A UNIFIED FRAMEWORK FOR FILTERBANK AND TIME-FREQUENCY BASIS VECTORS IN ASR FRONTENDS

Xiaoyu Liu, Stephen A. Zahorian

Department of Electrical and Computer Engineering, Binghamton University, Binghamton, NY, 13902, USA

ABSTRACT

For many years, filterbanks have been widely used as one step of frontend feature extraction for Automatic Speech Recognition (ASR). In this paper, we propose a unified framework for ASR frontends, by first moving the nonlinear amplitude scaling, and then combining the filterbank weights with the cosine basis vectors. As part of this framework, we also show that the delta terms used to encode feature dynamics can also be viewed as one realization of a set of "unified" basis vectors over time. With this framework, frontends can be developed, analyzed and evaluated through their basis vectors over frequency and time.

Index Terms— Filterbank, spectro-temporal, unified, basis vector, frontend

1. INTRODUCTION

For many years, filterbanks, implemented as weighted sums of FFT magnitudes, are widely used as a frontend processing step for ASR systems. Figure 1(a) is a block diagram of the filterbank-based feature extraction approach. One commonly used version of this approach is to compute Mel Frequency Cepstral Coefficients (MFCCs) [1]. The MFCC features are computed using a set of triangular bandpass filters approximately logarithmically spaced above 1 kHz to map the short time power in the Hertz domain to the Mel domain. In recent years, various enhanced MFCC algorithms have been developed. In [2], a Smooth MFCC (SMFCC) algorithm incorporates the pitch frequency information in building the filterbank, and in [3], the spectral envelope of the voiced frames is enhanced to improve the noise-robustness of the MFCCs.

To extract features from the amplitude-scaled output of the filterbank, the Discrete Cosine Transform (DCT) is computed using "half" cosine multiple basis vectors. The feature calculation using these "regular" cosine basis vectors is given by equation (1) as:

$$DCTC(i) = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} a\big(P(j)\big) \cos\left(\frac{\pi i}{N}(j - 0.5)\right) \qquad (1)$$

where $DCTC(i)$ is the ith DCT coefficient, N is the total number of filter channels, $P(j)$ is the output power of the jth channel, and $a(\,)$ is the amplitude scaling function. The DCT coefficients are similar to the principal components of the spectrum. In [4], a Distributed DCT method is presented to remove the correlation between filterbank outputs more completely, which leads to a more compact set of cepstral features.

As pointed out in [5,6,7,8], the delta and acceleration terms of the DCTCs greatly help to improve the recognition accuracy since these time derivatives capture the dynamic behavior of adjacent coefficients. The delta terms are computed through equation (2), where Θ is the window length in frames, and higher order terms are the deltas of lower order ones.

$$\Delta(t) = \frac{\sum_{\theta=1}^{\Theta} \theta (DCTC_{t+\theta} - DCTC_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \qquad (2)$$

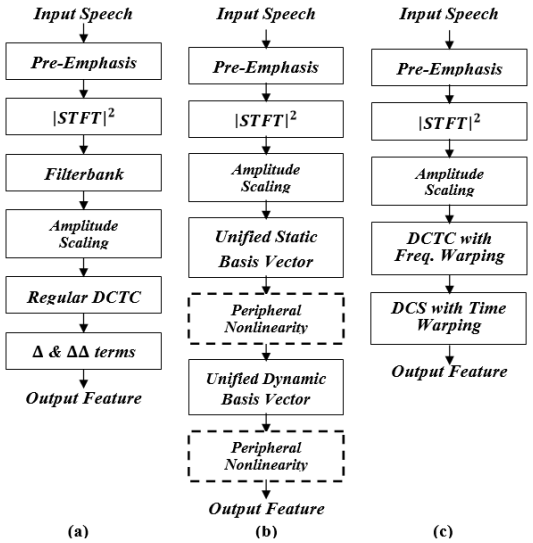

Fig.1. Block diagrams of the filterbank-type frontend (a), the unified structure (b), and the spectro-temporal system (c) in [9].

