FILTERBANK LEARNING USING CONVOLUTIONAL RESTRICTED BOLTZMANN MACHINE FOR SPEECH RECOGNITION

Hardik B. Sailor and Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar-382007, Gujarat, India

hardik_sailor@daiict.ac.in, hemant_patil@daiict.ac.in

ABSTRACT

Convolutional Restricted Boltzmann Machine (ConvRBM) as a model for speech signal is presented in this paper. We have developed ConvRBM with sampling from noisy rectified linear units (NReLUs). ConvRBM is trained in an unsupervised way to model speech signal of arbitrary lengths. Weights of the model can represent an auditory-like filterbank. Our proposed learned filterbank is also nonlinear with respect to center frequencies of subband filters similar to standard filterbanks (such as Mel, Bark, ERB, etc.). We have used our proposed model as a front-end to learn features and applied to speech recognition task. Performance of ConvRBM features is improved compared to MFCC with relative improvement of *5%* on TIMIT test set and *7%* on WSJ0 database for both Nov'92 test sets using GMM-HMM systems. With DNN-HMM systems, we achieved relative improvement of *3%* on TIMIT test set over MFCC and Mel filterbank (FBANK). On WSJ0 Nov'92 test sets, we achieved relative improvement of *4-14%* using ConvRBM features over MFCC features and *3.6-5.6%* using ConvRBM filterbank over FBANK features.

Index Terms— Convolutional RBM, rectified linear units, pooling, filterbank.

1. INTRODUCTION

Features for speech processing applications specifically in speech recognition area, are based on properties of human auditory processing [1], [2]. Many years of research have been done to design auditory-based features for speech processing applications. Some of the approaches are engineering-based and some are based on representation learning, i.e., data-driven models [2]. Representation learning is a growing research area under machine learning where underlying model can automatically discover features needed for detection or classification from the raw signals [3], [4]. Unsupervised learning is the most important form of representation learning since human learning is largely unsupervised [4], [5].

The authors would like to thank Dept. of Electronics and Information Technology (DeitY), Govt. of India for sponsoring two consortium projects, (1) TTS Phase II (2) ASR Phase II and authorities of DA-IICT.

Earlier works on unsupervised learning to learn filterbank include use of Independent Component Analysis (ICA) applied to samples of speech signals in small windows [6], [7]. In [8], Non-negative Matrix Factorization (NMF) was applied to speech power spectra to learn auditory-like filterbank. Recently, Restricted Boltzmann Machine (RBM) with rectified linear units (ReLUs) was used for representation learning in ASR task [9]. To learn invariant representation and improve scalability of RBM, Convolutional RBM was proposed [10]. ConvRBM with sigmoid units was applied in speech processing applications using spectrograms to learn temporal modulation features [11]. All previous unsupervised learning methods for speech signals perform similar to MFCC and filterbank features. However, these methods did not show improvements over these features. These methods rely on processing with smaller windows of speech signals. We have developed ConvRBM to model full length *1-D* speech signals using NReLUs to avoid problems with block-based processing [12]. Convolutional models applied to speech signals in time-domain provide shift invariance [12].

Recently, supervised learning has become quite popular for learning filterbanks and acoustic modeling from raw speech signals such as study reported in [13], [14], [15], [16]. Mel-like filterbank was learned from FFT spectra of speech using DNN in [17]. We have separated unsupervised filterbank learning using ConvRBM and supervised acoustic modeling using GMM and DNN for ASR as done in [9].

In this paper, we have developed novel convolutional RBM with rectified linear units that can model speech signal of any length in an unsupervised way. We have shown that our learned filterbank has similar properties of hand-crafted filterbanks. Experiments on TIMIT and WSJ0 speech recognition task shows improved performance compared to standard spectral features such as MFCC and Mel-filterbank.

2. CONVOLUTIONAL RBM FOR SPEECH SIGNALS

In this Section, we describe ConvRBM developed for modeling speech signals of varying lengths. ConvRBM has two layers, namely, visible layer and hidden layer [10]. The input to ConvRBM is an entire speech signal of length n-samples.

