DNN derived filters for processing of modulation spectrum of speech

Jan Pešán[1], *Lukáš Burget*[1], *Hynek Heřmanský*[1 2], *Karel Veselý*[1]

[1]Brno University of Technology, Speech@FIT group and IT4I Centre of excellence, Czech Republic
[2]Johns Hopkins University, Baltimore, MD, USA
{ipesan,burget,iveselyk}@fit.vutbr.cz, hynek@jhu.edu

Abstract

We propose a novel approach to design modulation frequency filters for the first stage processing of critical band spectrum of speech using deep neural network (DNN). These filters replace conventional modulation frequency filters currently used in state-of-the-art BUT speech recognition system and yield about 10% relative improvement in phoneme recognition accuracy. The resulting filters are consistent with some known temporal properties of higher levels of mammalian auditory processing and suggest more efficient scheme for pre-processing of speech for ASR.

Index Terms: deep neural network, convolutive layer, modulation filters, mammalian auditory processing

1. Introduction

In early days of ASR, typical features were derived from rather short 10-20 ms segments of the signal. However, over the years, ASR field gradually started to use longer and longer temporal contexts, and features of some of state-of-the-art systems are derived from temporal contexts of hundreds of ms. Such longer temporal context carries information about spectral dynamics rather than merely using a short-time spectrum of speech [1]. This brings the concept of so called modulation spectrum of speech [2], which is spectrum of changes of spectral envelopes in speech.

One of the early techniques, which utilize modulation spectrum is RASTA processing [3], where band-pass filters with time constants of the order of a couple of hundred ms are applied to deal with harmful effects of linear distortion in the signal, and enhancing modulation frequency components with syllabic frequencies. Success of RASTA spurred further research. The TRAP technique [4] simply uses long temporal trajectories of spectral energies as inputs to frequency-localized nonlinear classifiers. Linear discriminant analysis (LDA), applied to derive filters directly from labeled speech data, yields impulse responses, which resemble temporal derivatives of Gaussian function, again emphasizing modulation frequency components in the neighborhood of 5 Hz [5]. Somehow surprisingly, the LDA-derived filters are almost identical at all carrier frequencies, suggesting appropriateness of frequency-invariant processing of information in speech. This means that similar information-carrying features in dynamic speech spectrum are being extracted at all frequencies, just as it is hypothesized that similar shapes of edges of objects are extracted at all positions in a picture in the first stages of image recognition.

Longer temporal contexts would be consistent with dominant temporal properties of spectro-temporal auditory cortical receptive fields (STRFs) [6]. Various emulations of STRFs with the long context emerged [7] for use in ASR.

To extract these frequency-invariant dynamic features, MRASTA [8] uses temporal derivatives of Gaussian functions with varying variance, thus effectively extracting various components of the modulation spectrum. Similarly, the bottleneck features [9] use projections of 300 ms long temporal trajectories of spectral energies in individual critical bands on Hamming window weighted cosine bases as inputs to subsequent DNN, again implying various band-pass filters on the modulation spectrum.

In this paper, we experiment with a similar setup where the Hamming window weighted cosine bases are used as modulation frequency filters. The question we are addressing in this paper is to what extent are such filters optimal as pre-processing steps in deriving information from spectral dynamics. To answer such question, the modulation filters are learned directly from acoustic data by training their impulse responses using the DNN training framework. We report improved recognition performance obtained with the learned filters. We also analyze the frequency response of the filters.

2. Temporal context in DNN training

Many state-of-the-art speech recognition systems use Deep Neural Network (DNN) to map speech features frame-by-frame into phone (or phone state) posteriors, which are in turn used in HMM based decoder to decode phoneme or word sequences. The speech features usually used with the DNN classifiers are MFCC [10] or the log Mel filter bank outputs. Typically, features from some temporal context around the current frame are taken as the DNN input to predict a phone. Traditionally, to incorporate the information about the temporal context in GMM-HMM based recognizers, raw features were augmented with the temporal derivatives (delta and acceleration coefficients). However, these coefficients represented rather short context of few frames (few tens of millisecond). In current DNN-HMM based systems, much longer context up to several tens of frames (hundreds of miliseconds) is used. All the frames from the required context can be simply stacked into one vector to form the DNN input. However, this results in very high-dimensional vector with lots of redundant and noisy information. Typically, some dimensionality reduction is applied to ”smooth” the information encoded in the time context.

