COMPARATIVE INVESTIGATION OF DIFFERENT FRONT-END FEATURES FOR ROBUST SPEECH RECOGNITION

Suyoun Kim
Advisor: Ian Lane

F-14, 11-751, Speech Recognition and Understanding
Carnegie Mellon University
suyoun@cmu.edu

Abstract

Noise robustness in the Automatic Speech Recognition (ASR) systems is becoming increasingly important in the real world applications. Considerable robust feature extraction algorithms especially motivated by human auditory process have been proposed to improve recognition accuracy in noisy conditions. In accordance with introduction of new acoustic model, DNN, using less engineered front-end feature representation, i.e. fbank and focusing on robust modeling have been suggested recently. However, there is no comparative investigation of two approaches in terms of noise robustness, especially in mismatched environment or with reverberation. In this work, we explore the state-of-the-art front-end features, such as PNCC, FBANK, as well as traditional features, MFCC, PLP based on both DNN and GMM acoustic models and then evaluate the performance on noisy version of the RM corpus.

Index Terms: Robust speech recognition

1. Introduction

Noise robustness in the Automatic Speech Recognition (ASR) systems is becoming increasingly important in the real world applications. Neverthless, most speech recognition systems still remain sensitive to the nature of the acoustical environments within which they are deployed, and their performance deteriorates sharply in the presence of sources of degradation such as additive noise, linear channel distortion, and reverberation. One of the most challenging problems is that recognition accuracy degrades significantly if the test environment is different from the training environment and/or if the acoustical environment includes disturbances such as additive noise, channel distortion, speaker differences, reverberation, and so on.

Over the years dozens hundreds of algorithms have been introduced to address this problem and they generally fall into one of two approaches: feature space and model space methods. Feature space methods attempt to remove the corrupting noise from the observations prior to recognition [9] or design a new robust feature inspired by some attributes of auditory physiology and perception [15, 14]. Model space methods leaves the observations unchanged and instead updates the model parameters of the recognizer to be more representative of the observed speech, e.g. [2, 13].

Recently, in addition to the traditional approaches for noise robustness, many researchers begin to use less engineered features and training with better acoustic models to avoid discarding unknown useful information and learn more invariant to the effects of noise. Over the last few years, advances in both machine learning algorithms and computer hardware have led to use Deep Neural Networks (DNNs) that contain many layers of non-linear hidden units as an acoustic model of input features, instead of using Gaussian mixture models (GMMs), to produces posterior probabilities over hidden Markov models (HMMs) states, and DNNs have shown to outperform GMMs on a variety of speech recognition benchmarks [7, 10]. This change of acoustic modeling also led to a resurgence of interest in input representation because DNNs do not require uncorrelated data. For example, it has been found that simple less enginerred filterbank feature outperforms traditional mel frequency cepstral coefficients (MFCCs) [3] or perceptual linear prediction (PLP) coefficients [6] computed from the raw waveform and their first- and second-order temporal differences [4]. The work reported in [1] applied local convolutional filters with max-pooling to the frequency shows that DNNs offers noise robustness.

However, there is no comparative investigation of the physiologically motivated feature representations with DNN acoustic model in terms of noise robustness, especially in mismatched environment or in the presence of reverberation. In this work, we investigate the noise robustness performance of the various state-of-the-art front-end features, PNCC, fbank, as well as traditional features, MFCC, PLP using GMM and DNN based acoustic models. We also investigate model-space noise-adaptive training approach which combines various noisy speeches so that it can learn higher level features that are more invariant to the effects of the environmental distortion during network training itself. We evaluate the performance on RM dataset with a broad range of conditions of noise and reverberation.