Spectro-temporal frontends provide much more detailed representations of the temporal patterns in speech than the time derivative terms. The work in [9] proposes non-uniform time resolution within time blocks of the static features using a set of Discrete Cosine Series (DCS) expansion, and in [10], parallel and hierarchical structures are developed based on a temporal filterbank and in [11], two-dimensional Gabor features are obtained to capture the diagonal spectro-temporal patterns.

In our work, we propose a unified framework for ASR frontends, which is built upon a set of unified time-frequency basis vectors. The nonlinear amplitude scaling is moved to immediately after the FFT magnitude step. Under this framework, frontend systems, such as (but not limited to) [9, 11], can be characterized entirely through the unified basis vectors, which gives a common yardstick for analyzing frontends. We also discuss other potential benefits of this perspective.

2. A UNIFIED FRAMEWORK

2.1. Moving the amplitude scaling to the 'front'

It's interesting to note that if we move the nonlinear amplitude scaling in Figure 1(a) to before the filterbank, the filterbank weights can then be combined with the "regular" half cosine basis vectors by a simple matrix multiplication. However, this modification should be justified by inherent auditory properties as well as ASR experiments.

Physiologically, different frequency components in a travelling wave cause maximum displacement of the basilar membrane at different positions. The membrane vibration "fires" the neurons through hair cells, and the firing rate as a function of sound intensities is modeled by the nonlinear amplitude scaling. A commonly used nonlinearity is the logarithmic compression, as in the Seneff model [12]. More sophisticated auditory models such as [13], indicate that this perceptual loudness mapping can be better approximated by a power-law function [14], and should be frequency-dependent due to the sensitivity of the hair cells. Directly mapping the original spectrum with the nonlinearity inherently eliminates this frequency distinction.

However, we place the nonlinearity before the filterbank since (1) frequency-independency simplification is widely made and experimentally justified by ASR systems, such as MFCC and PLP [15], which uses an equal-loudness curve to compensate for the simplification, (2) based on (1), there is no compelling evidence as to where the nonlinearity should be placed, (3) experimentally, we will show that it does not affect ASR performance much if the nonlinearity is moved to before the filterbanks, and (4), as discussed below it allows the system unification which brings benefits.

2.2. Unified basis vectors

First, with the amplitude scaling moved, it's straightforward to create a set of "unified" static basis vectors by a matrix product. Suppose the rows of the matrix $\boldsymbol{W}$ contain the filterbank channel response, and the rows of $\boldsymbol{BVF_{reg}}$ contain the regular cosine basis vectors, the unified version $\boldsymbol{BVF_{uni}}$ is given in Eq. (3), and the amplitude-scaled FFT spectrum is weighted by $\boldsymbol{BVF_{uni}}$ to obtain the static DCTCs.

$$\boldsymbol{BVF_{uni}} = \boldsymbol{BVF_{reg}}\boldsymbol{W} \tag{3}$$

In the standard MFCC framework, the dynamic (Δ) features are computed from the static DCTCs, using Eq.(2). Here we show that the Δ terms can also be computed using basis vector manipulations. From Eq.(2), to compute any nth order differential term, its basis vector with respect to the previous lower order (neglecting the constant denominator) is given by $\boldsymbol{bv}_n = [-\theta_n, -\theta_n + 1, \ldots, 0, 1, \ldots \theta_n]$, where θ_n is the window length. If we view $\boldsymbol{bv}_n$ as a discrete signal, with each element representing both the amplitude and the time index (i.e. [-2,-1,0,1,2] gives a signal whose magnitude is -2 at index -2, and -1 at index -1, etc.), then, the nth order basis vector with respect to the DCTCs (i.e. absolute time) can be computed as:

$$\boldsymbol{bvT_n} = \boldsymbol{bv_1} \circledast \boldsymbol{bv_2} \ldots \circledast \boldsymbol{bv}_n \tag{4}$$

where $\circledast$ is the convolution operator, and each $\boldsymbol{bv_i}$ is the ith order basis vector in terms of its previous lower order. Thus, putting all $\boldsymbol{bvT_n}$, including the zeroth order, into rows of a unified dynamic basis vector matrix $\boldsymbol{BVT_{uni}}$, the final feature matrix $\boldsymbol{F}$ at the output is given by Eq.(5), where $a(\boldsymbol{X})$ is the amplitude-scaled FFT spectrum.

