP features on TIMIT and AURORA-4 databases. We also include a comparison with the Gabor filter bank features investigated in [15]. The rest of the paper is as follows: Section 2 details the MS feature extraction and other baseline features together with the DNN architecture used for evaluation. Section 3 details the evaluation setup followed by the results and discussion in Section 4. Section 5 concludes the paper along with some suggestions for future work.

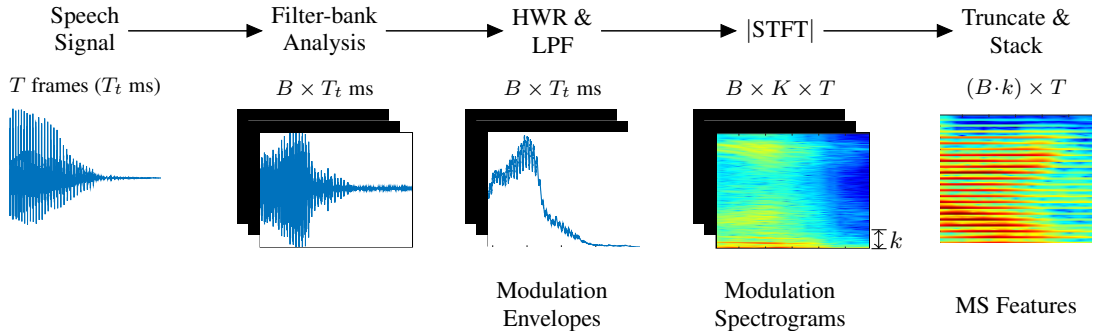


Figure 1: Block diagram overview of the processing steps to obtain the proposed MS features. The corresponding sizes of each of the representation are also shown.

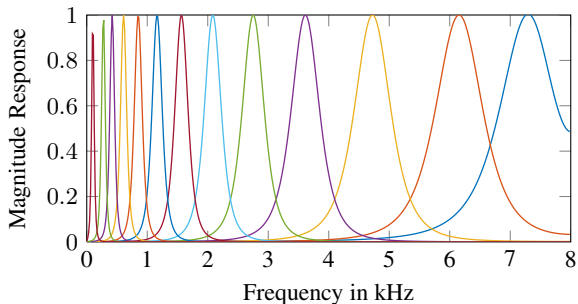


Figure 2: Frequency response of the equivalent rectangular bandwidth filters used to model the basilar membrane.

2. Methods

2.1. MS features

The MS representation was proposed as part of a computational model for human hearing which relies on the low frequency amplitude modulations within various frequency bands [9]. The processing chain used to obtain the MS features is depicted in Figure 1.

To obtain the MS features, the input speech signal is first filtered using a filter-bank having B channels to model the poor frequency resolution of the basilar membrane. This is implemented using an equivalent rectangular bandwidth (ERB) filter bank whose center frequencies are equally spaced along the log-frequency axis that also model the non-linear frequency resolution property of cochlea as defined in [16]. In this work we used ERB filter bank implemented using Gammatone filters [17]. The frequency response of these filters are shown in Figure 2. The resulting B band-limited signals are half-wave rectified to model non-negative nerve firings. The modulation envelopes are obtained by low-pass filtering these rectified sub-band signals at a 3 dB cut-off frequency of around 30 Hz, since human speech contains very small amplitude modulations.

From these envelopes which contain only low frequency signals, the modulation spectrograms are obtained by taking the magnitude STFT, resulting in B modulation spectrograms [10] of size $K \times T$ each, where K is the number of modulation frequency bins used to obtain the STFT and T is the number of frames in the signal. As there is a low-pass filtering operation, it is possible to truncate each of these modulation spectrograms to their lowest few, say k , bins [18,19], i.e., each modulation spectrogram now has size $k \times T$. Only the positive half of the magnitude modulation spectrogram is considered. To obtain a compact two-dimensional representation, we stack these modu-

lation spectrograms originating from B channels to a matrix of size $(B \cdot k) \times T$. These are then log compressed to model the non-linear intensity to loudness variation of the ear. These are referred to as the MS features. Notice that k denotes the number of amplitude modulation frequencies within each frequency sub-band that are used in the model.

