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
{hardik_sailor, hemant_patil}@daiict.ac.in

Introduction

- Human auditory processing → design features
- Representation learning to discover features
- Unsupervised learning to learn filterbanks
- Convolutional models avoid block-based processing
- Convolutional RBM to learn filterbanks directly from speech signals

Convolutional RBM for Speech Signals

- ConvRBM has two layers: visible layer and hidden layer [1], [2].
- 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 subband filters) are shared between visible and hidden units [1].
- The response of the convolution layer is given as:

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

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 and $\tilde{w}$ denote flipped array.

- The energy function for ConvRBM is given as,

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

where convolution length $l = n - m + 1$, $\sigma_x = 1$ and c is a shared visible bias.

- Hidden units are sampled using noisy ReLUs as done in [3].
- Single-step contrastive divergence for model learning.
- Following are the sampling equations for hidden and visible units (to reconstruct speech signal x_{recon}):

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

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.

Feature Representation

- Pooling is applied to reduce representation of ConvRBM filter responses in temporal-domain.
- Pooling is performed across time and separately for each filter using *25* ms window length (wl) and *10* ms shift (ws).
- Logarithmic non-linearity compresses the dynamic range of features.

Acknowledgements

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, Gandhinagar, India.

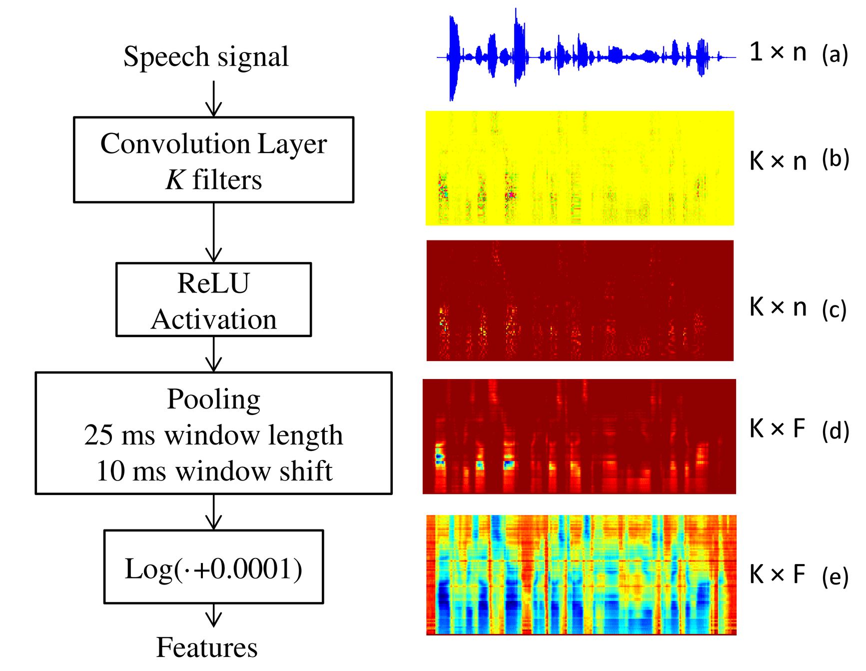

Fig. 1 Block diagram of stages in feature representation using trained ConvRBM. To shows figures on right side, filters were arranged in increasing order of center frequency. (a) speech signal, (b) and (c) responses from convolution layer and ReLU nonlinearity, respectively, (d) pooling operation, (e)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

- The feature extraction steps involved in this ordering resembles the auditory processing in human ear [4].

Analysis of Filterbank

- Weights of ConvRBM were initialized randomly and there is no constraint on filter shapes.
- Many filters are very similar to auditory gammatone filters.
- Filters with lower center frequencies are highly localized in frequency-domain while filters with higher center frequencies are more broad in terms of bandwidth.
- Mimic the human perception for hearing.