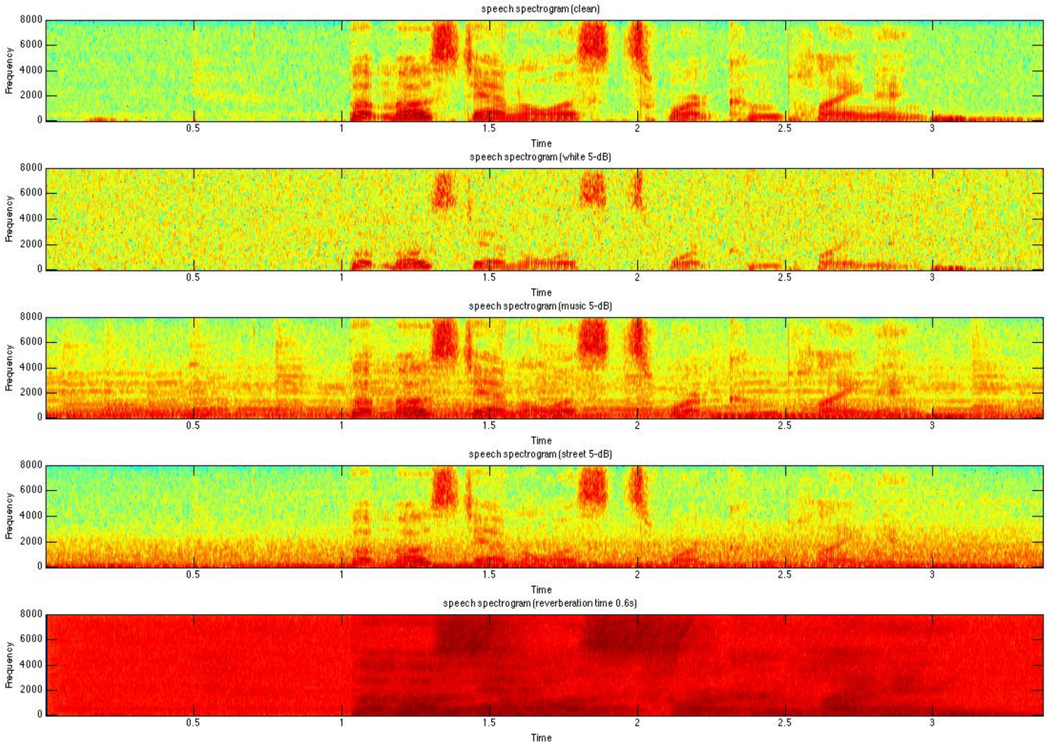


Figure 1: Spectrogram of an utterance with additive noise. As an example of the noise effect that may be derived from different noise types, first row depicts the spectrogram of an utterance from the WSJ database for clean speech and from second to the last row depict spectrograms of the same utterance in the presence of additive white noise, music noise, street noise, at an SNR of 5 dB, and reverberation at 0.6 reverberation time, respectively.

The rest of this paper is organized as follows. Section 2 describes the features based on auditory physiology, power normalized cepstral coefficients (PNCC), and presents the network architecture for combining various front-end features using DNNs and the configurations for evaluating the performances of various features. Section 3 describes the experimental results showing the biologically inspired features help in improving RM word error rate results. Finally, Sections 4 concludes the paper and discusses future work.

2. Experiments

In this section we present our experiments that are intended to explore various state-of-the-art front-end features in a wide variety of acoustic environments. We begin in 2.1 with the description of generating noisy dataset, and in 2.2 we describes the features and we provides the experimental procedures that were used in 2.3.

2.1. Additive Noise Dataset

The experiments were carried out on the DARPA Resource Management (RM) database. The recognition system was trained on 1,600 utterances of clean or degraded speech for training and was tested with 600 utterances.