The dimensionality of the MS features depends on the value of K which is approximately equal to the window-length used during the STFT step, the sampling frequency f_s and the 3 dB cut-off frequency of the low-pass filter f_{3dB} used to obtain the modulation envelope. The value of k thus will be roughly $\geq f_{3dB} \cdot K / f_s$. i.e., a higher value of K and/or k can be used to capture more temporal amplitude modulation frequencies.

2.2. Baseline features

In this work, we compare the proposed set of features with the conventional Mel, short-time Fourier transform (STFT) and the perceptual linear prediction (PLP) features. We also include the comparison with another physiologically inspired features using Gabor filters, dubbed GBFB features [15]. GBFB features are computed by processing the log-Mel spectrogram with 31 frequency channels by a number of 2D modulation filters. In this setup, the 2D Gabor filters are defined as the product of a complex sinusoidal function and a Hann envelope function, such that they cover a wide range of spectro-temporal modulation patterns [15]. With the setting described in [8], 59 spectro-temporal filters are used per Mel channel which resulted in a total of 1829 components. These are then reduced to 657 features per frame by removing redundant features. For further details, we refer the reader to [8,15].

2.3. DNN Decoder

The evaluations are done using the ‘‘recipe’’ DNN-HMM-based recognizer in the Kaldi toolkit [20]. A DNN is simply a multi-layer perceptron with multiple hidden layers between its inputs and outputs. Performing back-propagation training on such a network can result in a poor local optimum with a randomly initialized network weights. To circumvent this, a pre-training is done first by considering each pair of adjacent layers as restricted Boltzmann machines (RBM) [21] and then a back propagation training is done over the entire network such that it provides posterior probability estimates for the HMM states [22]. All DNNs used are comprised of 6 hidden layers with 2 048 sigmoid neurons per layer. The input layer used a temporal context of 11 frames.

To perform ASR using a DNN-HMM-hybrid setting, the state emission likelihoods generated by the GMMs are replaced

Setting	K	k	Mod. freqs. taken (Hz)	Size
$MS_{1024;5}$	1024	5	0, 15, 30, 45, 60	200
$MS_{1024;3}$	1024	3	0, 15, 30	120
$MS_{2048;5}$	2048	5	0, 7.5, 15, 22.5, 30	200

Table 1: Summary of the MS settings evaluated along with the modulation frequencies considered and the number of features per frame.

by the pseudo-likelihoods or scaled-likelihoods generated by the DNN.

3. Evaluation Setup

3.1. TIMIT database

TIMIT is a benchmark database for evaluating and comparing the phone recognition accuracy of various ASR systems in clean conditions. The training set of the database contains 3 696 utterances recorded from 462 speakers with 8 utterances per speaker. For evaluation, we used the core test set, which contains 192 utterances with 8 sentences each from 24 speakers. The development set of the database contains 400 utterances from 50 speakers. The phone error rates (PER) in % are reported for evaluations on the TIMIT database.

3.2. AURORA-4 database

AURORA-4 database is a large vocabulary continuous speech recognition database based on the WSJ0 corpus of read English speech. It contains six additive noise versions with channel matched and mismatched conditions. The multicondition training data containing 7 138 utterances with channel variations and added noise is used for training the DNNs. The training data contains all the six noise types added at varying SNRs between 10 to 20 dB in steps of 1 dB.

The test set of the database contains 14 sets (test01-test14), each containing 330 utterances. Test01 (or test A) contains the clean utterances recorded with a single microphone and test02-test07 sets (or collectively test B) contain its noisy versions added with the six noise types at varying SNRs between 5 to 15 dB in steps of 1 dB. Test08 (or test C) contains the clean utterances recorded with multiple microphones and test09-test14 (or collectively test C) sets contain its noisy versions same as in test B. A development set of the same structure as of the test set is provided, but with a different set of 330 utterances, for parameter tuning and cross-validation. Word error rates (WER) in % is used to compare the various systems evaluated on this database.