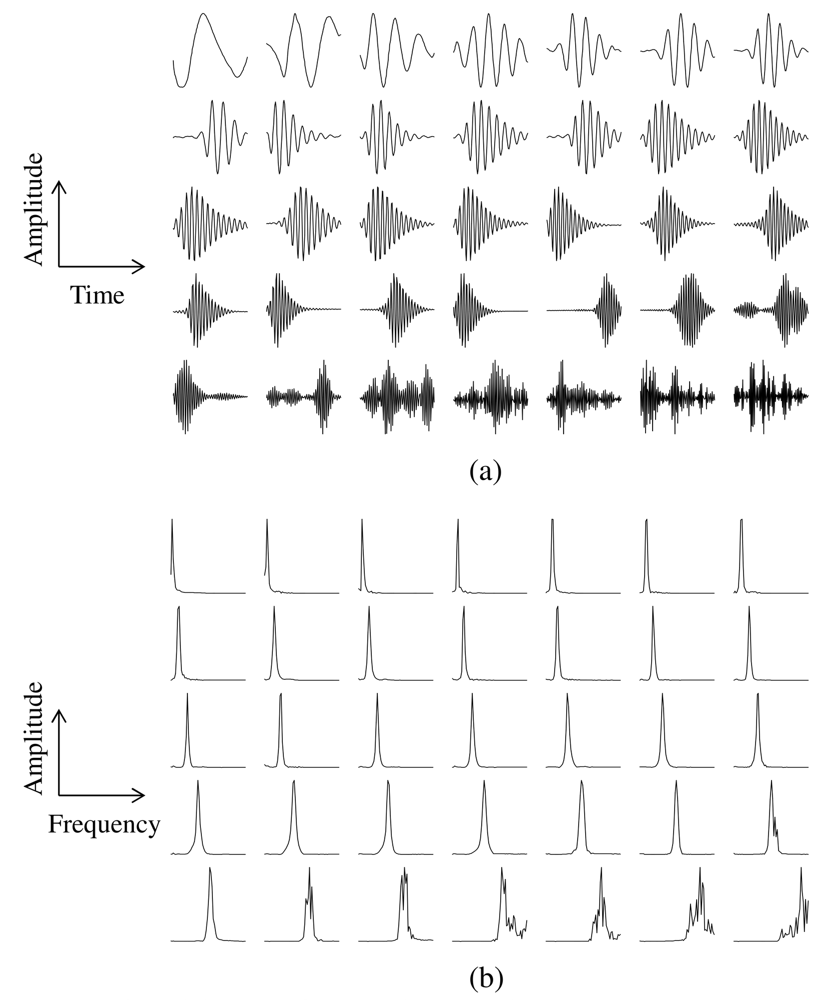

Fig. 2 Examples of subband filters learned using ConvRBM: (a) filters in time-domain (i.e., impulse responses), (b) filters in frequency-domain (i.e., frequency responses).

- Our model can also accurately reconstruct speech signal even after ReLU non-linearity.

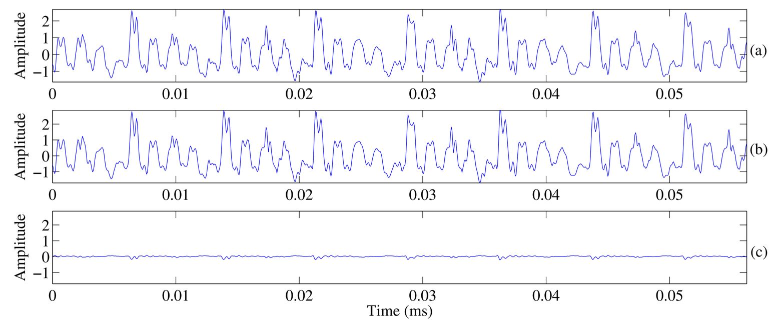

Fig. 3 (a) Segment of speech, (b) reconstructed from model, (c) residual error. Root Mean Squared Error (RMSE) between original and reconstructed speech is *0.0453*.