Hidden layer consists of *K*-groups (i.e., number of filters) with filter length *m*-samples in each. Weights (also called as filters or subband filters with respect to speech perception [18]) are shared between visible and hidden units among all the locations in each group [10]. If we denote b_k as the hidden bias for k^{th} group, then response of the convolution layer is given as:

$$I_k = (x * \tilde{w}^k) + b_k, \quad (1)$$

where $x = [x_1, x_2, ..., x_n]$ are samples of speech signal, $w^k = [w_1^k, w_2^k, ..., w_m^k]$ is a weight vector (i.e., k^{th} filter) and $\tilde{w}$ denote flipped array [10]. The energy function for ConvRBM is given as,

$$E(\mathbf{x}, \mathbf{h}) = \frac{1}{2\sigma_x^2} \sum_{i=1}^{n} x_i^2 - \frac{1}{\sigma_x} \sum_{k=1}^{K} \sum_{j=1}^{l} h_j^k I_k - \frac{c}{\sigma_x^2} \sum_{i=1}^{n} x_i, \quad (2)$$

where convolution length $l = n - m + 1$ and c is a shared visible bias. Each speech signal is normalized to zero mean and unit variance. Hence, variance (σ_x) in (2) is set to *1*. Hidden units are sampled using noisy ReLUs as suggested in [19]. We have used single-step contrastive divergence for model learning [20]. Following are the sampling equations for hidden and visible units (to reconstruct speech signal x_{recon}):

$$\begin{aligned} h^k &\sim max(0, I_k + N(0, \sigma(I_k))), \\ x_{recon} &\sim \mathcal{N}\left(\sum\nolimits_k (h^k * w^k) + c, 1\right), \end{aligned} \quad (3)$$

where $N(0, \sigma(I_k))$ is a Gaussian noise with mean zero and sigmoid of I_k as a variance and $\mathcal{N}(\mu, 1)$ is Gaussian distribution with mean μ and variance *1*.

After ConvRBM is trained, pooling is applied to reduce representation of ConvRBM filter responses in temporal-domain. Our model is different than used in [11] where probabilistic max-pooling was used in inference stage itself for binary hidden units. Our approach resembles the method used in [21] where time-domain gammatone responses were reduced using average-based framing, a pooling-like operation. For signal of sampling frequency *Fs = 16* kHz, pooling is applied using *25* ms (*400* samples) window length (wl) and *10* ms (*160* samples) shift (ws). We have used this setup to compare standard spectral features (e.g., MFCC) extracted using same windowing parameters. Pooling is performed across time and separately for each filter. Speech signal with *n*-samples has $F = \frac{n - wl + ws}{ws}$ number of frames. We have experimented with both average and max-pooling and found better results with average pooling. After pooling operation, stabilized logarithmic non-linearity $\log(\cdot + 0.0001)$ is applied as done in [22].

The block diagram of the model described above is shown in Figure 1. During feature extraction stage, we have used deterministic ReLU non-linearity $max(0, I_k)$ as activation function of hidden units. Pooling operation reduce temporal resolution from $K \times n$ samples to $K \times F$ frames. Logarithmic non-linearity compresses the dynamic range of features which was found to improve performance in ASR [22]. The feature extraction steps involved in this ordering resembles the processing in human ear auditory representations [23].

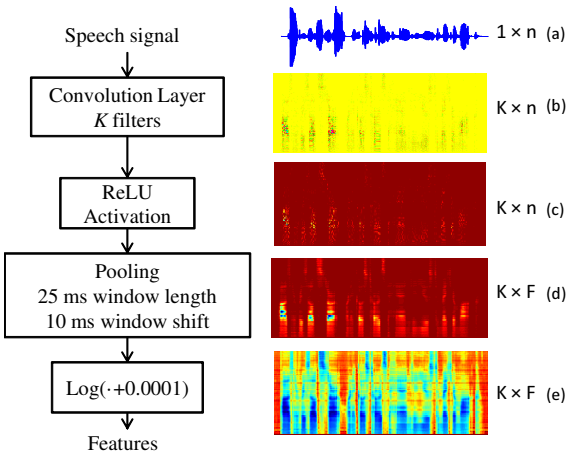

Fig. 1. Block diagram of stages in feature representation using trained ConvRBM. To shows figures on right side, filters were arranged in increasing order of center frequency. (a) speech signal, (b) and (c) responses from convolution layer (same length) and ReLU nonlinearity, respectively, (d) pooling operation, (e) logarithmic compression.

3. ANALYSIS OF FILTERBANK

3.1. Analysis of subband filters

Examples of subband filters learned using ConvRBM are shown in Figure 2. Filters were arranged according to increasing order of center frequencies. Weights of ConvRBM were initialized randomly and there is no constraint on filter shapes, still the model is able to learn meaningful representation. Impulse responses of filters in time-domain are shown in Figure 2(a). We can see that many filters are very similar to auditory gammatone filters. Unlike the filters derived using RBM [9], our filters resemble more like auditory filters for speech signals [7]. This may be due to the fact that RBM was trained on randomly selected smaller windows of speech signal and hence, they were in any random temporal phase [9]. We have trained our model on speech signals in time-domain without windowing to learn filters and pooled later to get short-term spectrum representation. Figure 2(b) shows frequency-domain representation of corresponding time-domain impulse responses. We can see that all filters are localized in frequency-domain with different center frequencies. Filters with lower center frequencies are highly localized in frequency-domain while filters with higher center frequencies are more broad in terms of bandwidth and hence, mimic human perception for hearing.

Our model can also accurately reconstruct speech signal even after ReLU non-linearity. Small segment of original speech (about *500* samples) from TIMIT database, segment of a reconstructed speech from model (eq. (3)) and residual

error is shown in Figure 3. From residual error, we can see very accurate reconstruction of speech signal.

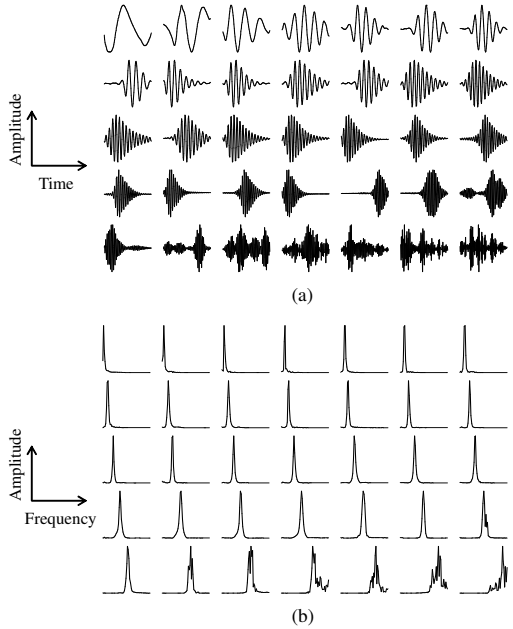

Fig. 2. Examples of subband filters learned using ConvRBM: (a) filters in time-domain (i.e., impulse responses), (b) filters in frequency-domain (i.e., frequency responses).

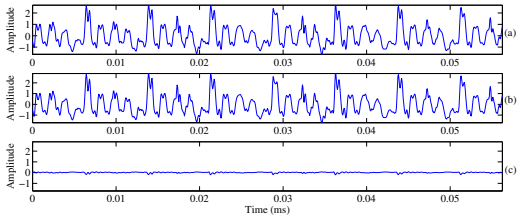

Fig. 3. (a) Segment of speech, (b) reconstructed from model, (c) residual error. Root Mean Squared Error (RMSE) between original and reconstructed speech is *0.0453*.

3.2. Comparison with standard filterbanks

In order to compare our filterbank with standard auditory filterbanks, we have shown center frequency *vs*. subband filter index plot in Figure 4. We can see that our filterbank has also nonlinear relationship between center frequencies and filter ordering (and hence, bandwidth of filters) similar as other auditory filterbanks. More number of subband filters are required for lower frequencies compared to higher frequencies. Hence, our learned filters can represent frequency tuning in human cochlea which can be modeled using a bank of subband filters. The spectrum representation of filters obtained following the steps in Figure 1, is compared with log-Mel spectrogram in Figure 5. Similar as log-Mel spectrogram, ConvRBM spectrogram indeed represent spectrum information such as formant contours, voiced and unvoiced sounds.

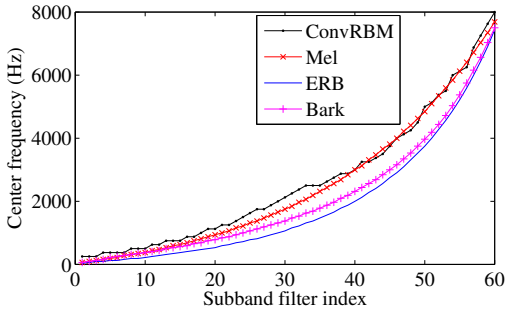

Fig. 4. Comparison of filterbank learned using ConvRBM with auditory filterbanks.

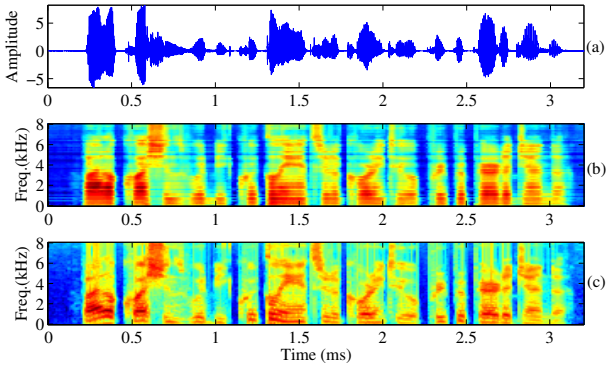

Fig. 5. (a) Speech signal, (b) spectrogram using ConvRBM filterbank, (c) log-Mel spectrogram.

4. EXPERIMENTAL SETUP

4.1. Speech databases

Speech recognition experiments were conducted on TIMIT [24] (for phone recognition task) and Wall Street Journal WSJ0 database [25]. In TIMIT database, all SA category sentences (same sentences spoken by all speakers) were removed as they may bias the speech recognition performance. Training data contains utterances from *462* speakers. Development and test sets contain utterances from *50* and *24* speakers, respectively. WSJ0 SI-84 training data consists of *14* hours of speech data which includes *7138* utterances spoken by *84* speakers. Two Nov'92 evaluation sets, namely, *5K*-word and *20K*-word vocabulary denoted as eval92_5K and eval92_20K, respectively, were used for testing.

4.2. Training of ConvRBM and feature extraction

Mean-variance normalized speech signals were applied to ConvRBM. Learning rate was chosen to be *0.005* which was fixed for first *10* epochs and decayed later at each epochs for stable learning. We observed that with NReLUs, only *25-35* training epochs were sufficient. For first five training epochs, momentum was set to *0.5* and after that it was set to *0.9*. We have trained model with different lengths of ConvRBM filters and with different number of filters. After model was trained, features were extracted from speech signal as shown

in Figure 1. To reduce dimension and compare with MFCC feature set, Discrete Cosine Transform (DCT) was applied (except in filterbank experiments) and only first *13*-D were retained. Delta and delta-delta features were also appended resulting in *39*-D feature vector.

4.3. ASR system building

For both databases, baseline monophone GMM-HMM systems were built using *39*-D MFCC features. MFCCs were extracted from windows of speech signal with *25* ms length and *10* ms shift similar as parameters of pooling. For TIMIT database, *48* phones were used for training and mapped to *39* phones during scoring [26]. Language modeling (LM) was performed using bi-gram language model. For WSJ0 database, tri-gram language model was used. DNN-HMM systems were built using Karel's recipe (without pre-training) in Kaldi [27] and results are reported with parameters: *3* layers, *1500* hidden units and *11* frame context-window.

5. EXPERIMENTAL RESULTS

5.1. Experiments on TIMIT database

The effect of number of filters, filter length and pooling type is verified through experiments on TIMIT database using GMM-HMM systems and results are reported in Table 1. We can see that optimal filter length corresponding to least Phone Error Rate (PER) is *128* samples on development (Dev) and test set. Filter length *128* samples, i.e., *8* ms is sufficient to capture small temporal variations in speech signals [7]. In our case, average pooling works better than max-pooling. Since we are using rectifier nonlinearity, it eliminates cancellations between neighboring filter outputs when combined with average pooling [28]. Best performance is obtained with *60* filters, *128* samples filter length and using average pooling.

Table 1. Comparison of number of subband filters, filter length and pooling type on TIMIT database in % PER

No. of filters	Filter length	Pooling type	Dev	Test
40	128	Avg	32.0	32.6
60	128	Avg	**31.2**	**31.8**
80	128	Avg	31.5	31.9
60	96	Avg	31.4	32.5
60	160	Avg	31.7	33.0
60	256	Avg	32.8	33.5
60	128	Max	32.6	33.5

Avg=Average, Max=Maximum

The comparison of MFCC and ConvRBM features using GMM-HMM systems are shown in Table 2. We can see that ConvRBM features perform better than MFCC features giving an absolute reduction of *1.5%* in PER on development set and *1.7%* on test set. We also experimented on hybrid DNN-HMM system with forced aligned labels obtained from corresponding GMM-HMM systems. To compare with Mel-filterbank features (FBANK), we have used ConvRBM trained on *40* filters even though less improvement in GMM-HMM systems compared to *60* filters. Table 2 shows that for DNN-HMM systems there is relative improvement of *3%* on test set using ConvRBM features and *2.6%* using ConvRBM filterbank over MFCC and FBANK features, respectively.

Table 2. Results on TIMIT database in % PER

Feature set	System	Dev	Test
MFCC (39-D)	GMM-HMM	32.7	33.5
ConvRBM (39-D)	GMM-HMM	**31.2**	**31.8**
MFCC (39-D)	DNN-HMM	23.0	24.0
ConvRBM (39-D)	DNN-HMM	**21.9**	**23.3**
FBANK (120-D)	DNN-HMM	22.2	23.4
ConvRBM-filterbank (120-D)	DNN-HMM	**21.5**	**22.8**

5.2. Experiments on WSJ0 database

For WSJ0 database, results of ASR experiments are reported in Table 3 in terms of % Word Error Rate (WER). Performance is improved using ConvRBM features compared to MFCC features. For GMM-HMM system, there is an absolute reduction of *0.99%* WER on eval92_5K test set and *1.92%* WER on eval92_20K test set over MFCC features. For DNN-HMM systems, lowest WER *5.85%* (*3.6%* relative improvement) for 5K test is achieved with ConvRBM filterbank while improvement is less using ConvRBM features. For 20K test set, ConvRBM features and ConvRBM filterbank yielded almost similar WER (although both have different number of filters). However, relative improvement of *14.6%* over MFCC and *5.6%* over FBANK features is achieved.

Table 3. Results on WSJ0 database in % WER

Feature set	System	eval92_5K	eval92_20K
MFCC(39-D)	GMM-HMM	13.95	27.72
ConvRBM(39-D)	GMM-HMM	**12.96**	**25.80**
MFCC(39-D)	DNN-HMM	6.30	15.70
ConvRBM(39-D)	DNN-HMM	**6.05**	**13.40**
FBANK (120-D)	DNN-HMM	6.07	14.32
ConvRBM-filterbank(120-D)	DNN-HMM	**5.85**	**13.52**

6. SUMMARY AND CONCLUSIONS

In this paper, convolutional RBM using NReLUs is developed to model raw speech signals. Filters learned in ConvRBM resembles auditory filters in human cochlea and also comparable with other auditory filterbanks. Features extracted from ConvRBM were used for speech recognition task. Experiments on TIMIT and WSJ0 databases shows that ConvRBM features perform better than standard spectral features in both GMM-HMM and hybrid DNN-HMM systems. Our future work will involve developing deep speech model where second layer of ConvRBM can model auditory cortex and learn *2D* Gabor-like subband filters. We will also use our model for low resource and noise robust ASR task.

7. REFERENCES

[1] R.M. Stern and N. Morgan, "Hearing is believing: Biologically inspired methods for robust automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 34–43, Nov 2012.

[2] H. Hermansky, J.R. Cohen, and R.M. Stern, "Perceptual properties of current speech recognition technology," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 1968–1985, Sept 2013.

[3] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, Aug 2013.

[4] Y. Bengio Y. LeCun and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[5] G. Hinton, "Where do features come from?," *Cognitive Science*, vol. 38, no. 6, pp. 1078–1101, 2014.

[6] J. Lee, H. Jung, T. Lee, and S. Lee, "Speech feature extraction using independent component analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2000, vol. 3, pp. 1631–1634.

[7] M. S. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, no. 4, pp. 356–363, 2002.

[8] A. Bertrand, K. Demuynck, V. Stouten, and H. Van hamme, "Unsupervised learning of auditory filter banks using non-negative matrix factorisation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP) 2008, Las Vegas, Nevada*, 2008, pp. 4713–4716.

[9] N. Jaitly and G. Hinton, "Learning a better representation of speech soundwaves using restricted Boltzmann machines," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2011, pp. 5884–5887.

[10] H. Lee, R. B. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26^{th} Annual International Conference on Machine Learning, (ICML), Canada, June 14-18*, 2009, pp. 609–616.

[11] H. Lee, P. T. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *23^{rd} Annual Conference on Neural Information Processing Systems, Canada, 7-10 December*, 2009, pp. 1096–1104.

[12] E. Smith and M. S. Lewicki, "Efficient coding of time-relative structure using spikes," *Neural Computation*, vol. 17, no. 1, pp. 19–45, Jan. 2005.

[13] Z. Tüske, P. Golik, R. Schlüter, and H. Ney, "Acoustic modeling with deep neural networks using raw time signal for LVCSR," in *INTERSPEECH*, Singapore, Sept. 2014, pp. 890–894.

[14] D. Palaz, M. Magimai-Doss, and R. Collobert, "Convolutional neural networks-based continuous speech recognition using raw speech signal," in *40^{th} International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD*, 2015, pp. 4295–4299.

[15] P. Golik, Z. Tüske, R. Schlüter, and H. Ney, "Convolutional neural networks for acoustic modeling of raw time signal in LVCSR," in *INTERSPEECH*, Dresden, Germany, 6-10 Sep 2015, pp. 26–30.

[16] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *INTERSPEECH*, Dresden, Germany, 6–10 Sep 2015, pp. 1–5.

[17] T. N. Sainath, B. Kingsbury, A.-R. Mohamed, and B. Ramabhadran, "Learning filter banks within a deep neural network framework," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2013, pp. 297–302.

[18] J. B. Allen, "How do humans process and recognize speech?," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 567–577, Oct 1994.

[19] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010, pp. 807–814.

[20] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[21] J. Qi, D. Wang, Y. Jiang, and R. Liu, "Auditory features based on gammatone filters for robust speech recognition," in *2013 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2013, pp. 305–308.

[22] Y. Hoshen, R. J. Weiss, and K. W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 4624–4628.

[23] X. Yang, K. Wang, and S.A. Shamma, "Auditory representations of acoustic signals," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 824–839, March 1992.

[24] Garofolo et al., "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, pp. 27403, 1993.

[25] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the Workshop on Speech and Natural Language*, Stroudsburg, PA, USA, 1992, HLT '91, pp. 357–362, Association for Computational Linguistics.

[26] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, Nov 1989.

[27] D. Povey et al., "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Dec. 2011.

[28] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?," in *IEEE 12^{th} International Conference on Computer Vision*, Sept 2009, pp. 2146–2153.