FILTER-BANK OPTIMIZATION FOR FREQUENCY DOMAIN LINEAR PREDICTION

Vijayaditya Peddinti[†] *and Hynek Hermansky*[†‡]

[†]Center for Language and Speech Processing,
[‡]Human Language Technology Center of Excellence,
Johns Hopkins University, USA
{vijay.p,hynek}@jhu.edu

ABSTRACT

The sub-band Frequency Domain Linear Prediction (FDLP) technique estimates autoregressive models of Hilbert envelopes of sub-band signals, from segments of discrete cosine transform (DCT) of a speech signal, using windows. Shapes of the windows and their positions on the cosine transform of the signal determine implied filtering of the signal. Thus, the choices of shape, position and number of these windows can be critical for the performance of the FDLP technique. So far, we have used Gaussian or rectangular windows. In this paper asymmetric cochlear-like filters are being studied. Further, a frequency differentiation operation, that introduces an additional set of parameters describing local spectral slope in each frequency sub-band, is introduced to increase the robustness of sub-band envelopes in noise. The performance gains achieved by these changes are reported in a variety of additive noise conditions, with an average relative improvement of 8.04% in phoneme recognition accuracy.

Index Terms— cochlear filters, spectral differentiation, robust speech recognition

1. INTRODUCTION

The FDLP technique ([1]) is used to extract the amplitude modulation (AM) component of a signal using autoregressive estimates of Hilbert envelopes, computed by linear prediction on the discrete cosine transform (DCT) of the signal. Sub-band FDLP ([2]), extends this technique to estimate the Hilbert envelopes of the sub-band signals, using windowed DCT coefficients. These sub-band envelope estimates stacked horizontally provide a time-frequency decomposition of the signal (reminiscent of the spectrograms computed by the Spectrograph™), as an alternative to the conventional short-term Fourier transform derived spectral representation.

A series of efforts have been made to identify the optimal FDLP parameters, for speech and speaker recognition tasks. Important parameters in the FDLP feature extraction framework are the filter-bank, characterized by the shape, bandwidth and number of windows, and the pole order. Effect of these parameters on the performance of FDLP features has been explored previously, for reverberant environments. Among such efforts, Thomas *et al.* [3], show that uniform windows on linear axis (96 bands) are optimal. Mallidi *et al.* [4], further explored the effect of model order, envelope expansion factor and bandwidth of the DCT windows,in the linear filter bank, on word recognition accuracy in reverberant environments. However no effort has been so far made either to identify the optimal shape of cosine transform widows, that determine implied signal filtering, or to optimize FDLP performance in noise. In this work we attempt to fill this void.

Gaussian cosine transform windows, applied in FDLP so far, are very crude emulations of known properties of cochlear filtering in human hearing. Filter shapes, which more closely emulate the filtering in human cochlea might be preferable. A variety of cochlear filter banks have previously been used for improving robustness of processing in noise ([5],[6],[7],[8]).

Distortions, which often result in constant shifts of FDLP-estimated spectro-temporal envelopes (e.g., additive noise) or their logarithmic counterparts (e.g., convolutive distortions), can be tackled by filtering these envelopes. These techniques include temporal modulation spectrum filtering operations like RASTA ([9]) and simple filtering operations along frequency axis ([10], [11], [12], [13], [14], [15]). Gain normalization (GN) [16] and dynamic compression [17], which are temporal modulation filtering operations used in the FDLP framework, have been reported to be effective on additive ([17]) and convolutive ([4]) distortions. In this paper we introduce spectral differentiation, which is a high pass filtering operation along spectral axis, into FDLP to further increase robustness of this representation.

In this paper a cochlear filter-bank is introduced into the FDLP framework. Broader bandwidths, asymmetric shapes and an over-sampled frequency axis are some of the characteristic features of these cochlear filters. Each of these features is applied to the filter-bank and corresponding changes to the phoneme recognition accuracy are reported. Further, spectral differentiation is also introduced. Optimal parameters of the cochlear filter-bank for speech recognition in noisy environments are identified.