We generated four additional noisy datasets in our work: (1) digitally-added white noise, (2) digitally-added noise that had been recorded live on urban streets, (3) digitally-added background music, and (4) passage of the signal through simulated reverberation. The street noise was obtained from [8] which was recorded on streets with steady but moderate traffic, and background music was selected from music segments from the original DARPA Hub 4 Broadcast News database. The reverberation simulations were accomplished using the *Room Impulse Response* open source package [5]. The virtual room size used was 5 x 4 x 6 meters, the position of the microphone was 2 x 1.5 x 2 of the room, and the position of the speaker was assumed to be 2 x 3.5 x 2. Each additive noise was included at different SNR, 20 dB, 15 dB, 10 dB, 5 dB, and 0 dB, and three reverberation times, 300 ms, 600 ms, and 1 s were used for generating the noisy conditions to be a progressively difficult.

2.2. Features based on Auditory Physiology

Feature extraction methods based on an understanding of both auditory physiology and psychoacoustics have been incorporated into ASR systems for decades. In recent years, there has been a renewed interest in the development of signal processing procedures based on much more detailed characterization of hearing by humans and other mammals. It is becoming increasingly apparent that the careful implementation of physiologically based and perceptually based signal processing can provide substantially increased robustness in situations in which speech signals are degraded by interfering noise of all types, channel effects, room reverberation, and other sources of distortion. And the fact that humans can hear and understand speech, even under conditions that confound our current machine recognizers, makes us believe that there is more to be gained through a greater understanding of human speech recognition [15, 14]. One of the features incorporating physiological phenomena is power normalized cepstral coefficients (PNCCs) [8]. PNCC processing includes 1) traditional pre-emphasis and short-time Fourier transformation, 2) integration of the squared energy of the STFT outputs using gammatone frequency weighting, 3) medium-time nonlinear processing that suppresses the effects of additive noise and room reverberation, 4) a power-function nonlinearity with exponent 1/15, and 5) generation of cepstral like coefficients using a discrete cosine transform (DCT) and mean normalization. For the most part, noise and reverberation suppression is accomplished by a nonlinear series of operations that accomplish running noise suppression and temporal contrast enhancement, working in a medium-time context with analysis intervals on the order of 50 - 150 ms.
In addition to PNCC, we also evaluated three more types of features, FBANK, MFCC, and PLP.

2.3. DNN Architecture

We employ the Kaldi toolkit as ASR back-end system [12] using the RM bigram language model and about 1,500 context-dependent triphone-states in the acoustic model. In a first step, we set up a GMM-HMM baseline system with 13 mean and variance normalized MFCCs with Delta and acceleration, followed by +-3 frame splicing, linear discriminant analysis (LDA), maximum likelihood linear transform (MLLT), and feature-space maximum likelihood linear regression (fMLLR). For the fair comparison, the alignment of the training data to the HMM states is then extracted from the clean training data using MFCC and used later for training on noisy dataset with other features of the DNN-HMM system. The hybrid DNN-HMM Kaldi system is based on “Dan’s implementation” [16] using a maxout network with 2-norm nonlinearities/activation functions and 2 hidden layers, each one with an input dimension of 1000 and an output dimension of 200.

3. Results

In this section, we explore three different condition to evaluate the performance of noise robustness for the different front-end features. These conditions are 1) training and test with matched condition data (Matched Condition), 2) training with clean speech and test with various noisy data (Mismatched Condition), and 3) training and test with multi-condition data (Combined Condition).

3.0.1. Matched Condition

The performance of these systems is shown in Figure 2 (first column). As the results in the figure indicate, in all matched conditions, the DNN produces substantial improvements compared to the baseline GMM-HMM system. This indicates that DNN acoustic model itself increase the noise robustness. In addition, further gains are achieved by using log mel filterbank features instead of cepstra at an SNR of 0 dB, 5 dB, and reverberation at 0.6 and 1.0 reverberation time. The more degraded input signal, fbank feature shows more robustness rather than MFCC or PLP.

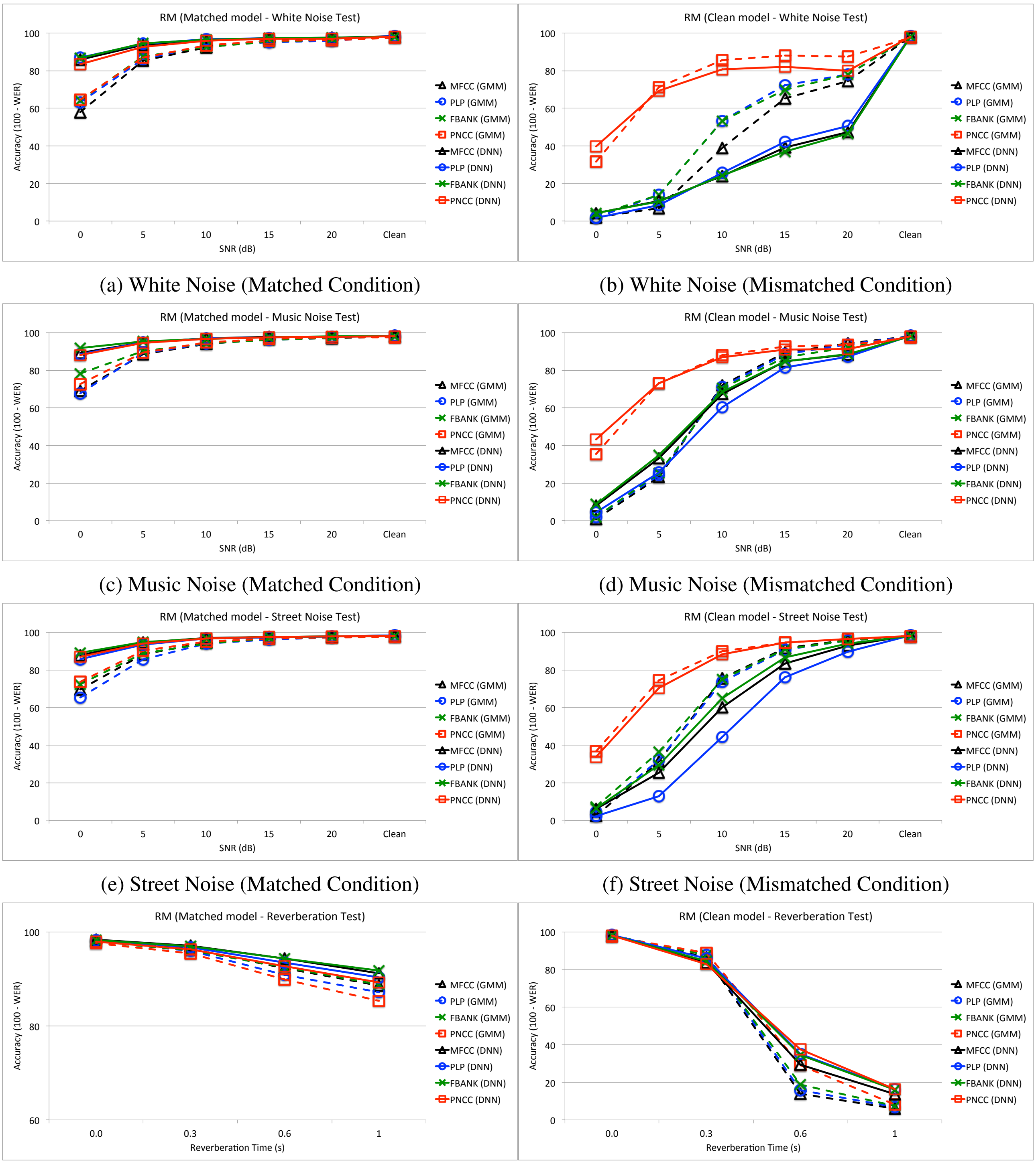
3.0.2. Mismatched Condition

To evaluate the performance in mismatched condition, we trained on features extracted from clean speech data and then tested with same noisy test sets. The results of this experiments are shown in Figure 2 (second column). In these cases, whatever type of the acoustic model was used, both PNCC based on DNN-HMM and PNCC based on GMM-HMM significantly outperform other features.

3.0.3. Combined Condition

Next, we examined the performance in combined conditions, which was trained with the combined all noise types with different SNR and then tested with the combined all noise test sets. As Figure 3 indicates, incorporating PNCC features to DNN acoustic model performs the best.

Finally, in Table 1, the improvement gap between the averaged WER in mismatched condition (second column) and



(a) White Noise (Matched Condition)

(b) White Noise (Mismatched Condition)

(c) Music Noise (Matched Condition)

(d) Music Noise (Mismatched Condition)

(e) Street Noise (Matched Condition)

(f) Street Noise (Mismatched Condition)

(g) Reverberation (Matched Condition)

(h) Reverberation (Mismatched Condition)

Figure 2: Comparison of recognition accuracy for MFCC, PLP, FBANK, and PNCC based on GMM and DNN using the RM corpus in matched (first column) and mismatched (second column) condition.

WER of combined condition (third column) were much larger in DNN model than in GMM model. For example, the gap of PNCC (GMM) is about 4%, whereas the gap of PNCC (DNN) is about 11%. This indicates that training a DNN on multi-condition data enables the network to learn higher level features that are more invariant to the effects of noise. In this case, we can view the deep neural network as a combination of nonlinear feature extractor and nonlinear classifier

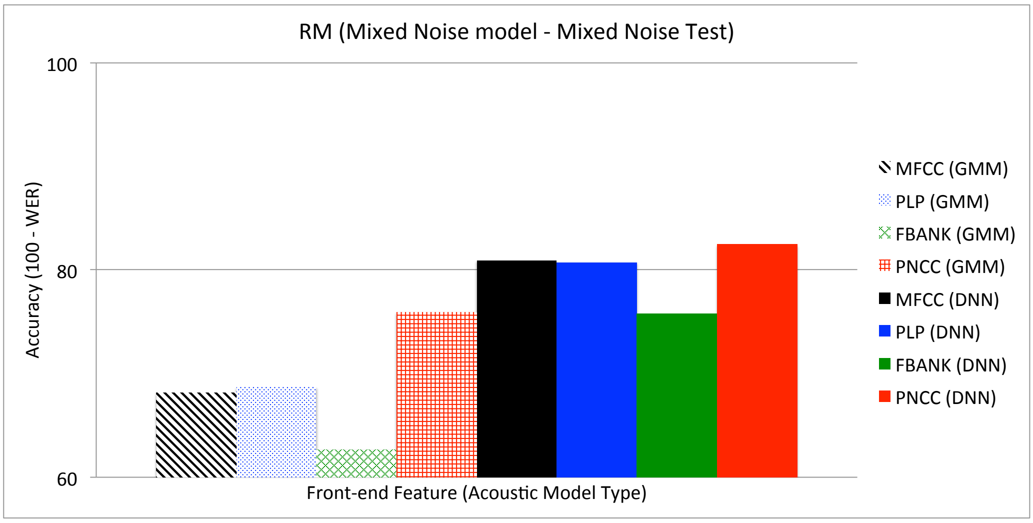


Figure 3: Comparison of recognition accuracy for MFCC, PLP, FBANK, and PNCC based on GMM and DNN using the RM corpus in combined training

Averaged WER(%)			
Feature (Acoustic Model)	Matched Condition	Mismatched Condition	Combined Condition
MFCC (GMM)	10.80	49.04	31.86
PLP (GMM)	11.13	47.16	31.29
FBANK (GMM)	9.69	46.76	37.33
PNCC (GMM)	10.07	28.33	24.08
MFCC (DNN)	5.13	52.63	19.07
PLP (DNN)	5.39	55.38	19.25
FBANK (DNN)	**4.81**	51.44	24.17
PNCC (DNN)	5.79	**28.70**	**17.55**

Table 1: Averaged WER of combinations of features (MFCC, PLP, FBANK, and PNCC) and acoustic models (GMM, and DNN) in matched condition, mismatched condition and combined condition. The bold fonts represent the lowest WER.

where the lower layers are implicitly seeking discriminative features that are invariant across the many acoustic conditions present in the training data. These results clearly demonstrate that the inherent robustness of the combination of feature space and model space approach can be maximized.

4. Conclusions

Through a series of experiments on the the RM database task, we show that the fbank feature with DNN acoustic model has remarkable noise robustness, especially in matched condition. However, it turns out that the PNCC feature with DNN acoustic model outperforms almost test cases in environmental distortion and multi-condition training case. These experimental results indicate that the careful implementation of physiologically based and perceptually based signal processing is still needed for increasing robustness of ASR in situations in which speech signals are degraded by interfering noise of all types, channel effects, room reverberation, and other sources of distortion.

5. Future Work

We have described how biologically inspired features combining DNN acoustic model achieved significant improvements in word accuracy of ASR system in a variety mismatched environment, and we believe that there is more to be gained through a greater understanding of both human auditory physiology and DNN acoustic model. We therefore are planning to model additional key attributes of human auditory system, lateral suppression, which is that the response to signals at a given frequency may be suppressed or inhibited by energy at adjacent frequencies, for more complex signals than pure tones. The presence of a second tone over a range of frequencies surrounding the characteristic frequency (CF) inhibits the response to the probe tone at CF, even for some intensities of the second tone that would be below threshold if it had been presented in isolation. As this form of lateral suppression enhances the response to changes in the signal content with respect to frequency, just as overshoots and undershoots in the transient response have the effect of enhancing the response to changes in signal level over time, we expect that the impact of the additive noise is substantially reduced. Also, we are planning to use noisy version of the Wall Street Journal (WSJ) corpus, Aurora4 dataset [11].

References

[1] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn. Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4277–4280. IEEE, 2012.

[2] A. Acero, L. Deng, Y. Gong, J. Li, and D. Yu. High performance hmm adaptation with joint compensation of additive and convolutive distortions, May 15 2012. US Patent 8,180,637.

[3] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366, 1980.

[4] S. Furui. Cepstral analysis technique for automatic speaker verification. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 29(2):254–272, 1981.

[5] E. A. Habets. Room impulse response generator. *Technische Universiteit Eindhoven, Tech. Rep*, 2(2.4):1, 2006.

[6] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4): 1738–1752, 1990.

[7] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

[8] C. Kim and R. M. Stern. Power-normalized cepstral coefficients (pncc) for robust speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4101–4104. IEEE, 2012.

[9] D. Macho, L. Mauuary, B. Noé, Y. M. Cheng, D. Ealey, D. Jouvet, H. Kelleher, D. Pearce, and F. Saadoun. Evaluation of a noise-robust dsr front-end on aurora databases. In *INTERSPEECH*, 2002.

[10] A.-r. Mohamed, G. E. Dahl, and G. Hinton. Acoustic modeling using deep belief networks. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(1):14–22, 2012.

[11] N. Parihar, J. Picone, D. Pearce, and H.-G. Hirsch. *Performance analysis of the Aurora large vocabulary baseline system*. Citeseer, 2004.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al. The kaldi speech recognition toolkit. In *Proc. ASRU*, pages 1–4, 2011.

[13] M. L. Seltzer, A. Acero, and K. Kalgaonkar. Acoustic model adaptation via linear spline interpolation for robust speech recognition. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 4550–4553. IEEE, 2010.

[14] R. Stern and N. Morgan. Hearing is believing: Biologically-inspired feature extraction for robust automatic speech recognition. *IEEE Signal Processing Magazine*, 29(34-43):170, 2012.

[15] R. M. Stern and N. Morgan. Features based on auditory physiology and perception. *Techniques for Noise Robustness in Automatic Speech Recognition*, pages 193–227, 2012.

[16] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur. Improving deep neural network acoustic models using generalized maxout networks. In *Proc. ICASSP*, 2014.