$$\boldsymbol{F} = \boldsymbol{BVT_{uni}} \cdot [\boldsymbol{BVF_{uni}} \cdot a(\boldsymbol{X})]^T \tag{5}$$

2.3. Discussion

In this section, we present a detailed discussion on the significance/applications of this unified frontend perspective, whose block diagram is depicted in Figure 1(b).

First, it's important to note that $\boldsymbol{BVT_{uni}}$ and $\boldsymbol{BVF_{uni}}$ in Eq.(5) can take on any generalized forms, though they are derived from a specific category of frontends. On a higher level, Eq.(5) shows that features can be viewed as a series of linear transformations of the spectrum scaled by an auditory nonlinearity, with optional peripheral nonlinearities in between (dashed blocks in the diagram). These linear transformations are represented by the unified basis vectors. Filterbanks (or other parts) exert their impact on system quality by shaping the basis vectors implicitly. Thus, the unified basis vectors determine the time-frequency properties of a frontend. In this sense, the scheme gives us a common "yardstick" to analyze and compare frontends which appear to be different or similar based on the properties of the unified basis vectors.

The first example to illustrate this point is the comparison between the "standard" MFCC and the spectro-temporal system in [9], whose diagram is given in Figure 1(c). It's important to emphasize that in the unified framework, both systems compute features in a mathematically identical manner, and the only difference lies in the unified basis vector forms. In [9], in computing the DCTCs, the ith basis vector $\phi_i(f)$ over frequency f is given by Eq.(6):

$$\phi_i(f) = \cos[\pi i \cdot g(f)] \cdot \frac{dg}{df} \tag{6}$$

where $g(f)$ is a frequency warping function. In Figure 2, we plot the first 3 basis vectors (left) with $g(f)$ set to a Mel-shape warping (right), and in Figure 3, we also plot the first 3 unified basis vectors for MFCC using a 26-channel Mel filterbank.

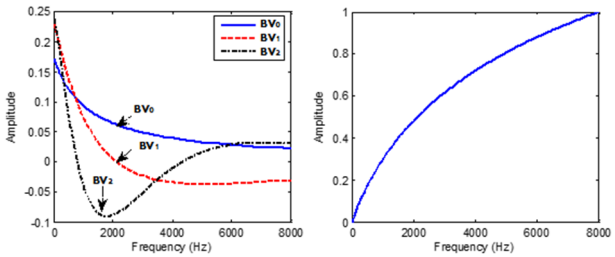

Fig2. First 3 DCTC basis vectors (left) with an embedded Mel-shape warping (right) in the system of [9].

The unified basis vectors produced by the Mel filterbank are less smooth than the ones generated by the Mel-shape warping. In Figure 2, the Mel scale is implemented in a continuous manner, with the envelope (BV_0) representing the frequency resolution *dg/df*; however, for the case shown in Figure 3, the basis vectors are computed using a 26 step quantized Mel scale, using filter bandwidth to represent frequency resolution. Thus, we might expect a finer frequency resolution characterization for

the continuous Mel-shape warping approach, which might lead to a higher recognition accuracy. However, the difference should be small, since they are essentially two ways of implementing a Mel scale, as shown by the similarities in the basis vectors.

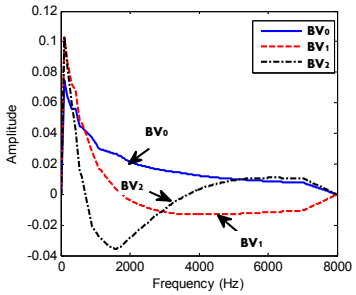

Fig.3. First 3 unified DCTC basis vectors for the standard MFCC frontend. A 26-channel Mel filterbank is combined with regular cosine basis vectors.

To obtain dynamic features, the system in [9] uses a set of Discrete Cosine Series (DCS) basis vectors to weight the time blocks of the DCTCs. The *i*th DCS basis vector is defined as:

$$\psi_i(t) = \cos[\pi i \cdot h(t)] \cdot \frac{dh}{dt} \qquad (7)$$

where $h(t)$ is a time warping function. Again, the first 3 DCS basis vectors are plotted in Figure 4 (left) with a continuous Kaiser window for the *dh/dt* term, and in the right panel, the first 3 differential basis vectors are presented, with BV_0 refers to the zeroth order in both cases.

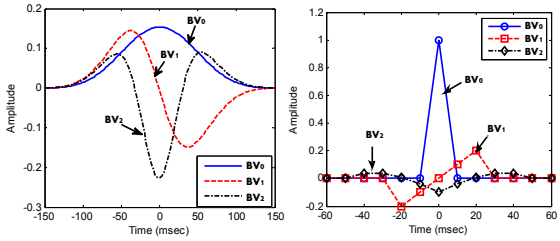

Fig.4. First 3 DCS (left) and differential (right) basis vectors in computing the dynamic features in system [9] and MFCC.

Clearly, the discrete differential (indicated by markers) and continuous DCS basis vectors are very different, both from their "look" and the logic used to derive them. However, as they are put into the same unified framework, we are able to analyze system properties through basis vectors. If we compare the zeroth order term, the DCS method encodes feature trajectories using a long segment of context. Speech frames near the current "observation point" are assigned more weight to determine the speech content, and those far from the center are smoothed. This relative importance of speech information is reflected by the time resolution, i.e. *dh/dt* in Eq.(7). However, the differential case uses only the block center term. Hence, the DCS method may provide better features for the time patterns in speech using non-uniform time resolution in long speech segments.

Another system that can be analyzed in the unified framework is found in [11,16], where a set of Gabor filters are proposed to capture the Localized Spectro-Temporal Features (LSTFs). However, the work of [17] shows that the directionality of LSTFs can be obtained through weighting the spectrum by a set of rotated basis vectors. Thus, the LSTFs can be studied and evaluated by the unified framework. In [17], the basis vectors are tuned toward different angles, and corresponding phonetic recognition results are obtained to evaluate the effects.

Potentially superior features can also be developed through the unified concept. As one example, the static and dynamic basis vector steps could be interchanged to allow the use of frequency-dependent dynamic basis vectors. The time resolution, i.e. *dh/dt*, built into the basis vectors for higher frequencies could use a more "peaky" window shape than that of lower frequencies, to allow higher time resolution. This modified framework accounts for the effects of the auditory time-frequency trade-off, revealed by psychoacoustic [18] and neurophysiological [19] known facts.

Empowered by the frontend unification approach, a higher level systematic unification can be envisioned, which will potentially push state-of-the-art speech recognition. For example, conceptually, a frontend should only compute static features, and the temporal patterns should be modeled by the recognizer. Indeed, human ears (the frontend) only do spectral analysis whereas higher levels of processing in the human brain (the recognizer) characterize the spectral-temporal information. Thus, it can be foreseen that modeling of the "hidden" spectral-temporal patterns can be exploited by the data-driven training of a state-of-the-art recognizer, such as a Deep Neural Network (DNN), which has the power of performing "deep learning."

Finally, there are limitations to the unified frontend in this work. It is not intended to replace specific frontends, nor even accounts for all of them (e.g. PLP). However, for many cases, it reveals the essence of features with a straightforward tool, the unified basis vectors, as a linear transformation. Possibly more effective systems can be developed. For frontends which might not fully fall into this structure, their system properties can still be studied with the view presented here. Also, the filterbank and the regular basis vectors can still be implemented in two separate steps as needed, to allow various techniques, such as the Power Normalized Cepstral Coefficient (PNCC) algorithms [14,20] to be inserted.

3. EXPERIMENTAL EVALUATION

The goal of this section is to present the system performance purely in terms of the unified basis vectors built from various filterbanks and the system in [9]. Extensive tests were also conducted to determine the effects of moving the nonlinearity.

3.1. Phonetic level recognition task

A 39 phoneme recognition task with TIMIT was conducted. 3696 and 1344 utterances (SA sentences removed) were used for training and testing respectively. 48 3-state monophone GMM/HMMs were trained by HTK 3.4, and a phonetic bigram language model was used for decoding. Throughout this subsection, the optimal frame length/space for frontends using differential dynamic basis vectors were 25ms/10ms respectively, and for other frontends using the DCS dynamic basis vectors, including the example spectral-temporal system in [9] were 8ms/2ms. The optimal block length/space for computing DCS were 302ms/8ms. 26 and 40 channels were used for Mel and gammatone filterbank derived basis vectors respectively. The gammatone was implemented as in [21].

In Table 1, we examine the effect of placing the amplitude scaling before the filterbank. 32 GMM mixtures were used. Logarithmic and power law functions were tested. The power exponent in the power law was 0.1. For the Mel and gammatone cases, the static and dynamic basis vectors were 12 regular cosine as in Eq.(1) (plus 1 log-energy) and delta/acceleration (39 features in total). For more thorough tests, the PLP frontend was also implemented, though the detailed analysis of this frontend was not done in our proposed framework. MATLAB code to obtain the PLP results can be found in [22], where 16 trapezoids were used as the filterbank with an equal-loudness curve built into the weights, and the power value was 0.33. 12 static terms were obtained from the LPC cepstral recursion. The dynamics were delta and acceleration. In the baseline cases (bolded), the amplitude scaling was placed after the filterbanks.

Table 1. Phonetic accuracy (%) of placing the amplitude scaling before/after the filterbanks

FB Type	Scaling Type	Scaling pos.	Accuracy
Mel	**log**	**After FB**	**69.8**
Mel	log	Before FB	69.6
Mel	**power (0.1)**	**After FB**	**69.8**
Mel	power (0.1)	Before FB	69.7
Gammatone	**log**	**After FB**	**70.3**
Gammatone	log	Before FB	70.1
Gammatone	**power (0.1)**	**After FB**	**68.9**
Gammatone	power(0.1)	Before FB	69.6
Trapezoids in PLP	**power (0.33)**	**After FB**	**70.0**
Trapezoids in PLP	power (0.33)	Before FB	69.9

Moving the amplitude scaling to before the FB results in only a negligible decrement in performance. Table 2 shows various combinations of static/dynamic basis vectors and numbers of dynamic terms. 13 static unified basis vectors were built with either filterbanks or a continuous Mel-shape warping. 96 GMMs were used. The baselines are again bolded. A logarithmic nonlinearity was placed before filterbanks.

Table 2. Phonetic accuracy (%) using different unified static/dynamic basis vectors

Static BV Type	Dynamic BV Type	Feature Num.	Accuracy
Mel+regular cosine	**delta, acceleration**	**39**	**71.4**
Gammatone+regular cosine	3 DCS	39	72.6
FFT+DCTC with Mel warp	3 DCS	39	72.4
Mel+regular cosine	**delta, acc. & third order**	**52**	**71.4**
Gammatone+regular cosine	4 DCS	52	73.5
FFT+DCTC with Mel warp	4 DCS	52	73.7
Gammatone+regular cosine	5 DCS	65	73.9
FFT+DCTC with Mel warp	5 DCS	65	74.2

First, with the same number of features, the combination of FFT+DCTC with Mel-shape warping and DCS cases are better than the bolded baselines (larger difference with 52 features). This is consistent with the finer frequency resolution reflected by the static basis vectors (compare Figure 2 and 3), and also better time resolution of the dynamic basis vectors (Figure 4). Also, note that adding more DCS basis vectors brings relatively significant improvements over the 39 feature cases, whereas more differential terms provide no improvements. This relative improvements again support the superiority of the non-uniform time resolution reflected in the unified dynamic basis vectors.

3.2. Word level recognition task

In this section, we report word (actually character) level recognition to confirm the findings with the phonetic experiments. 37116 utterances spoken by 78 women speakers from the 863 Mandarin Chinese database were used as training data (about 40 hours in total), and another 5 women speakers (3125 utterances) were used as a test data. 16-mixture cross-word triphones and a 5868-word bigram model were trained for decoding. Throughout this section, we use character percentage accuracy as the evaluation measurement.

In Table 3, we repeated the cases in Table 1 to further confirm the validity of moving the amplitude scaling. The setup parameters for the frontends were identical to those in Table 1. The baselines are bolded.

Table 3. Character accuracy (%) of placing the amplitude scaling before/after the filterbanks

FB Type	Scaling Type	Scaling Position	Accuracy
Mel	**log**	**After FB**	**86.1**
Mel	log	Before FB	86.6
Mel	**power (0.1)**	**After FB**	**85.1**
Mel	power (0.1)	Before FB	85.8
Gammatone	**log**	**After FB**	**85.8**
Gammatone	log	Before FB	87.0
Gammatone	**power (0.1)**	**After FB**	**83.3**
Gammatone	power (0.1)	Before FB	85.9
Trapezoids in PLP	**power (0.33)**	**After FB**	**85.9**
Trapezoids in PLP	power (0.33)	Before FB	86.5

These results strengthen the validity of moving the amplitude scaling. In Table 4, we present two pairs of comparisons with different static/dynamic settings. The optimal frame length/space for the DCS scenarios were 10ms/2ms (for the Mel+regular cosine case) and 25ms/2ms (for the FFT+DCTC with Mel warping case). The optimal block length/space of DCS were 142ms/14ms for both. A logarithm scaling was placed after the filterbanks. Baselines are bolded.

Table 4. Character accuracy (%) using different unified static/dynamic basis vectors

Static BV Type	Dynamic BV Type	Feature Num.	Accuracy
Mel+regular cosine	**delta, acceleration**	**39**	**86.1**
Mel+regular cosine	6 DCS	78	86.7
FFT+DCTC with mel warp	**delta, acceleration**	**39**	**87.1**
FFT+DCTC with mel warp	6 DCS	78	87.8

Again, the FFT Mel warping is better than the filterbank Mel warping. The DCS is superior to differential dynamic basis vectors. We predict that with such high-dimensional features, the improvements would be more obvious with higher order mixture models, as shown in Table 2 for phonetic recognition.

4. CONCLUSIONS AND FUTURE WORK

In this work, we developed a unified framework by moving the amplitude scaling and modifying the basis vectors. Insights were discussed in detail using examples. Extensive experiments confirmed the rearrangement of the nonlinearity. Also, various basis vector combinations were examined to show their determinant impacts on the frontend performance. Advanced frontend features and systematic unifications for state-of-the-art recognition are under investigation in our ongoing work.

5. ACKNOWLEDGEMENT

This research is sponsored by the Air Force Research Laboratory under agreement number FA87501210093. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Lab or the U.S. Government.

6. REFERENCES

[1] J.S. Bridle and M.D. Brown, "An Experimental Automatic Word-Recognition System," *JSRU Report*, no.1003, Joint Speech Research Unit, Ruislip, England, 1974.

[2] J. Wu and J. Yu, "An Improved Arithmetic of MFCC in Speech recognition System," in *IEEE Int. Conf. Electro., Comm., and Control,* Sept.2011, pp.719-722.

[3] K. Kaewtip, L.N. Tan and A. Alwan, "A Pitch-Based Spectral Enhancement Technique for Robust Speech Processing," in *INTERSPEECH-2013*, Aug.2013, pp.3284-3288.

[4] M.A. Hossan, S. Memon and M.A. Gregory, "A Novel Approach for MFCC Feature Extraction," in *IEEE 4th Int. Conf. on Signal Processing and Communication Systems,* Dec.2010.

[5] S. Memon, M. Lech and N. Maddage, "Speaker Verification Based on Different Vector Quantization Techniques With Gaussian Mixture Models," in *Third Int. Conf. on Network and System Security,* 2009, pp.403-408.

[6] H.S. Jayanna and S.R.M. Prasanna, "Fuzzy Vector Quantization for Speaker Recognition under Limited Data Conditions," *TENCON 2008-IEEE Region 10 Conference,* 2008, pp.1-4.

[7] J. Chen, K.K. Paliwal, M. Mizumachi and S. Nakamura, "Robust MFCCs Derived From Different Power Spectrum," in *Eurospeech 2001,* Scandinavia, 2001.

[8] C. Wang, Z. Miao and X. Meng, "Differential MFCC and Vector Quantization Used for Real-Time Speaker Recognition System," in *IEEE Congress on Image and Signal Processing,* 2008, pp.319-323.

[9] S.A. Zahorian, H. Hu, Z. Chen and J. Wu, "Spectral and Temporal Modulation Features for Phonetic Recognition," in *INTERSPEECH-2009*, Sept. 2009, pp.1071-1074.

[10] F. Valente and H. Hermansky, "Hierarchical and Parallel Processing of Modulation Spectrum for ASR Applications," in *ICASSP-2008,* April 2008, pp.4165-4168.

[11] M. Kleinshmidt, "Localized Spectro-Temporal Features for Automatic Speech Recognition," in *Eurospeech 2003,* Sept. 2003, Switzerland.

[12] S. Seneff, "A Joint Synchrony/Mean Rate Model of Auditory Speech Processing," *Journal of Phonetics*, 16, pp.55-76, 1988.

[13] X. Zhang, M.G. Heinz, I.C. Bruce, and L.H. Carney, "A Phenomenological Model for the Response of Auditory-Nerve Fibers: I. Nonlinear Tuning With Compression and Suppression," *J.Acoust. Soc. Am.*, vol.109, no.2, pp.648-670, Feb.2001.

[14] C. Kim and R.M. Stern, "Feature Extraction for Robust Speech Recognition Using a Power-Law Nonlinearity and Power-Bias Subtraction," in *INTERSPEECH-2009*, Sept.2009, pp.28-31.

[15] H. Hermansky, "Perceptual Linear Prediction Analysis of Speech," *J. Acoust. Soc. Am.,* vol.87, no.4, pp.1738-1752, Apr, 1990.

[16] B.Meyer, S.V. Ravuri, M.R. Schadler and N. Morgan, "Comparing Different Flavors of Spectro-Temporal Features for ASR," in *INTERSPEECH-2011*, Aug.2011, pp.1269-1272.

[17] W. Ge, "Two Modified Methods of Feature Extraction for Automatic Speech Recognition," Master thesis, Department of Electrical and Computer Engineering, Binghamton University, Dec.2013.

[18] H. Duifhuis, "Consequences of Peripheral Filter Selectivity for Nonsimultaneous Masking," *J. Acoust. Soc. Am.,* vol.54, no.6, pp.1471-1488, 1973.

[19] G.M. Bidelman and A.S. Khaja, "Spectrotemporal Resolution Tradeoff in Auditory Processing as Revealed by Human Auditory Brainstem Responses and Psychophysical Indices," *Neuroscience Letters,* vol. 572, pp. 53-57, 2014.

[20] C.Kim and R.M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition," in *ICASSP-2012,* March 2012, pp.4101-4104.

[21] M. Slaney, "Auditory Toolbox Version 2," *Interval Research Corporation Technical Report,* no.10, 1998.

[22]. D. Ellis. (2006) PLP and RASTA (and MFCC, and inversion) in MATLAB using melfcc.m and invmelfcc.m. [Online].Available: http://labrosa.ee.columbia.edu/matlab/rastamat/