2.1. Baseline system

Figure 1 shows the scheme of our baseline system, where the DNN input is derived from 40 log Mel filter bank outputs. The temporal context of 31 frames (current frame ± 15 frames) is considered. For each of the 40 frequency bands, the 31 point temporal trajectory of the log Mel filter bank output is projected into 16 Hamming window weighted DCT bases (see first row of

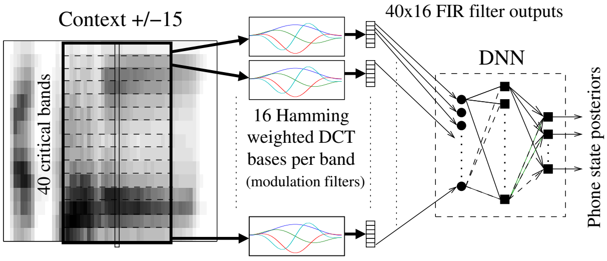

Figure 1: *Scheme of our baseline system.*

Figure 2) resulting in $40 \times 16 = 640$ dimensional DNN input. This (or similar) processing has proven to provide state-of-the-art performance for many tasks in our earlier experiments [11]. The rationale behind using the Hamming window weighted DCT bases was as follows: The Hamming window makes DNN to pay more attention to (or to represents in greater detail) the information around the current frame, for which DNN tries to predict the phone state class. The DCT bases are used to ”smooth” the temporal trajectories log filter bank outputs over time by discarding information about their fast changes. At the same time DCT makes the resulting coefficients more decorrelated, which may be beneficial for DNN training [12].

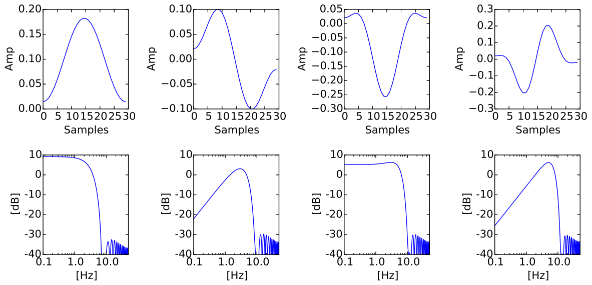

Figure 2: *4 DCT bases with their corresponding frequence responses.*

3. Learning modulation frequency filters

Now, let us see the frame-by-frame log filter bank outputs of one band as samples of a signal. As described above, we project 31 points of such signal (around the current frame) into the Hamming-DCT bases. Since we repeat this projection for every frame, we can see this operation as a convolution between the signal and the bases. In other words, we can see the bases as impulse responses of FIR filters operating on the temporal trajectories of log filter bank outputs. Such filters are known as *modulation frequency filters* as they shape temporal modulation of energies in individual frequency bands. In the second row of Figure 2, we can see the frequency responses of the modulation filters, which are all low-pass or band-pass filters discarding the fast temporal changes in the log filter bank outputs (i.e. discarding high modulation frequencies). In our case, the same set of modulation filters is used for all the frequency bands.

In our baseline system, the modulation filters impulse responses are precomputed and fixed as the Hamming-DCT bases. However, in the framework of DNN training, the modulation filters can be trained together with the other DNN parameters. In fact, the modulation filters impulse responses can be represented as an additional linear layer in the DNN, where the matrix of weights has a sparse structure. In the experiments described in section 5, the modulation filters are trained in such a way that the same set of filters is trained for each frequency band. This is achieved by sharing the weights (accumulating the back-propagated gradients) of the corresponding filters operating on different frequency bands.

4. Experimental setup

In our experiments, we use the baseline system described in section 2.1. The DNN has 640 dimensional input and 135 outputs corresponding to phoneme state posterior probabilities (3 states for each of 45 phonemes). It consists of 3 hidden layers with 2500 neurons in each layer. The *Sigmoid* activation functions are used in hidden layers. *Softmax* is used for the output layer.