Comparison With Standard Filterbanks

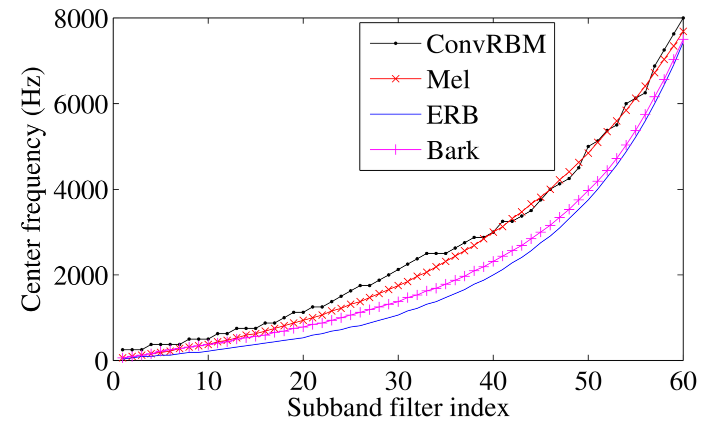

Fig. 4 Comparison of filterbank learned using ConvRBM with auditory filterbanks.

- 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.
- Learned filters can represent frequency tuning in human cochlea.

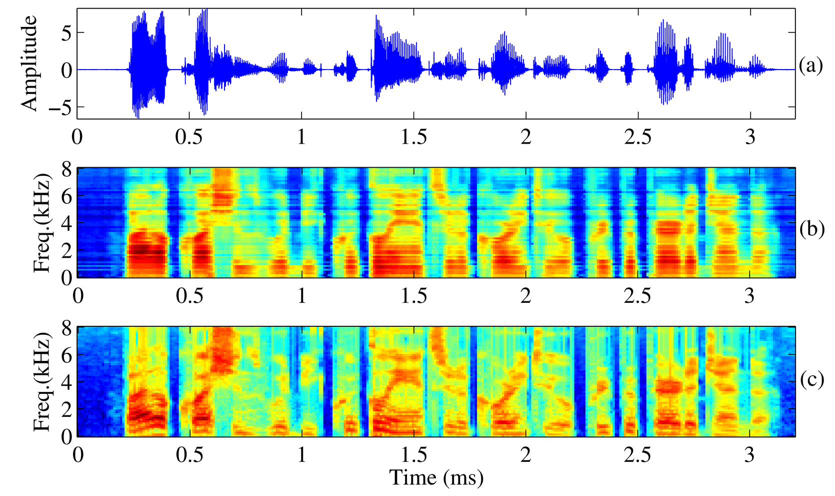

Fig. 5 (a) Speech signal, (b) spectrogram using ConvRBM filterbank, (c) log-Mel spectrogram.

- ConvRBM spectrogram indeed represent spectrum information such as formant contours, voiced and unvoiced sounds.

Experimental Setup

- Speech recognition experiments were conducted on TIMIT [5] (for phone recognition task) and Wall Street Journal WSJ0 database [6].

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.
- For first five training epochs, momentum was set to *0.5* and after that it was set to *0.9*.

ASR System Building

- Baseline monophone GMM-HMM and hybrid DNN-HMM system systems were built using *39-D* MFCC and *120-D* Mel filterbank features features.
- Results are reported on GMM-HMM and hybrid DNN-HMM systems with parameters: *3* hidden layers, *1500* hidden units and *11* frame context-window.

Experimental Results

Table 1: ConvRBM parameter tuning 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

- Filter length *128* samples, i.e., *8* ms is sufficient to capture small temporal variations in speech signals.

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**

- Relative improvement of *3%* on TIMIT test set over MFCC and Mel filterbank (FBANK).

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**

- Relative improvement of *4-14%* using ConvRBM features over MFCC features and *3.6-5.6%* using ConvRBM filterbank over FBANK features.

References

[1] 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 26th Annual International Conference on Machine Learning, (ICML), Canada, June 14-18*, 2009, pp. 609–616.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

Presented by Hardik B. Sailor in 41st IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP) 20-25 March 2016, Shanghai, China.