The organization of this paper is as follows: Section 2 briefly outlines the sub-band FDLP technique. Section 3 details cochlear window design. Section 4 describes the experimental setup. Finally, results and conclusions are presented in the Sections 5 and 6 respectively.

2. FILTER-BANKS FOR SUB-BAND FDLP

Let **x** be a real column vector that represents a discrete signal $x[n]$ of length N. Let **D** represent the DCT matrix whose elements are defined by

$$D[k,n] = a(k)cos(\frac{\pi(2n-1)(k-1)}{2N}) \quad k,n = 1,2,...,N \quad (1)$$

The research presented in this paper was partially funded by the DARPA RATS program under D10PC20015, IARPA BABEL project under W911NF12-C-0013 and the JHU Human Language Technology Center of Excellence. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reect the views of DARPA, IARPA or the JHU HLTCOE.

$$\text{where } a(k) = \begin{cases} \frac{1}{\sqrt{N}} & k = 1 \\ \sqrt{\frac{2}{N}} & 2 \leq k \leq N \end{cases}$$

For this unitary DCT matrix we have $\mathbf{D}^{-1} = \mathbf{D}^T$. The vector $\mathbf{Dx}$, represents the DCT coefficients of the signal $\mathbf{x}$.

Let $\mathbf{w}$ represent a column vector, defining a window in the DCT domain. The operation $\mathbf{Dx} \odot \mathbf{w}$, where $\odot$ represents the Hadamard (element-wise) product, corresponds to the symmetric convolution of the signal $x[n]$ with the sequence $\mathbf{D}^{-1}\mathbf{w}$. This symmetric convolution operation can be interpreted as circular convolution of a linear phase filter $h_w[n]$, formed after implicit left-sided symmetric extension of $\mathbf{D}^{-1}\mathbf{w}$, with the symmetrically padded data sequence $x[n]$ [18].

Thus, to implement a filter in the DCT domain, it is sufficient to compute the Hadamard product of the DCT of the signal, with the filter's frequency response, computed at corresponding frequencies [19]. The resulting product corresponds to the filtered signal, exactly in case of a linear phase filter and approximately for a non-linear phase filter. Hence any filter-bank can be emulated in the sub-band FDLP technique by designing windows which represent the magnitude responses of the filters in the filter-bank. The problem of filter-bank design is transformed to a window design problem, with shape, position and number of windows as the design parameters.

In the following sections a filter bank with the desired magnitude responses is implemented in the sub-band FDLP technique through the design of windows. The parameters of these windows are optimized for robust phoneme recognition. The phoneme recognition setup is described in Section 4.

Figure 1 summarizes the sub-band FDLP technique and the feature extraction procedure. In the first stage, the signal $x[n]$ is transformed to the DCT domain. The DCT of the signal is then windowed, using a choice of windows which emulate a desired filter-bank. Spectral differentiation, if performed, is done by subtracting the corresponding DCT components from consecutive windows. The windowed DCT components are then used used to estimate the Hilbert envelope of the sub-band signal, which is gain-normalized. The sub-band envelopes are integrated in 25 ms frames with a hop of of 10 ms, to obtain the sub-band energy representation. These outputs are then compressed and downsampled (along frequency axis) as described in Section 4, to obtain the final feature representation presented to the neural network.

The proposed method changes the sub-band decomposition block (Stage I) and post-processing block (Stage II) of the FDLP pipeline.

3. EMULATING A COCHLEAR FILTERBANK

Sub-band FDLP technique has so far used Gaussian windows to optimize the spectro-temporal properties for analysis. Athineos *et al.* [20] argue that better spectral auto-correlation estimates, derived from use of these windows, lead to better estimates of the Hilbert envelopes. These windows, emulating the mel filter-bank, consist of overlapping Gaussian windows, with a variance of 1 on the mel scale. We will refer to this set of windows here as $\mathbf{\Gamma}$.

Cochlea-like filtering with broader and asymmetric magnitude responses (as implemented e.g. in [21]) has been shown to provide robustness under mismatched conditions in speech processing applications. In this section, a set of asymmetric windows are designed to emulate the cochlear filter bank. The shape, bandwidth and number of these windows is experimentally selected for optimal performance in noisy conditions. Further, differentiation along frequency axis ([10],[11],[12]) was implemented to improve performance with these asymmetric windows.

Asymmetric windows: Compared to mel filter bank, the key difference in a cochlear-like filter bank is the highly asymmetric magnitude response of the filters. Hence windows emulating these filters, would have asymmetric shapes. To allow for systematic optimization, it is desirable to provide a parametric representation of such asymmetric windows, that can be controlled by these explicit set of parameters. Hence we use windows defined in perceptual linear prediction (PLP) technique ([21]), for spectral integration. These windows are piece-wise continuous and asymmetric, with parameters to control the lower frequency decay, higher frequency decay and the bandwidth of the filter. These piece-wise continuous windows are defined on the Bark frequency scale, Ω, as

$$\Psi(\Omega;\Omega_c) = \begin{cases} 10^{\alpha(\Omega-\Omega_c+\Omega_w/2)} & \Omega-\Omega_c \leq -\Omega_w/2, \\ 1 & -\Omega_w/2 < \Omega-\Omega_c < \Omega_w/2, \\ 10^{-\beta(\Omega-\Omega_c-\Omega_w/2)} & \Omega-\Omega_c \geq \Omega_w/2, \end{cases} \tag{2}$$

where Ω_c is the center frequency of the current window, α and β are the lower frequency and higher frequency steepness factors and Ω_w is the width of the flat top. The Bark-Hertz transformation used here is ([22])

$$\Omega(\omega) = 6ln\{\omega/1200\pi + [(\omega/1200\pi)^2 + 1]^{0.5}\} \tag{3}$$

Equation 2 is used to define a set of windows $\mathbf{\Psi}(\Omega;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Omega_w})$ at various center frequencies Ω_c, where $\boldsymbol{\alpha},\boldsymbol{\beta}$ or $\boldsymbol{\Omega_w}$ could also be functions of Ω_c. The parameters $\boldsymbol{\Omega_w}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ of these windows are determined to emulate a cochlear filter bank. The set of windows $\mathbf{\Psi}(\Omega;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Omega_w})$ is represented by the shorter form $\mathbf{\Psi}$, throughout this paper.

To emulate the heavy asymmetry towards the lower frequencies in typical cochlear windows (e.g., [13]) the α parameter in the filters ($\mathbf{\Psi}$) is reduced exponentially with Ω_c. Thus the filters centered at higher frequencies have a wider lower frequency spread than filters at the lower frequencies, on the bark frequency scale.

Performance of asymmetric windows, $\mathbf{\Psi}$, is compared to the symmetric windows ,$\mathbf{\Gamma}$, in the phoneme recognition task, described in Section 4. The Gaussian windows were distributed one per mel, with variance of 1 mel. The asymmetric windows were designed to match the number of Gaussian windows. The value of Ω_w is chosen as 0.8 bark, with exponentially decreasing α and constant β of 2.5. Figure 2 compares the windows from $\mathbf{\Psi}$ and $\mathbf{\Gamma}$ for a center frequency of ω_c.

Table 1 compares the average performance of these two windows for 5 different noise conditions, at 5 different SNRs. It can be seen that the cochlear windows ($\mathbf{\Psi}$) perform comparatively better than the Gaussian windows ($\mathbf{\Gamma}$). The average relative increase in accuracy, across all conditions at all SNRs, on the use of cochlear windows is 5.38%.

Spectral differentiation: Spectral differentiation (SD), which is a difference operation along spectral axis, highlights the rapid discontinuities along the spectral axis, and alleviates constant trends in the spectrum. Initial experiments applying spectral differentiation on Gaussian windows resulted in severe drop in the performance, hence the asymmetric windows were chosen for its application. It is implemented as a difference of the windowed DCT ($\mathbf{Dx} \odot \mathbf{w}$), for neighbouring windows ($\mathbf{w}$) in $\mathbf{\Psi}$. It is to be noted that features extracted from sub-band envelopes, generated after spectral differentiation, were used as an independent feature stream and not combined with non spectral-differentiated features.

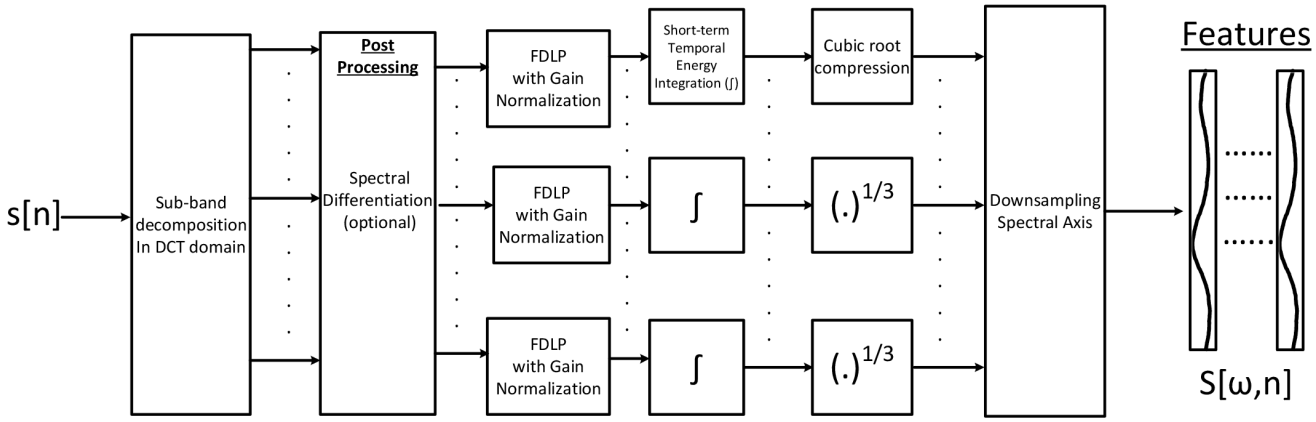

Fig. 1. *Block Schematic of sub-band FDLP technique*

Table 1. *Average Phoneme Recognition accuracy across 5 noises, at 5 different SNRs (in %)*

SNR	Filters				
	$\mathbf{\Gamma}$ (47 filters/8KHz)	$\mathbf{\Psi}$ (47 filters/8KHz)	$\mathbf{\Psi}$ + SD (47 filters/8KHz)	$\mathbf{\Psi}_{best}$ ($\Omega_w = 0.2$,3 filters/bark)	Relative Change in accuracy ($\mathbf{\Psi}_{best}$ vs $\mathbf{\Gamma}$)
clean	71.4	69.3	67.6	69.7	-2.38
20	59.62	61.03	61.96	63.23	6.06
15	51.95	54.46	55.75	56.55	8.85
10	42.91	45.34	46.96	47.19	9.97
5	34.34	36.14	37.55	37.69	9.74
0	27.77	29.01	30.28	30.58	10.10

From Table 1, it can be seen that SD is helpful in many conditions when performed on cochlear windows. The performance of cochlear windows (with SD) is better than the symmetric windows, in all noise conditions, with an average relative increase of 8.03%, across all noise conditions. However there was a drop in the performance in clean condition. On analysis of the phoneme confusion matrices, corresponding to the two types of windows, it was observed that stop consonants were the worst affected phone class, in the clean condition, on the use of cochlear windows. This was attributed to the loss of temporal resolution due to the use of cochlear windows with discontinuities. Currently efforts are being made to verify if the use of smoother asymmetric windows can alleviate this loss.

The bandwidth and number of cochlear windows was varied, to optimize for robustness, while reducing the effect on performance in clean condition. Bandwidth of the filters implied by these windows, is controlled by the flat top width (Ω_w) of the window, since Ω_w is related to the bandwidth of the filter by an additive constant, for a given α and β (see Equation 2). Highest recognition accuracies were observed at 3 filters/bark (63 filters/8KHz) with 0.2 bark flat width. However, the number of filter bank outputs was reduced to 1 filters/bark, after performing the spatial differentiation and gain normalization operations at higher frequency resolution, to reduce the dimensionality of the input feature vector.

Shapes of the windows $\mathbf{\Psi}$ for various center frequencies (ω_c), with $\alpha(\omega_c)$ decreasing exponentially, $\beta(\omega_c) = 2.5$ and $\Omega_w(\omega_c) = 0.2$ are shown in Figure 3.

4. EXPERIMENTAL SETUP

Database: Experiments were done using TIMIT database and the phoneme recognition system is trained on clean speech sampled 16 KHz. The training data consists of 3400 utterances from 475 speakers, cross-validation data set consists of 296 utterances from 37 speakers and the test data set consists of 1344 utterances from 168 speakers. The TIMIT database, which is hand-labeled using 61 labels is mapped to the standard set of 39 phonemes.

TIMIT test set corrupted with five different noise conditions from NOISEX-92 database [23] (factory floor noise (I), speech babble noise, fighter jet (F16) cockpit noise, car interior (Volvo 340) noise and military tank (leopard) noise), forms the development set to optimize the parameters of the asymmetric filter-bank. 11 different noise conditions (benz, buccaneer(I), buccaneer(II), car, destroyer-ops, exhibition-hall, factory(II), m109, restaurant, street and subway), not used in optimization, form the test data in the final evaluation.

Back-End: The phoneme recognition system is based on Hidden Markov Model - Artificial Neural Network (HMM-ANN) paradigm [24]. The system uses a MLP-based hierarchical phoneme posterior estimator ([25]) which generates the posteriors used for phoneme decoding. The hierarchical phoneme posterior estimator consists of a cascade of two three-layer MLPs, each with 1500

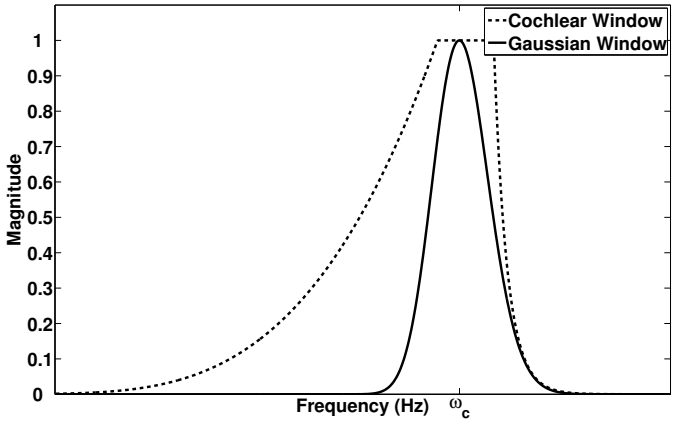

Fig. 2. *Symmetric (Gaussian) and Asymmetric (cochlear) windows for particular center frequency ω_c, warped to the linear frequency scale ($\boldsymbol{\omega}$)*

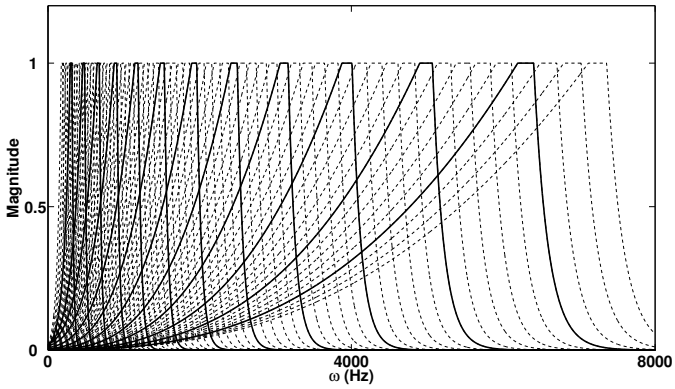

Fig. 3. *Shapes of windows $\boldsymbol{\Psi}$ with variable α, $\beta = 2.5$, $\Omega_w = 0.2$ bark and with 3 filters/bark (on a linear frequency scale)*

hidden neurons and 40 output neurons, representing the phoneme classes (including silence). The two ANNs are trained using the standard back propagation algorithm with cross entropy as the training error. For the decoding step, all phonemes are considered equally probable (no language model). The performance of phoneme recognition is measured in terms of phoneme accuracy (excluding silence). The HMM-ANN recognition system was trained only on the original clean train set and tested on the clean and noisy versions of the test. This training procedure helps test the robustness of proposed feature extraction method.

Front-End: Mohamed *et al.* ([26]) successfully argue that logarithmic spectral features are preferred over cepstral representations as inputs for neural nets, as the discriminative information is spread over all the coefficients. Further, cubic compressed spectra were found to perform better than log compressed spectra [21]. Hence cubic compressed spectral energies from short term integrated subband envelopes, along with their first and second temporal differentials, were used as input feature vector. The first MLP was trained with a temporal context of 9 frames. The input feature vector for the second MLP was formed from the posterior estimates generated by the first MLP with a temporal context of 23 frames ([25]).

5. RESULTS

Cochlear windows $\boldsymbol{\Psi}$ (with spectral differentiation (SD), 3 windows/bark and Ω_w=0.2) were compared with the Gaussian windows $\boldsymbol{\Gamma}$ (1 window/mel and variance=1 mel). Average phoneme recognition accuracy was measured, in 11 different noise conditions from NOISEX-92 database [23] at 5 different SNRs. These noise conditions do not overlap with the previous conditions, and form an unseen test set. The average relative increase in recognition accuracy was 8.04%, across all these conditions.

Average phoneme recognition accuracies at various SNRs are tabulated in Table 2.

Table 2. *Average Phoneme Recognition accuracy across 11 noises at 5 different SNRs (in %)*

SNR (dB)	Window Type		Relative Change
	$\boldsymbol{\Gamma}$	$\boldsymbol{\Psi}$+SD	in accuracy
∞	71.35	69.66	-2.37
20	59.56	62.46	5.04
15	51.96	55.68	7.16
10	42.94	46.66	8.66
5	33.85	37.03	9.39
0	25.88	28.61	10.55

Maximum gains due to the spectral differentiation process were observed in band-limited noises such as car, benz and volvo. Further, it was observed that gains were consistent, especially at lower SNRs. Reduced accuracy in the clean condition, as highlighted before, was due to the recognition errors in consonants, especially plosives. Vowels on the other hand showed increase in recognition accuracy, even in the clean condition.

6. CONCLUSION AND FUTURE WORK

In this paper an asymmetric cochlear filter-bank was designed for application in the FDLP framework. This asymmetric filter-bank ($\boldsymbol{\Psi}$) was shown to be suitable for application of spectral differentiation. Parameters of this filter-bank ($\boldsymbol{\Psi}$) were optimized for robustness to additive noise distortions, with minimal effect on performance in clean condition. This asymmetric filter-bank, $\boldsymbol{\Psi}$, provides an average relative improvement of 8.04% in phoneme recognition accuracy over the mel filter-bank ($\boldsymbol{\Gamma}$).

Pole-order and envelope compression factor were shown to be influential parameters in the FDLP extraction technique, in previous studies for reverberant environments. Optimal values of these parameters have to be identified for the asymmetric filter bank, in the additive noise condition. Further experiments are being conducted to verify if the gains observed are consistent when combined language models or in scenarios with multi-condition training.

The use of asymmetric filters and spectral differentiation brings FDLP processing closer to cochlear processing in humans, hence this representation is seen as an ideal candidate for operations such as modulation filtering, which emulate cortical processing stage in humans. Experiments on modulation filter-bank design using this new representation are being conducted.

7. REFERENCES

[1] Marios Athineos and Daniel P.W. Ellis, "Autoregressive Modeling of Temporal Envelopes," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5237–5245, Nov. 2007.

[2] Marios Athineos and Daniel P.W. Ellis, "Frequency-domain linear prediction for temporal features," in *Automatic Speech Recognition and Understanding, (ASRU'03). IEEE Workshop on*. 2003, pp. 261–266, IEEE.

[3] Samuel Thomas, Sriram Ganapathy, and Hynek Hermansky, "Hilbert Envelope Based Features for Far-Field Speech Recognition," in *Machine Learning for Multimodal Interaction*, Andrei Popescu-Belis and Rainer Stiefelhagen, Eds., number 2, pp. 119–124. Springer Berlin / Heidelberg, 2008.

[4] Sri Harish Mallidi, Sriram Ganapathy, and Hynek Hermansky, "Modulation spectrum analysis for recognition of reverberant speech," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[5] Oded Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," *Computer Speech & Language*, vol. 1, no. 2, pp. 109–130, Dec. 1986.

[6] M. Hunt and C. Lefebvre, "Speech recognition using a cochlear model," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, apr 1986, vol. 11, pp. 1979 – 1982.

[7] Qi Li and Yan Huang, "An Auditory-Based Feature Extraction Algorithm for Robust Speaker Identification Under Mismatched Conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1791–1801, Aug. 2011.

[8] R. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, may 1982, vol. 7, pp. 1282 – 1285.

[9] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis technique," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, mar 1992, vol. 1, pp. 121 –124.

[10] Shihab A. Shamma, "Speech processing in the auditory system II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve," *The Journal of the Acoustical Society of America*, vol. 78, no. 5, pp. 1622, Nov. 1985.

[11] Pratibha Jain and Hynek Hermansky, "Beyond a single critical-band in trap based asr," in *INTERSPEECH*, 2003.

[12] Frantisek Grézl and Hynek Hermansky, "Local averaging and differentiating of spectral plane for trap-based asr," in *INTERSPEECH*, 2003.

[13] 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, Mar. 1992.

[14] L. Rabiner and J. Wilpon, "On the use of bandpass liftering in speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 7, pp. 947–954, July 1987.

[15] S.K. Nemala, K. Patil, and M. Elhilali, "Multistream Bandpass Modulation Features for Robust Speech Recognition," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011, pp. 1277–1280.

[16] S. Thomas, S. Ganapathy, and H. Hermansky, "Recognition of Reverberant Speech Using Frequency Domain Linear Prediction," *IEEE Signal Processing Letters*, vol. 15, pp. 681–684, 2008.

[17] Sriram Ganapathy, Samuel Thomas, and Hynek Hermansky, "Temporal envelope subtraction for robust speech recognition using modulation spectrum," in *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*. Dec. 2009, pp. 164–169, IEEE.

[18] S. Martucci, "Symmetric convolution and the discrete sine and cosine transforms," *IEEE Transactions on Signal Processing*, vol. 42, no. 5, pp. 1038–1051, May 1994.

[19] B. Chitprasert and K.R. Rao, "Discrete cosine transform filtering," in *International Conference on Acoustics, Speech, and Signal Processing*, apr 1990, pp. 1281 –1284 vol.3.

[20] Marios Athineos, Hynek Hermansky, and D.P.W. Ellis, "LP-TRAP: Linear predictive temporal patterns," in *Proc. of ICSLP*. 2004, Citeseer.

[21] Hynek Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am*, vol. 87, no. April, pp. 1738–1752, 1990.

[22] M.R. Schroeder, "Recognition of complex acoustic signals," *Life Sciences Research Report*, vol. 5, pp. 324, 1977.

[23] A. Varga, H.J.M. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," Tech. Rep., DRA Speech Research Unit, Malvern, 1992.

[24] Hervé Bourlard and Nelson Morgan, *Connectionist speech recognition: a hybrid approach*, Springer, 1994.

[25] Joel Pinto, Sivaram Garimella, Mathew Magimai-Doss, Hynek Hermansky, and Hervé Bourlard, "Analysis of MLP-Based Hierarchical Phoneme Posterior Probability Estimator," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 225–241, Feb. 2011.

[26] A. Mohamed, G. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, march 2012, pp. 4273 –4276.