The DNN is trained using stochastic gradient descent to minimize frame-by-frame cross-entropy with weights randomly initialized (including the modulation frequency filter parameters to be trained). For all experiments, the same training strategy and learning rate scheduler (*Newbob*) [13] is applied. We report results in terms of phone recognition accuracy and frame classification accuracy (with phoneme states as classes).

The experiments were carried out on a proprietary database containing American English read and spontaneous speech data with medium background music and babble noise. The training set consists of 6 hours of transcribed speech from 32 speakers. One hour (10 speakers) is reserved for cross-validation set and one hour (10 speakers) for the test set. Data were manually transcribed on the word level and the phoneme state level transcriptions were obtained using forced alignment with a GMM-HMM based LVCSR system [14]. To facilitate evaluation of progress on such proprietary database, we compare our new results to the results from our baseline speech recognizer described in section 2.1.

As described before, the modulation filters can be represented by an aditional layer prepended to the DNN. In such configuration, it is necessary to modify the learning rates for the first two layers (the new layer with filters and the first layer of the original DNN). Nonlinearity between layers limits values on the neuron outputs to a fixed range. The absence of the limiting nonlinearity can cause oscillations during the training as the gradients of weights connected through the linear units are highly correlated (multiplying weights in one layer by scalar and in the other layer by its inverse leads to an equivalent solution). We scaled the learning rate on the first and the second layer by factor 0.01 and 0.1, respectively.

5. Results

In our baseline system described in section 2, 16 Hamming-DCT modulation filters were used with temporal context of 31 frames. In the following expepriments, we also experiment with different numbers of modulation filters starting from a single filter (first column of Fig. 2) going up to 32 Hamming-DCT bases. Next, we experimented with systems, where the modulation filters were trained together with the parameters of the DNN. Again, we experimented with different numbers of trained modulation filters (up to 31).

On the left of Fig.4, we can see frame accuracies on the cross validation set for different tested systems. When comparing systems with the same number of modulation filters, the trained filters always outperform fixed Hamming-DCT bases. On the right, we can see phoneme accuracies on the test set. It is obvious, that trained temporal filters outperforms the base-

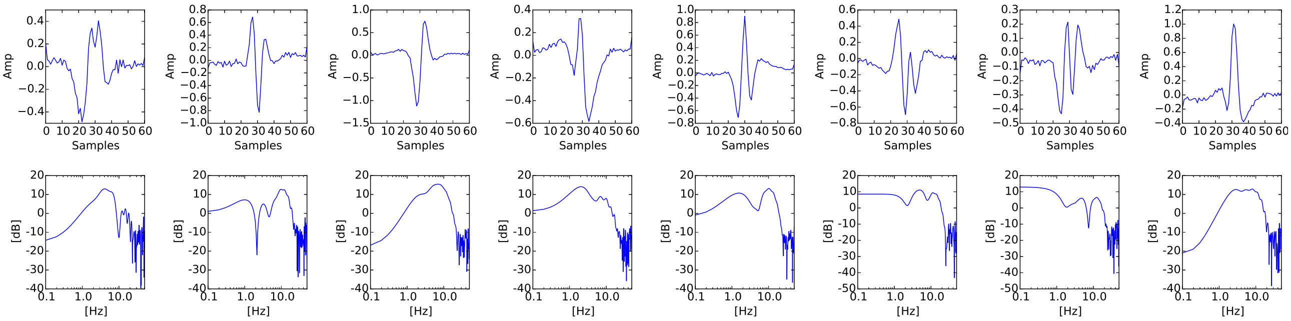

Figure 3: *Frequency and impulse responses of 8 derived filters*

line configuration. For modulation filters set to Hamming-DCT bases, the best performance is obtained with full set of 31 filters, where the frame accuracy is 44.0% and phoneme accuracy is 55.5% The trained modulation frequency filters perform the best with frame accuracy 44.1% and phoneme accuracy 55.7% with only 12 filters as can be seen on Fig. 4.