3.3. Feature extraction

All the testing and training data are first pre-processed using a DC removal filter and a pre-emphasis filter of coefficient 0.97 before extracting the features. The STFT features are obtained by taking the STFT of the signal with a window length of 25 ms and a window shift of 10 ms with $F = 512$ bins. The absolute values of the positive half of the STFT is taken to obtain STFT features of size 256 per frame. To obtain the Mel features, the STFT features are Mel integrated with $B = 40$ channels resulting in 40 Mel features per frame. The log compressed Mel and STFT features are used to train the DNNs as they were found to yield better results than the raw format. The PLP features are extracted using the Kaldi feature extraction script with 40 Mel channels and 13 PLP coefficients per frame are computed.

Features	PER in %
<i>Mel</i>	21.5
<i>Mel</i> _{splice7}	21.8
<i>STFT</i>	22.1
<i>PLP</i>	21.6
<i>GBFB</i>	21.0
$MS_{1024;5}$	19.6
$MS_{1024;3}$	19.8
$MS_{2048;5}$	20.6

Table 2: Average PER in % obtained for the TIMIT speaker-independent phone recognition task with DNNs trained on various input features.

The GBFB features are extracted using the code provided in [8] which yielded 657 Gabor features per frame.

To obtain the MS features, equivalent rectangular bandwidth filter bank containing $B = 40$ channels, implemented using Slaney’s toolbox [23], is used to obtain the sub-band signals. These are then half-wave rectified and low-pass filtered at a 3 dB cut-off frequency of 30 Hz to obtain the modulation envelopes within each sub-band. Then an analysis is made for various choices of K , which decides the resolution of the modulation frequencies used, and k which decides the set of amplitude modulation frequencies considered. Two choices of K are used; a window length of 64 ms with $K = 1024$, and a window length of 128 ms with $K = 2048$. The resolution of the modulation spectra will be roughly 15 Hz and 7.5 Hz with $K = 1024$ and 2048, respectively. The evaluations are then made for various choices of k . The settings evaluated are summarised in Table 1.

Since the alignments used for the DNN training are taken from a GMM-based back-end which used a shorter window length (25 ms) than the ones used by the MS, there will be a state-frame misalignment when MS features are used. It is found that the MS features with window length 64 ms and 128 ms lead the GMM features by 2 and 4 frames respectively and the alignments are corrected by delaying the MS features by the respective number of frames. Also notice that the MS features take into account a temporal context of 165 ms when 11 consecutive frames are used for DNN training, whereas all the baseline features span only 115 ms context. For a fair comparison, we also include another baseline system based on Mel features which uses a temporal context of 15 frames (splice = 7) which adds to 165 ms context (denoted as *Mel*_{splice7}).

4. Results and Discussion

4.1. Results on TIMIT database

The PER results obtained for various settings on the TIMIT database are presented in Table 2. Notice that no speaker adaptation is done on any of these features. It is observed that using a splice of 7 frames with Mel features is found to be performing worse than the splice equal to 5 setting. It can be seen that both the perceptually motivated models (GBFB and MS) outperform the traditional features and the MS features yield the best result with a phone recognition accuracy of more than 80 % on the TIMIT database. Given the splice 5 vs. splice 7 comparison with the Mel features, this improvement cannot be attributed to a longer temporal context used by the MS features.

It can also be seen that including more modulation frequencies ($MS_{1024;5}$ vs. $MS_{1024;3}$) indeed can benefit the PER per-

Features	Mic 1							Mic 2							Avg.
	01	02	03	04	05	06	07	08	09	10	11	12	13	14	
<i>Mel</i>	3.5	4.3	6.7	9.7	8.2	6.6	8.7	10.3	14.5	20.9	24.6	24.4	20.1	24.4	13.3
<i>Mel</i> _{splice7}	3.3	4.5	7.1	9.8	8.3	6.7	8.6	9.8	14.1	20.5	24.6	23.0	19.6	24.4	13.2
<i>STFT</i>	3.3	4.6	6.9	9.6	8.9	6.7	8.9	10.8	14.9	21.3	25.6	24.5	20.5	25.2	13.7
<i>PLP</i>	3.9	5.3	7.8	10.7	9.8	7.8	10.0	10.4	14.9	22.6	25.9	26.0	21.4	25.8	14.5
<i>GBFB</i>	3.5	4.6	6.7	9.1	8.0	6.8	8.2	8.0	12.9	18.8	23.2	20.7	18.1	20.4	12.1
<i>MS</i> _{1024;5}	2.6	4.0	6.6	8.8	8.4	6.4	8.7	8.7	11.8	20.0	23.7	21.5	18.0	22.0	12.2
<i>MS</i> _{1024;3}	2.8	3.9	6.5	8.8	7.7	6.7	8.0	10.0	14.6	20.7	23.6	22.6	18.7	23.1	12.7

