ROBUST SPEECH RECOGNITION USING WARPED DFT-BASED CEPSTRAL FEATURES IN CLEAN AND MULTISTYLE TRAINING

M. J. Alam, P. Kenny, P. Dumouchel, D. O'Shaughnessy

CRIM, Montreal, Canada
ETS, Montreal, Canada
INRS-EMT, Montreal, Canada

ABSTRACT

This paper investigates the robustness of the warped discrete Fourier transform (WDFT)-based cepstral features for continuous speech recognition under clean and multistyle training conditions. In the MFCC and PLP front-ends, in order to approximate the nonlinear characteristics of the human auditory system in frequency, the speech spectrum is warped using the Mel-scale filterbank, which typically consists of overlapping triangular filters. It is well known that such nonlinear frequency transformation-based features provide better speech recognition accuracy than linear frequency scale features. It has been found that warping the DFT spectrum directly, rather than using filterbank averaging, provides a more precise approximation to the perceptual scales. WDFT provides non-uniform resolution filter-banks whereas DFT provides uniform resolution filter-banks. Here, we provide a performance evaluation of the following variants of the warped cepstral features: WDFT, and WDFT-linear prediction-based MFCC features. Experiments were carried out on the AURORA-4 task. Experimental results demonstrate that the WDFT-based cepstral features outperform the conventional MFCC and PLP both in clean and multistyle training conditions in terms of recognition error rates.

Index Terms— Warped DFT, speech recognition, multi-style training, spectrum enhancement, linear prediction

1. INTRODUCTION

Mel-frequency cepstral coefficients (MFCCs) [1] and *perceptual linear prediction* (PLP) [21] have proven to be effective features for speech and speaker recognition tasks. MFCCs are usually computed by integrating short-term spectral power using a Mel-scaled filterbank (MelFB), typically consisting of overlapping triangular filters. The short-term power spectrum is warped according to the Mel scale to mimic the non-uniform frequency resolution property of the human auditory system. MFCC and PLP features perform well under matched training and test conditions but the performance gap between automatic speech recognizers (ASRs) and human listeners in real world settings is significant [2, 3]. Different operating conditions during signal acquisition - channel response, handset type, additive background noise, reverberation and so on - lead to feature mismatch across training and test utterances, thereby degrading the performance of the MFCC- and PLP-based recognizers. We focus on additive noise degradation.

There is a large body of research on improving the robustness of speech recognition systems under adverse acoustic environments. Environment compensation methods can be implemented at the front end (feature domain) [4-16], back end (model domain) [17-19] or both. Here, we focus on front-end techniques.

The goal of this paper is to compare several features utilizing *warped* DFT (introduced in [25]). This includes WDFT-MFCC (MFCC computed from the WDFT spectrum), and WDFT-LP (MFCC computed from the WDFT-based linear prediction spectrum) for a robust speech recognition task.

To evaluate and compare the performances of the WDFT cepstral features speech recognition experiments are performed on the AURORA-4 [22] LVCSR task both in clean and multistyle training conditions and the results are reported on the four evaluation conditions mentioned in section 4.1. For comparative purposes, the following front-ends are also included: standard MFCC [1], standard PLP [21]. Warped DFT-based features are found to provide lower recognition error rates than the DFT-based cepstral features.

2. MFCC AND PLP FRONT-ENDS

In the conventional MFCC front-end, processing of a speech signal begins with pre-processing (DC removal and pre-emphasis using a first-order high-pass filter with transfer function $\left(1-\alpha z^{-1}\right)$). Short-time Fourier transform (STFT) analysis is then performed using a finite duration (25 ms) Hamming window with a frame shift of 10 ms to estimate the power spectrum of the signal. The N-point windowed DFT, denoted by $S[k]$, is given by:

$$S[k]=\sum_{n=0}^{N-1}s[n]w[n]W^{kn} \leftarrow \begin{matrix} W=e^{-j\frac{2\pi}{N}} \\ k=0,1,...,N-1 \end{matrix}, \quad (1)$$

where k is the frequency bin index, n is the time index, $w[n]$ is the window function, and $s[n]$ is the short-time speech signal. Here, we choose $w[n]$ as the Hamming window. DFT provides a fixed frequency resolution, more specifically $2\pi/N$, over the whole frequency range [28]. In practice, DFT is implemented using the fast Fourier transform (FFT) algorithm. In order to approximate the nonlinear characteristics of the human auditory system in frequency, the speech spectrum is warped using the Mel-scale filterbank, which typically consists of overlapping triangular filters. It is well known that such nonlinear frequency transformation-based features provide better speech recognition accuracy than linear frequency scale features [1]. The mapping from linear frequency f (in Hz) to the Mel-frequency f_{mel} is performed using the following relation:

$$f_{mel}=2595\log_{10}\left(1+\frac{f}{700}\right). \quad (2)$$

Let $\mathbf{F}$ represent the $N_{\text{fb}}\times(N/2+1)$ filterbank matrix with N_{fb} Mel-filters and $\mathbf{C}$ the $N_{\text{ceps}}\times N_{\text{fb}}$ discrete cosine transformation matrix with N_{ceps} cepstral coefficients retained. Let M denote the number of frames. With these matrix notations, the N_{ceps}-dimensional MFCCs $\mathbf{c}$ can be obtained from the DFT-based speech spectrum matrix $\mathbf{S}$ of dimension $(N/2+1)\times M$ as:

$$\mathbf{c}=\mathbf{C}\left(\log\left(\mathbf{FS}\right)\right). \quad (3)$$

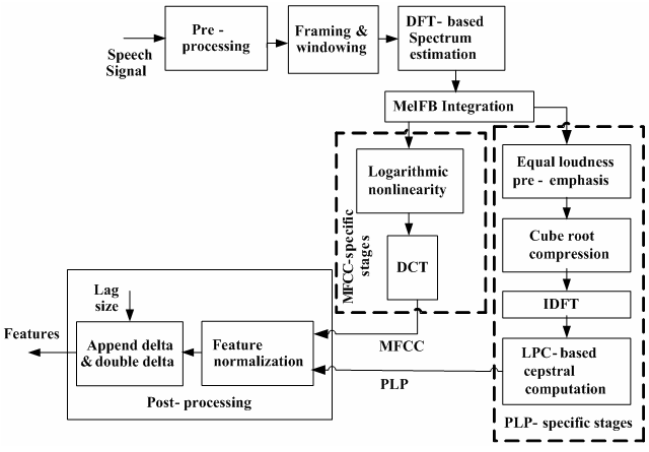

Fig. 1. Different steps of the MFCC and PLP front-ends.

PLP processing shares some common parts with MFCC processing, as shown in Fig. 1. In contrast to MFCC, pre-emphasis is performed based on an equal-loudness curve after Mel-frequency warping. Further, instead of logarithmic nonlinearity, cube root compression is performed in PLP to approximate the relationship between perceived loudness and the sound intensity [30]. After this stage, an inverse discrete Fourier transform (IDFT) is used for obtaining a perceptual autocorrelation sequence following linear prediction (LP) coefficient computation. Cepstral recursion is performed to obtain the final features from the LP coefficients [29]. Finally, the feature vector is augmented with time derivatives after being normalized by mean and variance normalization (MVN).

3. WARPED DFT-BASED CEPSTRAL FEATURES

Transforming a linear frequency scale to a non-linear frequency scale is called *frequency warping*. One method to achieve frequency warping is to apply a nonlinearly-scaled filterbank, such as a mel filterbank, to the linear frequency representation. Another way is to use a *conformal mapping*, such as the *bilinear transformation* [31-32], which preserves the unit circle. It is defined in the z-domain as:

$$H(z)=\frac{z^{-1}-\alpha'}{1-\alpha' z^{-1}},\ \forall\ -1<\alpha'<1 \quad (4)$$

where α' is the warp factor.

In warped DFT (WDFT) the locations of the frequency points are modified by applying an all-pass transformation to warp the frequency axis. Then, uniformly-spaced points on the warped frequency axis are equivalent to non-uniformly-spaced points on the original frequency axis. By choosing the warping parameters suitably, one can place some of the frequency samples close to each other to provide higher resolution in the frequency range of interest without increasing the length of the DFT [27]. With this frequency warping, one can improve the spectral representation of speech signals in the low frequency region [28].

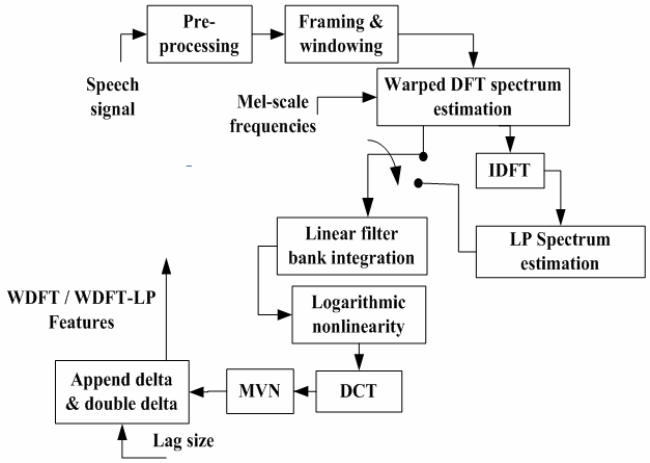

Fig. 2. Extraction of warped DFT-based cepstral features. Depending on the selection of the spectrum estimator, different variants of WDFT-based cepstral features are obtained, e.g., WDFT-LP when LP spectrum estimation is chosen.

For 8 kHz sampled signals, both the Mel and Bark scale can be approximated by warping factors $\alpha'=0.31$ and $\alpha'=0.42$, respectively [12]. Warping the DFT spectrum directly,

rather than using filterbank averaging, provides a more precise approximation to the perceptual scales [12].

The warped short-time speech spectrum is obtained by applying a warped DFT matrix $\tilde{\mathbf{W}}$, whose elements are given by $\tilde{W}^{\tilde{k}n} = e^{-j2\pi\tilde{k}n/N}$, $\tilde{k}$ being uniformly spaced on the Mel scale instead of the linear frequency (e.g., Hz) as in Eq. (1). Let $\mathbf{F}_l$ represent the $N_{\text{fb}} \times (N/2+1)$ linear filterbank matrix with N_{fb} linear filters, $\tilde{\mathbf{W}}$ the $(N/2+1)\times(N/2+1)$ WDFT matrix, $\mathbf{s}_w$ the framed and windowed speech signal matrix of size $(N/2+1)\times M$; then the warped cepstral features can be computed as:

$$\tilde{\mathbf{c}} = \mathbf{C}\left(\log\left(\mathbf{F}_u\left(\tilde{\mathbf{W}}\mathbf{s}_w\right)\right)\right), \tag{5}$$

where M is the number of frames.

The WDFT matrix $\tilde{\mathbf{W}}$ can be pre-computed and stored in a file (.mat file) to reduce the execution time. Since the spectrum is already pre-warped using Mel-frequency warping, the nonlinearly-spaced triangular-shaped Mel-frequency filterbank is replaced by a filterbank of uniformly spaced, half-overlapping triangular filters, to provide dimensionality reduction and spectral smoothing [21-22].

Fig. 3 shows running speech spectra of (a) clean and (b) noisy speech signals corrupted by babble noise with a signal-to-noise ratio of 6 dB, obtained using DFT, WDFT, and WDFT-LP spectrum estimators. Based on this visual examination, WDFT and WDFT-LP provide more robust spectral estimates compared to DFT and LP methods. Due to reduced degrees of freedom in all-pole modeling (model order p = 24 coefficients versus N = 256 bins), the WDFT-LP spectra are generally much smoother than the WDFT. This potentially results in improved noise robustness over WDFT [20].

In addition to WDFT- & WDFT-LP-based cepstral features, one can also compute WDFT-MVDR (minimum variance distortionless response) and WDFT-RMVDR (regularized MVDR) features using their corresponding all-pole model variants of MVDR [26] and regularized MVDR [13-15] coefficients. In this work we present only WDFT- and WDFT-LP-based cepstral features. (WDFT-MVDR and WDFT-RMVDR cepstral features are still in progress.)

Once the warped spectrum is obtained, the remainder of the feature extraction process in Fig. 2 can be summarized as follows:

(a) Apply inverse DFT (IDFT) on the warped power spectrum to compute a perceptual autocorrelation sequence.

(b) Compute LP coefficients by performing pth order LP analysis via Levinson-Durbin recursion using perceptual autocorrelation lags [29].

(c) Obtain WDFT-LP cepstral features from the LP spectral estimates followed by a linear-scale filterbank, logarithmic compression and DCT [20].

There are at least two possible ways to compute the cepstrum from the all-pole spectrum. The first way is to compute the all-pole model and derive the cepstra directly from the coefficients of the all-pole filter [11]. The second way is to compute the spectrum from the LP coefficients using DFT and compute the cepstral coefficients from the spectrum in the standard way (Fig. 2) by replacing the Mel filterbank with a linear-scale filterbank. In this paper, we choose the second approach because of the ease with which perceptual considerations can be incorporated [11].

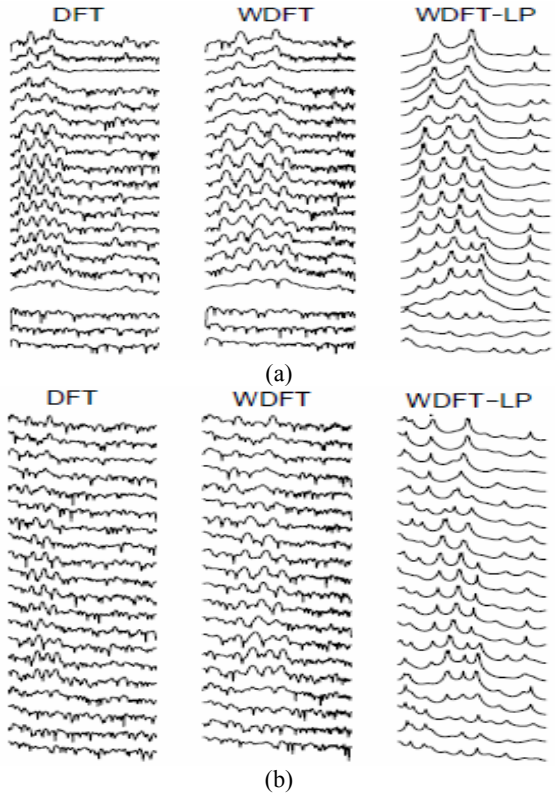

Fig. 3. Comparison of running spectra of (a) clean and (b) noisy (degraded with babble noise with a signal-to-noise ratio of 6 dB) speech signals [20]. Time runs from bottom up and the frequency axis from left to right. The frequency axis is linear for DFT and for WDFT (warped DFT) and WDFT-liner prediction (WDFT-LP) it is linear in the Mel scale. The model order (p) used for WDFT-LP is 24.

4. PERFORMANCE EVALUATION

Warped DFT (WDFT)- and WDFT-linear prediction (WDFT-LP)-based cepstral feature extractors, as presented in fig. 2, are evaluated and compared with the conventional MFCC, PLP front-ends on the AURORA-4 corpus in the context of speech recognition. Both clean and multistyle training modes are considered here. Word error rate (WER) is used as an evaluation metric.

4.1. Speech Corpus and Experimental Setup

The AURORA-4 [22] continuous speech recognition corpus consists of a clean training set, a multi-condition training set

and 14 evaluation (or test) sets. The 14 test sets are grouped into the following 4 evaluation conditions [22-23].

Test set A - clean speech in training and test, same channel (set 1), ***Test set B*** - clean speech in training and noisy speech in test, same channel (sets 2-7), ***Test set C*** - clean speech in training and test, different channel (set 8), ***Test set D*** - clean speech in training and noisy speech in test, different channel (sets 9-14). The number inside the brackets represents the test set number defined in the AURORA-4 corpus.

For the continuous speech recognition task on the AURORA-4 corpus, all experiments employed state-tied crossword speaker-independent triphone acoustic models with 4 Gaussian mixtures per state. A single-pass Viterbi beam search-based decoder was used along with a standard 5K lexicon and bigram language model with a prune width of 250 [23]. We use a HTK-based recognizer [24].

For our experiments, we use 13 static cepstral features (including the 0th cepstral coefficient) augmented with their delta and double-delta coefficients, making 39-dimensional feature vectors. The analysis frame length is 25 ms with a frame shift of 10 ms. The delta and double features are calculated using a 5-frame window. For all methods, presented in Table 1, extracted features are normalized using utterance-level mean and variance normalization (MVN).

4.2. Results and Discussion

Word error rate (WER) is used as an evaluation metric for performance evaluation and comparison of the warped DFT-based cepstral feature extraction methods. Plotted Spectra of a noisy speech signal in fig. 3 and the speaker recognition results of [20] suggest higher robustness of WDFT- and WDFT-LP-based features over the DFT-based MFCC and PLP features. To select the optimal model order for the all-pole variant WDFT-LP, we perform speech recognition experiments by varying p from 10 to 30. The model order that provided lowest WER was selected as the optimal model order. The optimal model order found in these experiments is 24. In [12], the optimal model order $p = 24$ was reported for the perceptual MVDR (PMVDR), a method similar to WDFT-MVDR. The difference between PMVDR and WDFT-MVDR is that in the former the Mel-scale filterbank is approximated by adjusting the warp factor of a bilinear transformation. A high-order model in the all-pole modeling is needed to model just enough detail necessary for accurate recognition [12]. Table 1 presents the WER (in %) obtained by the various front-ends considered in this work when the recognizer is trained using the clean training features and tested on the clean as well as noisy test features. None of the front-ends of Table 1 include any additional noise compensation method, such as speech enhancement or additional feature normalization beyond MVN. According to Table 1, WDFT-based cepstral features outperform MFCC, PLP features under mismatched conditions, as expected from prior literature [12, 20, 30]. WDFT-LP performs the best on average over all the other front-ends.

In Table 2 the WERs (in %) obtained by the various front-ends considered in this work are presented when the recognizer is trained on the multistyle (or multi-condition) training features and recognition is performed on the clean as well as noisy test features. Multistyle training is a very effective method for the compensation of mismatch between train/test environments. In multistyle training enough representation data (clean plus noisy) is included in the training phase to create somewhat matched training/test environments. It is observed from table 2 that the WDFT-based cepstral features outperformed, on the average, the DFT-based MFCC and PLP features. Comparing the results of tables 1 and 2 it can be said that WDFT-based cepstral features performed better than the MFCC and PLP both in clean and multi-condition training modes. It indicates that warping the DFT spectrum directly provides a more precise approximation to the perceptual scales than using filterbank averaging.

	A	**B**	**C**	**D**	**Avg.**
MFCC	**9.98**	50.81	28.88	64.55	38.56
PLP (HTK)	10.28	49.59	25.56	60.36	36.45
WDFT-MFCC	10.90	49.07	24.27	60.67	36.23
WDFT-LP	11.01	**43.08**	**23.65**	**54.68**	**33.10**

Table 1. Word error rates (WERs in %) obtained by the various feature extractors considered in this paper, on the AURORA-4 LVCSR corpus under clean training conditions. The model order selected in this task is: $p = 24$ for WDFT-LP and $p = 14$ for PLP. The lower the WER the better is the performance of the feature extractor.

	A	**B**	**C**	**D**	**Avg.**
MFCC	**14.62**	23.84	19.19	31.47	22.28
PLP (HTK)	16.10	24.98	18.27	**30.23**	22.40
WDFT-MFCC	15.43	**23.65**	17.97	30.41	**21.87**
WDFT-LP	15.46	23.98	**17.50**	30.66	**21.90**

Table 2. Word error rates (WERs in %) obtained by the various feature extractors considered in this paper, on the AURORA-4 LVCSR corpus under multistyle training condition. The model order selected in this task is: $p = 24$ for WDFT-LP and $p = 14$ for PLP. The lower the WER the better is the performance of the feature extractor.

5. CONCLUSION

Variants of the Mel-frequency warped discrete Fourier transform, a more robust warped frequency representation-based cepstral feature, are presented. MFCC features computed from the Mel-warped DFT spectrum-based front-ends (WDFT, WDFT-LP) provided lower recognition error rates

than the conventional MFCC and PLP on the AURORA-4 corpus. The presented speech spectra (fig. 3) and experimental speech recognition results on the AURORA-4 LVCSR task demonstrated the robustness of the WDFT- and WDFT-LP-based cepstral features. Our future work includes:

1. Computation of WDFT-MVDR (minimum variance distortionless response) and WDFT-RMVDR (regularized MVDR)-based features using their corresponding all-pole model variants of MVDR [26] and regularized MVDR [13-15].
2. Incorporation of auditory domain enhancement techniques [5, 6] into the warped DFT-based cepstral feature extraction framework to improve its robustness, specifically in clean training condition.

REFERENCES

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE TASLP*, vol. 28, no. 4, pp. 357–366, August 1980.

[2] Huang, X., Acero, A., Hon, H., Spoken Language Processing: A Guide to Theory, Algorithm and System development, Prentice-Hall PTR, Upper Saddle River, New Jersey, 2001.

[3] D. O'Shaughnessy, Speech Communications: Human and Machine, 2nd ed., IEEE Press, 2000.

[4] ETSI ES 202 050, Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; advanced front-end feature extraction algorithm; Compression algorithms; 2003.

[5] C. Kim and R. M. Stern., "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," Proc. ICASSP, pp. 4574-4577, March 2010.

[6] M. J. Alam, P. Kenny, D. O'Shaughnessy, "Robust Feature Extraction for Speech Recognition by Enhancing Auditory Spectrum," Proc. INTERSPEECH, Portland Oregon September 2012.

[7] J. van Hout, A. Alwan, "A novel approach to soft-mask estimation and log-spectral enhancement for robust speech recognition," Proc. of ICASSP, pp. 4105-4108, 2012.

[8] V. Mitra, H. Franco, M. Graciarena, A. Mandal, "Normalized Amplitude modulation features for large vocabulary noise-robust speech recognition," Proc. of ICASSP , pp. 4117-4120, 2012.

[9] M. J. Alam, P. Kenny and D. O'Shaughnessy, "Smoothed Nonlinear Energy Operator-based Amplitude Modulation Features for Robust Speech Recognition," Proc. NOLISP, LNAI 7911, pp. 168-175, Mons, Belgium, 2013.

[10] W. Zhu, D. O'Shaughnessy, "Incorporating frequency masking filtering in a standard MFCC feature extraction algorithm," Proc. ICSP, pp. 617-620, Beijing, Aug-Sep., 2004.

[11] S. Dharanipragada, B. D. Rao, "MVDR based Feature Extraction for Robust Speech Recognition", *Proc. ICASSP*, pp. 309-312, 2001.

[12] U. H. Yapanel, J.H.L. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," Speech Comm., Vol. 50, pp. 142-152, 2008.

[13] M. J. Alam, P. Kenny, D. O'Shaughnessy, "Speech recognition using regularized minimum variance distortion-less response spectrum-estimation based cepstral features," Proc. ICASSP, Vancouver, Canada, May, 2013.

[14] M. J. Alam, D. O'Shaughnessy, P. Kenny, "A novel feature extractor employing regularized MVDR spectrum estimator and subband spectrum enhancement technique," Proc. WOSSPA, Algiers, Algeria, May, 2013.

[15] M. J. Alam, P. Kenny, D. O'Shaughnessy, "Regularized MVDR Spectrum Estimation-based Robust Feature Extractors for Speech Recognition," Proc. INTERSPEECH, Lyon, France, 2013.

[16] J. Droppo, A. Acero, "Environmental robustness," in springer handbook of speech processing, Benesy, J.; