IMPROVEMENTS TO THE IBM SPEECH ACTIVITY DETECTION SYSTEM FOR THE DARPA RATS PROGRAM

Samuel Thomas[1], George Saon[1], Maarten Van Segbroeck[2] and Shrikanth S. Narayanan[2]

[1]IBM T.J. Watson Research Center, Yorktown Heights, USA
[2]SAIL, University of Southern California, Los Angeles, USA
{sthomas,gsaon}@us.ibm.com,{maarten,shri}@sipi.usc.edu

ABSTRACT

In this paper we describe improvements to the IBM speech activity detection (SAD) system for the third phase of the DARPA RATS program. The progress during this final phase comes from jointly training convolutional and regular deep neural networks with rich time-frequency representations of speech. With these additions, the phase 3 system reduces the equal error rate (EER) significantly on both of the program's development sets (relative improvements of 20% on dev1 and 7% on dev2) compared to an earlier phase 2 system. For the final program evaluation, the newly developed system also performs well past the program target of 3% P_{miss} at 1% P_{fa} with a performance of 1.2% P_{miss} at 1% P_{fa} and 0.3% P_{fa} at 3% P_{miss}.

Index Terms— Speech activity detection, acoustic features, robust speech recognition, deep neural networks.

1. INTRODUCTION

Speech activity detection (SAD) is the first step in most speech processing applications like automatic speech recognition (ASR), language identification (LID), speaker identification (SID) and keyword search (KWS). This important step allows these applications to focus their resources on the speech portions of the input signal. Given its importance, the DARPA RATS program has developed exclusive SAD systems to detect regions of speech in degraded audio signals transmitted over communication channels that are extremely noisy and/or highly distorted [1], in addition to building LID, SID and KWS applications for the same data.

During the course of the program, various sites have developed SAD systems [2, 3, 4, 5, 6, 7, 8] with an end goal of achieving performances better than the final program target of 3% P_{miss} at 1% P_{fa}. P_{miss} is defined as the ratio of the duration of speech missed to the entire duration of speech, while P_{fa} is the ratio between the duration of falsely accepted or inserted speech to the duration of total non-speech in a given set of audio data. Fig. 1 illustrates IBM's performances over 3 phases of the program towards achieving the final program target. Prior to the third and final phase of evaluation, the program ran two evaluations with targets at 5% P_{miss} at 3% P_{fa} (phase 1) and 4% P_{miss} at 1.5% P_{fa} (phase 2). In both these phase evaluations, IBM systems performed past the intermediate targets.

For these evaluations, our systems are trained on recordings from existing conversational telephone corpora (Fisher English and

This work was supported in part by Contract No. D11PC20192 DOI/NBC under the RATS program. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

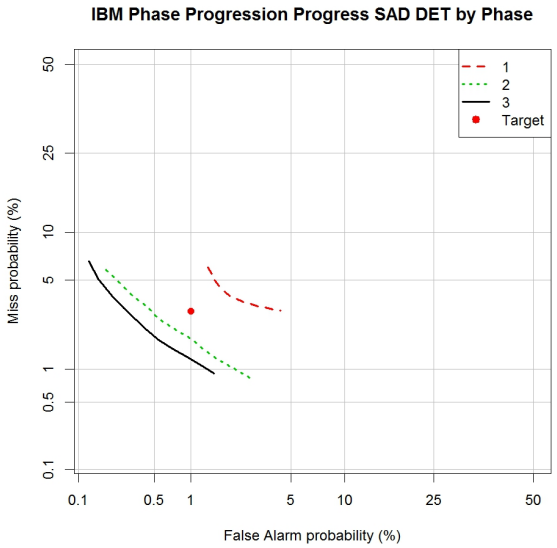

Fig. 1. IBM SAD DET curves for three phases of the RATS program along with the final program target [9].

Arabic Levantine) and new data in Arabic, Levantine, Pasto and Urdu distributed for the program by the Linguistic Data Consortium (LDC) in three incremental releases. The recordings are corrupted by transmitting the original clean audio over 8 different "degraded" radio channels, labeled A through H with a wide range of radio transmission effects [1]. In addition to audio, the corpus of about 2000 (∼250 hours of data per channel) hours of data is automatically annotated into regions of speech, non-speech or non-transmission by appropriately modifying the clean annotations based on unique shift and other transmission artifacts introduced by each channel.

The trained systems are internally evaluated on two official development sets (dev1 and dev2) which contain 11 and 20 hours of audio, respectively. The final evaluation at the end of each phase is performed on an evaluation set of about 24 hours of audio with unreleased transcripts. The results over the three phases of the program in Fig. 1 are based on this same evaluation set.

In section 2 we briefly describe the phase 1 and 2 systems (performances indicated by the dashed red and green lines in Fig. 1). Improvements to the phase 2 system [4] are then described in section 3. These improvements are validated by results from experiments on the dev1 and dev2 sets in section 4. The paper concludes with a discussion and future directions (section 5).

2. PHASE 1 AND 2 SAD SYSTEM ARCHITECTURES

A key design consideration in all the evaluation phases was to treat the segmentation of an audio signal into speech (S), non-speech (NS) and non-transmission (NT) as a simple ASR decoding problem with a three "word" vocabulary [4]. To perform the decoding, an HMM topology with 5 states for each word and a shared start and end state is employed. Each of the 5 states for a word has a self loop and the shared end state is connected back to the start state. For a test audio signal, frame-level scores for each word (S/NS/NT) are then generated from a trained acoustic model before a Viterbi decode is performed to generate segmentations.

Since the evaluation is closed in terms of the channels over which data is transmitted, a second consideration in our framework has been to create channel specific acoustic models for each of the 8 RATS channels. Although no data from any unseen channel needs to be analyzed during test, the channel identity of each utterance needs to be determined. Each utterance is hence processed by a channel detector to select the most appropriate channel model for segmentation. For all the phases, we use 8 channel-dependent GMMs. All Gaussians are scored for every frame and the GMM with the highest total likelihood determines the channel. This approach has 100% channel detection accuracy on both dev1 and dev2 [4].

To improve the performance of speech/non-speech detections by creating diverse systems, starting from phase 2, we use a multi-pass SAD pipeline. In this architecture [2], features used in the first stage of the pipeline are normalized to zero mean and unit variance using audio file-level statistics. The S/NS detections from the first stage are then used to derive statistics from only speech regions. These statistics are then used for feature normalization in the second stage. We focus on two key steps of these SAD systems - the feature extraction stage, for diverse feature representations that capture distinct properties of speech and non-speech and the acoustic modeling stage, for appropriate models that produce reliable acoustic scores using the employed features. For several acoustic features that we use, contextual information is added by appending consecutive frames together. The resulting high dimensional features are then projected to a lower dimension using linear discriminant analysis (LDA). Since the number of output classes is only three, we use a Gaussian-level LDA where we train 32 Gaussians per class and declare the Gaussians as LDA classes [4].

2.1. Phase 1 SAD System

For the single pass SAD system developed in this phase, relatively simple acoustic features and models are used. For each of the 3 classes - S, NS and NT, GMM models are trained on 13 dimensional PLP features extracted every 10 ms from 25 ms analysis windows. After the acoustic features have been normalized at the speaker level, contextual information is added by stacking up to ± 16 frames. A Gaussian-level LDA is finally applied to reduce the dimensionality of the features to 40. Log-likelihood scores from 1024 component GMM models trained on these features are then used as acoustic scores with the HMM based decoder described earlier.

Additionally, a shallow neural network with one hidden layer with 1024 hidden nodes is also trained on 9 consecutive frames of the 40 dimensional features used with the GMM models above, to generate posterior probabilities of the 3 target classes. Scores from the neural network models are then combined with the earlier GMM-based scores, using a weighted log-linear frame-level frame combination. These scores are then used along with the HMM based decoder to produce S/NS/NT segmentations.

2.2. Phase 2 SAD System

In the second phase of the program we build a multi-pass SAD system with a diverse set of features and acoustic models [4]. The acoustic features we use include -

1. *PLP features* - Similar to the features used in the phase 1 system, 13 dimensional PLP features are employed but with additional post-processing. The cepstral coefficients are not only normalized to be zero mean and unit variance using either file-based or speech-only based statistics but are also filtered using an ARMA filter [10] in a temporal window of ± 20 frames.
2. *Voicing features* - The YIN cumulative mean normalized difference [11], an error measure that takes large values for aperiodic signals and small values for periodic signals, is used as a single dimensional voicing feature. This feature is appended with normalized PLP features, yielding a 14 dimensional feature vector. After appending contextual information from 17 consecutive frames, the final vector is projected down to 40 dimensions (PLP+voicing feature) using a Gaussian-level LDA described above.
3. *FDLP features* - A second kind of short-term features [12] are extracted from sub-band envelopes of speech modeled using frequency domain linear prediction (FDLP) [13] . These 13 dimensional features are post-processed by a mean/variance normalization followed by an ARMA filtering, before $\pm$ 8 consecutive frames are spliced and projected down to 40 dimensions using a Gaussian-level LDA.
4. *Rate-scale features* - After filtering the auditory spectrogram [14] using spectro-temporal modulation filters covering 0-2 cycles per octave in the scale dimension and 0.25-25 Hz in the rate dimension [4], 13 dimensional cepstal features are extracted similar to other short-term features above. The rate-scale cepstra are further normalized to zero mean and unit variance and ARMA filtered, before $\pm$ 8 frames are concatenated and projected down to 40 dimensions with a Gaussian-level LDA transform.
5. *Log-mel features* - The log-mel spectra are extracted by first applying 40 mel scale integrators on power spectral estimates (0-8 kHz frequency range) in short analysis windows (25 ms) of the signal followed by the log transform, every 10 ms. In addition to a temporal context of 11 frames, the log-mel features are file/speech only normalized and augmented with their Δ and $\Delta\Delta$s as well.

To model these features, two kinds of acoustic models are trained. The first set of models are deep neural networks (DNNs) trained on fused feature streams obtained by adding various 40 dimensional features (FDLP/Rate-scale features) to the 40 dimensional PLP+voicing feature stream [4]. The input to the DNNs are 320 dimensional features obtained by augmenting the 80 dimensional fused features with their Δ, $\Delta\Delta$ and $\Delta\Delta\Delta$s. The second set of models are convolutional neural networks (CNNs) [15] trained on the 120 dimensional log-mel features. These networks have two convolutional layers using sliding windows of size 9×9 and 4×3 in the first and second layers respectively. Both of these models have 3 hidden layers with 1024 units in each layer and are discriminatively pre-trained before fully trained to convergence.

Using these features and acoustic models, a multi-pass SAD system is built by combining three sets of channel dependent networks using a weighted log-linear frame-level score combination [4]. The three models that were combined are: (i) a DNN trained on a fusion of PLP+voicing and rate scale features with file-based normalization, (ii) a DNN trained on a fusion of PLP+voicing and FDLP features with speech-based normalization, and (iii) a CNN trained on log mel spectral features with speech-based normalization. These models were trained on all the data (2000 hours) and significantly improve speech/non-speech detection (see Fig. 1).

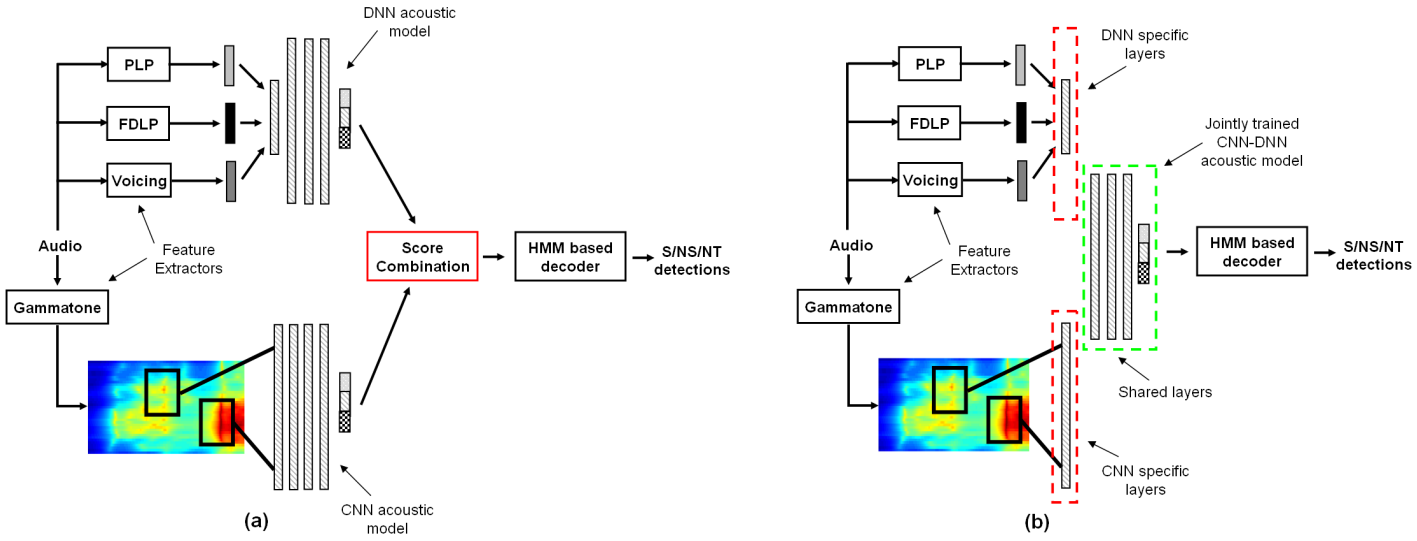

Fig. 2. Schematic of (a) separately trained DNN and CNN models combined at the score level in phase 2 versus (b) a jointly trained CNN-DNN model, on feature representations used in phase 3.

3. IMPROVEMENTS TO THE PHASE 2 SYSTEM

For the third phase of the program we focus again on two key steps - the feature extraction and acoustic modeling components, of the SAD system. On the acoustic modeling front, we work on a new acoustic modeling technique that better integrates training of the diverse feature representations we use. At the feature extraction level, we investigate the use of a Gammatone time-frequency representation that provides additional complementary information.

3.1. Joint training of DNN and CNN acoustic models

One of the primary reasons for considerable gains in the second phase was the adoption of neural network based acoustic models. Using a DNN model, multiple input features can be easily combined by concatenating feature vectors together. In our case we have used a combination of diverse cepstral features - PLP+voicing features along with FDLP or rate-scale based features. These kinds of features, however cannot be used along with the CNNs. CNNs achieve shift invariance by applying a pooling operation on the outputs of its convolutional layers. In order to achieve shift invariance in the feature domain, the features have to be topographical, such as log-mel features. Although the outputs of the CNN systems are quite complementary, their benefits are combined with the DNN models only at the score level using a simple weighted log-linear model. Given the acoustic modeling capabilities of these models, it would however be better if the benefits of a CNN (shift invariance) could be combined with the benefits of a DNN that can use diverse features, at a more earlier stage, by jointly training these diverse acoustic models.

In [16], a neural network graph structure is introduced which allows us to use both topographical (log-mel features) and non-topographical features (PLP+voicing/FDLP features) together. This is achieved by constructing a neural network model with both convolutional layers similar to the input layers of a CNN and input layers similar to that a DNN, followed by shared hidden layers and a single final output layer. The joint CNN-DNN model is trained by combining the outputs/gradients of all input layers during the forward/backward passes. Since most layers are shared, an additional benefit of this configuration is that it has much fewer parameters than separate DNN and CNN models, with only about 10% more parameters than the corresponding CNN. Preliminary experiments for SAD in [16], showed significant relative improvement in equal error rate (EER) from using this kind of jointly trained model over the separate models with score fusion. EER is defined as the point where P_{miss} coincides with P_{fa}. We use these models in phase 3 to build much larger acoustic models and replace individual DNN and CNN models which were previously trained separately.

3.2. The Gammatone feature representation

To improve the performance of the joint CNN-DNN acoustic model, we hypothesize that it is necessary to have a more diverse feature representation than the log-mel features as input for the CNN layers, since the PLP, FDLP and log-mel features have similar filter-bank representations and processing steps. Research in computational auditory scene analysis (CASA) motivates the use of the Gammatone auditory filter bank over the triangular shaped Mel-scale filter bank since the asymmetric shape of the Gammatone filters yield a better approximation of human cochlear filtering [17, 18]. With the Gammatone spectrum for feature extraction showing additional benefits compared to traditional features such as PLP or MFCCs, on several tasks like robust automatic speech recognition [19], speaker verification [20, 21] and language identification [22], we use this representation instead of the log-mel features as input for the CNN layers.

To extract these features the audio data is first downsampled to 8kHz. After pre-emphasizing, Hanning windowing, and framing into frames of 25 ms window length and 10 ms frame shift, the Fourier spectrum is filtered by a filter bank with 64 Gammatone filters. The spectrum is further post-processed by a cubed root compression and temporally smoothed using a second order ARMA filter. The final Gammatone features are also mean and variance normalized on a per utterance basis. Fig. 2 is a schematic of the proposed joint CNN-DNN architecture with Gammatone features.

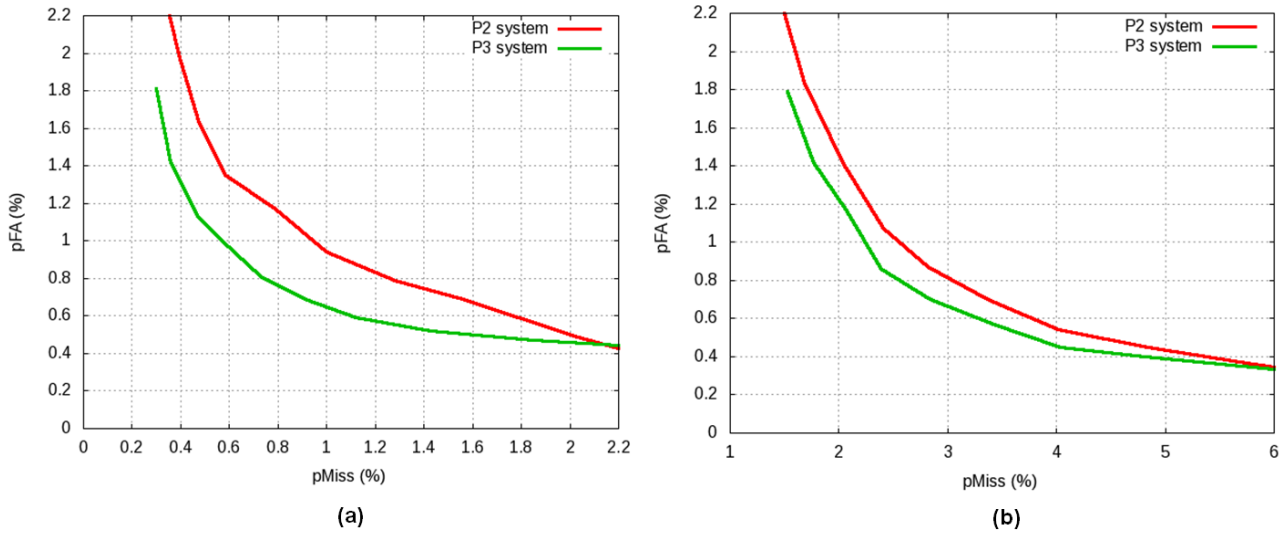

Fig. 3. ROC curves of Phase 2 and Phase 3 systems on the (a) dev1 and (b) dev2 sets.

4. EXPERIMENTS AND RESULTS

For the phase 3 evaluation we build a multi-pass SAD system with two jointly trained neural network based acoustic models for each of the 8 RATS channels on the entire 2000 hours of training data. As in the previous phase, while the first pass acoustic model uses file-level statistics, the second pass model relies on speech detections from the first pass to derive speech only statistics for feature normalization. The input to the DNN layers for both these models are 320 dimensional features obtained by augmenting the 80 dimensional fused features (40 dimensional PLP+voicing features with 40 dimensional FDLP features) with their Δ, $\Delta\Delta$ and $\Delta\Delta\Delta$s. For the CNN layers, 3 streams comprising of 64 dimensional Gammatone features, their Δ and $\Delta\Delta$s are used. The jointly trained model has 2 DNN specific hidden layers (1024 hidden units each) and 2 CNN specific convolutional layers (128 and 256 units each) followed by 5 shared hidden layers (1024 hidden units each) and a final output layer (3 units). All of the 128 nodes in the first convolutional layer of the CNN are attached with 9×9 filters that are two dimensionally convolved with the input representations. The second convolutional layer with 256 nodes has a similar set of 4×3 filters that process the non-linear activations after max pooling from the preceding layer. The non-linear outputs from the second CNN layer are then passed onto the following shared hidden layers. More details about these architectures, training and decoding settings can be found in [16, 23].

In our first set of experiments we compare the performance of a jointly trained acoustic model with the score combination of separately trained DNN and CNN trained systems on dev1. Table 1 shows the performance of 3 different SAD system configurations, each using speech based statistics for feature normalization. We obtain close to 12% relative improvement by jointly training a CNN-DNN system compared to score fusion of individual systems. In a second experiment we replace the CNN feature representation from log-mel to Gammatone based features. With an additional 9% relative improvement from using these diverse features, a total relative improvement of about 20% is achieved compared to the baseline.

In a second set of experiments we test the performance of the multi-pass phase 3 system on both official development sets. The final outputs of the multi-pass system are based on a combination of scores from the first pass and the second pass models. As discussed earlier, both these models are jointly trained CNN-DNN acoustic models. Fig. 3 shows the performances of the phase 2 and the proposed phase 3 models. The phase 2 system is a combination of 3 models as described earlier in section 2. The phase 3 system reduces the EER significantly on both sets with relative improvements of 20% on dev1 and 7% on dev2 compared to the phase 2 system. The improvements on both these developments also translate into significant improvements on the progress set during the phase 3 evaluation (solid black line in Fig 1).

Table 1. Performance (EER%) of DNN/CNN systems on dev1.

System	EER(%)
Score combination of DNN (PLP +voicing+FDLP) and CNN (log-mel)	0.97
Joint training DNN (PLP+voicing +FDLP) and CNN (log-mel)	0.85
Joint training DNN (PLP+voicing +FDLP) and CNN (Gammatone)	0.77

5. CONCLUSIONS

We have presented the IBM SAD system for the RATS phase 3 evaluation. This system achieved significant improvements over the systems developed for previous phases. The gains come from improved acoustic modeling using jointly trained CNN-DNN models and acoustic features that differ in type and normalization. Future work will address the effectiveness of these models on unseen channel conditions and adaptation to those channels.

6. ACKNOWLEDGMENTS

The authors thank Brian Kingsbury, Sriram Ganapathy, Hagen Soltau and Tomas Beran for useful discussions.

7. REFERENCES

[1] K. Walker and S. Strassel, "The RATS Radio Traffic Collection System," in *ISCA Odyssey*, 2012.

[2] T. Ng et al., "Developing a Speech Activity Detection system for the DARPA RATS Program," in *ISCA Interspeech*, 2012.

[3] S. Thomas et al., "Acoustic and Data-driven Features for Robust Speech Activity Detection," in *ISCA Interspeech*, 2012.

[4] G. Saon et al., "The IBM Speech Activity Detection System for the DARPA RATS Program," in *ISCA Interspeech*, 2013.

[5] A. Tsiartas et al., "Multi-band Long-term Signal Variability Features for Robust Voice Activity Detection," in *ISCA Interspeech*, 2013.

[6] M. Graciarena et al., "All for One: Feature Combination for Highly Channel-degraded Speech Activity Detection," in *ISCA Interspeech*, 2013.

[7] S.O. Sadjadi and J.H. Hansen, "Unsupervised Speech Activity Detection using Voicing Measures and Perceptual Spectral Flux," *IEEE Signal Processing Letters*, 2013.

[8] J. Ma, "Improving the Speech Activity Detection for the DARPA RATS Phase-3 Evaluation," in *ISCA Interspeech*, 2014.

[9] H. Goldberg and D. Longfellow, "The DARPA RATS Phase 3 Evaluation," in *DARPA RATS PI Meeting*, 2014.

[10] C.-P. Chen and J. Bilmes, "MVA Processing of Speech Features," *IEEE Transactions on Audio, Speech, and Language Processing*, 2007.

[11] A. de Cheveigne and H. Kawahara, "YIN, a Fundamental Frequency Estimator for Speech and Music," *The Journal of the Acoustical Society of America*, 2002.

[12] S. Thomas, S. Ganapathy, and H. Hermansky, "Phoneme Recognition using Spectral Envelope and Modulation Frequency Features," in *IEEE ICASSP*, 2009.

[13] A. Kumerasan and A. Rao, "Model-based Approach to Envelope and Positive Instantaneous Frequency Estimation of Signals with Speech Applications," in *The Journal of the Acoustical Society of America*, 1999.

[14] T. Chi, P. Ru, and S. Shamma, "Multiresolution Spectrotemporal Analysis of Complex Sounds," in *The Journal of the Acoustical Society of America*, 2005.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient based Learning applied to Document Recognition," *Proceedings of the IEEE*, 1998.

[16] H. Soltau, G. Saon, and T.N. Sainath, "Joint training of convolutional and non-convolutional nueral networks," in *IEEE ICASSP*, 2014.

[17] M. Slaney et al., "An Efficient Implementation of the Patterson-Holdsworth Auditory Filterbank," *Apple Computer, Perception Group, Tech. Rep*, 1993.

[18] E.A. Lopez-Poveda and R. Meddis, "A Human Nonlinear Cochlear Filterbank," *The Journal of the Acoustical Society of America*, 2001.

[19] Y. Shao, Z. Jin, D.L. Wang, and S. Srinivasan, "An Auditory-based Feature for Robust Speech Recognition," in *IEEE ICASSP*, 2009.

[20] Y. Shao and D.L. Wang, "Robust Speaker Identification using Auditory Features and Computational Auditory Scene Analysis," in *IEEE ICASSP*, 2008.

[21] M. Li, A. Tsiartas, M.V. Segbroeck, and S. Narayanan, "Speaker Verification using Simplified and Supervised i-vector Modeling," in *IEEE ICASSP*, 2013.

[22] M.V. Segbroeck, R. Travadi, and S. Narayanan, "UBM Fused Total Variability Modeling for Language Identification," in *ISCA Interspeech*, 2014.

[23] H. Soltau, H.K. Kuo, L. Mangu, G. Saon, and T. Beran, "Neural Network Acoustic Models for the DARPA RATS Program," in *ISCA Interspeech*, 2013.