Table 3: Average WERs in % obtained on each test set of the AURORA-4 database for DNNs trained on various input features.

Features	test A	test B	test C	test D	Avg.
<i>Mel</i>	3.5	7.4	10.3	21.5	13.3
<i>Mel</i> _{splice7}	3.3	7.5	9.8	21.0	13.2
<i>STFT</i>	3.3	7.6	10.8	22.0	13.7
<i>PLP</i>	3.9	8.6	10.4	22.8	14.5
<i>GBFB</i>	3.5	7.3	8.0	19.0	12.1
<i>MS</i> _{1024;5}	2.6	7.1	8.7	19.5	12.2
<i>MS</i> _{1024;3}	2.8	6.9	10.0	20.5	12.7

Table 4: Summary of results on the AURORA-4 database with DNNs trained on various input features.

formance. It is also seen that increasing the modulation spectral resolution by increasing K could be detrimental (ref. $MS_{2048;5}$) mainly because of its too long temporal context (11 frames constitute to 218 ms) which could cover multiple phones at a time and may result in a poorer classifier.

When compared to the GBFB features, MS features gave an absolute PER improvement of 1.4 % (7% relative). This is in fact one of the best results reported on the TIMIT database with the given DNN architecture without any speaker adaptive training.

4.2. Results on AURORA-4 database

Next, the noise robustness of the features are evaluated on the AURORA-4 database. The WERs obtained on each of the test sets in the AURORA-4 for various input features are detailed in Table 3. It can be seen that both GBFB and MS features yield a better robustness to both channel variation and noisy conditions over the Mel, STFT and PLP features. The $MS_{2048;5}$ is not evaluated as it gave poorer performance on the TIMIT database. MS features yielded the best performance on the single microphone cases. In particular, a significant WER improvement even on clean speech is obtained which is even better than the results obtained for the same DNN setting trained on Mel features extracted from the clean training data of the database (2.9 % reported in [14]).

The summary of WERs obtained on various test sets are presented in Table 4. It can also be seen that including more modulation frequencies improves the performance in channel mismatched conditions (ref. test C and test D results of $MS_{1024;5}$ vs. $MS_{1024;3}$). For multiple microphone cases GBFB features performed better because of its sophisticated design in which the features are chosen such that they exhibit robustness to channel variations and noisy conditions. However, no such adaptation is done for the MS feature extraction. Also notice that the MS features use fewer features per frame when compared to the GBFB features. These results reaffirm the effec-

tiveness of combining perceptually motivated rich features as inputs to the DNNs.

Additional experiments were also conducted by concatenating the Mel features with the MS features. However, the evaluations using these concatenated features (not shown) yielded more or less similar results as the MS features. It implies that the information provided to the DNN by the MS and Mel features are not complementary in general and no additional information is introduced by the Mel features.

5. Conclusions

In this paper, we evaluated the performance of the perceptually motivated modulation spectrogram features as input features to DNNs. The approach yielded a PER of 19.6 % on the TIMIT database which is among the best results published on the database without any speaker adaptive training. Further, the noise robustness of these features are evaluated and compared on the AURORA-4 database and it is shown that MS features yield robust performance in all cases when compared to the Mel, STFT and PLP features. When compared to the GBFB features, MS features gave a better performance on single microphone cases. These results reaffirm that DNNs can be effectively combined with perceptually motivated features to bridge the gap between the ASR and HSR performances.

Further evaluations of MS features with other choices of low-pass cut of frequencies f_{3dB} and other values of K , to vary the number amplitude modulation frequencies considered, are to be done. Other future work is to incorporate channel adaptation and speaker adaptation into the MS feature extraction framework.

6. References

- [1] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, Nov 2012.
- [2] 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 on*, vol. 20, no. 1, pp. 30–42, Jan 2012.
- [3] L. Deng, J. Li, J. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero, "Recent advances in deep learning for speech research at microsoft," in *Acoustics, Speech and Signal Processing, 2013 IEEE International Conference on*, May 2013, pp. 8604–8608.
- [4] M. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Acoustics, Speech and Signal Processing, 2013 IEEE International Conference on*, May 2013, pp. 7398–7402.

- [5] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: an overview," in *Acoustics, Speech and Signal Processing*, 2013 *IEEE International Conference on*, May 2013, pp. 8599–8603.
- [6] A. Mohamed, G. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling," in *Acoustics, Speech and Signal Processing*, 2012 *IEEE International Conference on*, March 2012, pp. 4273–4276.
- [7] V. Mitra, W. Wang, H. Franco, Y. Lei, C. Bartels, and M. Gra-ciarena, "Evaluating robust features on deep neural networks for speech recognition in noisy and channel mismatched conditions," in *INTERSPEECH*. ISCA, 2014, pp. 895–899.
- [8] A. Martinez, N. Moritz, and B. T. Meyer, "Should deep neural nets have ears? the role of auditory features in deep learning ap-proaches," in *INTERSPEECH*. ISCA, 2014.
- [9] C. Plack, *The sense of hearing*. Lawrence Erlbaum Associates Publishers, 2005.
- [10] S. Greenberg and B. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation of speech," in *Acoustics, Speech, and Signal Processing*, 1997 *IEEE International Confer-ence on*, vol. 3, 1997, pp. 1647–1650.
- [11] B. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram." *Speech Commu-nication*, vol. 25, no. 1-3, pp. 117–132, 1998.
- [12] P. Clark, G. Sell, and L. Atlas, "A novel approach using modula-tion features for multiphone-based speech recognition," in *Acous-tics, Speech and Signal Processing*, 2011 *IEEE International Conference on*, May 2011, pp. 5264–5267.
- [13] C. E. Schreiner and J. V. U., "Representation of amplitude mod-ulation in the auditory cortex of the cat. I. The anterior auditory field (AAF)," *Hearing Research*, vol. 21, pp. 227 – 241, 1986.
- [14] D. Baby, J. F. Gemmeke, T. Virtanen, and H. Van hamme, "Exemplar-based speech enhancement for deep neural network based automatic speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP)*, 2015 *IEEE International Confer-ence on*, April 2015.
- [15] B. T. Schädler, M. R. and Meyer and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *Journal of Acoustical Soci-ety of America*, vol. 131, no. 5, pp. 4134–4151, 2012.
- [16] R. D. Patterson, M. H. Allerhand, and C. Gigur, "Time-domain modeling of peripheral auditory processing- a modular architec-ture and a software platform," *Journal of Acoustical Society of America*, vol. 98, pp. 1890–1894, 1995.
- [17] M. Slaney, "An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank," in *Technical Report 35*. Ap-ple Computer, Inc., 1993.
- [18] T. Barker and T. Virtanen, "Non-negative tensor factorization of modulation spectrograms for monaural sound source separation," in *INTERSPEECH*. ISCA, 2013.
- [19] D. Baby, T. Virtanen, J. F. Gemmeke, T. Barker, and H. Van hamme, "Exemplar-based noise robust speech recogni-tion using modulation spectrogram features," in *Spoken Language Technology Workshop*, 2014 *IEEE*, South Lake Tahoe, USA, De-cember 2014.
- [20] 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 recog-nition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing So-ciety, Dec. 2011.
- [21] G. Hinton, "A practical guide to training restricted boltzmann ma-chines," in *Neural Networks: Tricks of the Trade (2nd ed.)*, 2012, pp. 599–619.
- [22] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequenc-discriminative training of deep neural networks," in *INTER-SPEECH*. ISCA, 2013, pp. 2345–2349.
- [23] M. Slaney, "Auditory toolbox version 2," *Interval Research Cor-poration*, vol. 10, 1998.