5.1. Extended time context

After the successful experiment with the 31 frames context, we tried to extend context size to 61 frames (30 frames on both sides around the central frame). In Fig. 5, we can see that the difference in performance between Hamming-DCT filters and modulation frequency trained filters is even larger. We can also see that for the Hamming-DCT filters, the performance with the context 61 frames is slightly worse compared to the context of 31 frames. With the same number of filters, the trained modulation filters again always perform better compared to the Hamming-DCT based filters.

For the Hamming-DCT based filters,the best performance is obtained with 28 filters, where frame accuracy is 41.1% and phoneme accuracy is 54.9%. However, such system is outperformed by the one with only 8 trained modulation filters with frame accuracy of 42.5% and phoneme accuracy 56.1%. Note that, for the trained modulation filters, extending the context from 31 frames to 61 frames helps to improve the phoneme recognition performance. On the other hand, the frame classification performance is slightly worse.

6. Analysis of learned filters

We choose time context of 61 frames for the further analysis of learned filters behavior. Improvements in recognition accuracy are important but even more interesting could be to see which solutions the DNN optimization came with. Fig 3 shows impulse and frequency responses of 8 FIR DNN-derived modulation filters. All filters significantly attenuate modulation frequencies that are higher than about 20 Hz. Most filters also suppress modulation frequencies that are lower than 1 Hz. Such band-pass character is consistent with sensitivity of hearing to sound modulations [15], and has been also observed in LDA designed FIR modulation filters [5].

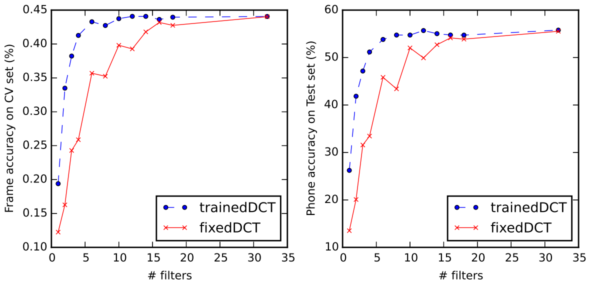

Figure 4: *Results with 31 frames of temporal context.*

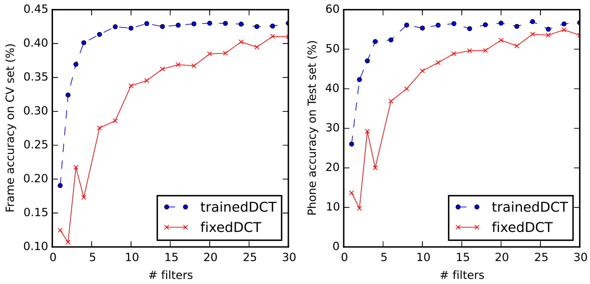

Figure 5: *Results with 61 frames of temporal context.*

To gain some insight to relative importance of various DNN-designed filters, we also show impulse and frequency responses of smaller set of filters. When only one filter is learned, such a filter is low-pass, computing weighted average of spectral values in 61 frames and passing modulation frequencies lower that 14 Hz as in Fig 6. The phone recognition accuracy obtained with the single filter at each frequency is only 26%. With two filters, the recognition accuracy increases to 42.3%. The filters are band-pass, suppressing modulation frequencies below 1 Hz see Fig 6. Three filters in Fig 7 yield 47.1% accuracy and we again observe one low-pass and two band-pass filters.

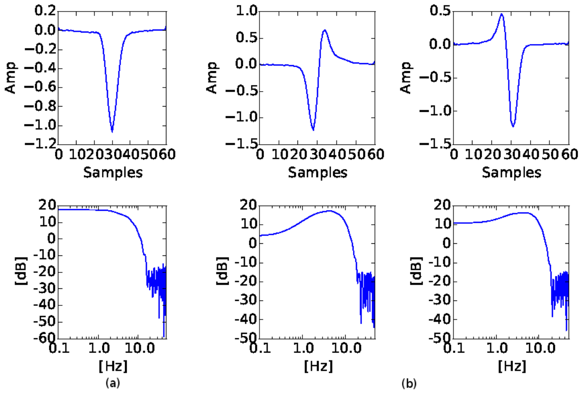

Figure 6: Frequency and impulse responses of (a) 1 filter, (b) 2 filters.

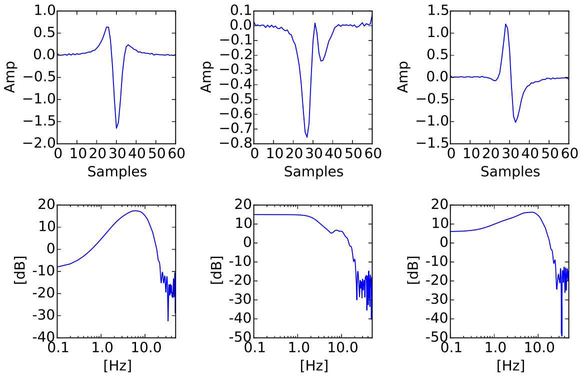

Figure 7: *Frequency and impulse responses of 3 derived filter.*

7. Conclusion

DNN paradigm was successfully used for design of modulation frequency FIR filters. The technique optimizes the whole process of deriving posterior probabilities of speech sound classes (three-state phonemes). The FIR filters of interest represent the first linear stage of the DNN. The filters have dominantly band-pass character, suppressing modulation frequencies higher that 20-25 Hz and lower than 1 Hz. Such filters are qualitatively consistent with FIR filters obtained by LDA, with human sensitivity to modulations, and with observed temporal properties of mammalian auditory cortical receptive fields. When applied in the state-of-the-art ASR system, where they replace the previously used cosine-based filters, the new filters consistently yield higher recognition accuracies. This tendency is the most pronounced when smaller numbers of filters are being used, supporting high efficiency of the derived FIR modulation filters.

8. Acknowledgments

This work was supported by the *DARPA RATS Program* under Contract No. HR0011-15-C-0038. The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

The work was also supported by *Technology Agency of the Czech Republic* project No. TA04011311 "MINT", *IT4Innovations Centre of Excellence* CZ.1.05/1.1.00/02.0070, *Intelligence Advanced Research Projects Activity* (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0013 and by *Google faculty award* to Hynek Hermansky

9. References

[1] H. Hermansky, "Speech recognition from spectral dynamics," *Sadhana*, vol. 36, no. 5, pp. 729–744, 2011.

[2] T. Houtgast and H. J. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," *Acta Acustica united with Acustica*, vol. 28, no. 1, pp. 66–73, 1973.

[3] H. Hermansky and N. Morgan, "Rasta processing of speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 4, pp. 578–589, 1994.

[4] S. R. Sharma, "Multi-stream approach to robust speech recognition," Ph.D. dissertation, 1999.

[5] S. van Vuuren and H. Hermansky, "Data-driven design of rasta-like filters." in *Eurospeech*, 1997.

[6] N. Mahajan, N. Mesgarani, and H. Hermansky, "Principal components of auditory spectro-temporal receptive fields," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[7] M. Kleinschmidt, D. Gelbart *et al.*, "Improving word accuracy with gabor feature extraction." in *INTERSPEECH*, 2002.

[8] H. Hermansky and P. Fousek, "Multi-resolution rasta filtering for tandem-based asr," IDIAP, Tech. Rep., 2005.

[9] K. Vesely, M. Karafiát, F. Grezl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 336–341.

[10] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech & Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[11] P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. dissertation, 2009. [Online]. Available: http://www.fit.vutbr.cz/research/view_pub.php?id=9132

[12] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 9–48.

[13] "Quicknet [online]," http://www1.icsi.berkeley.edu/Speech/qn.html.

[14] F. Grezl, M. Karafiát, and L. Burget, "Investigation into bottle-neck features for meeting speech recognition." in *INTERSPEECH*, 2009, pp. 2947–2950.

[15] R. Riesz, "Differential intensity sensitivity of the ear for pure tones," *Physical Review*, vol. 31, no. 5, p. 867, 1928.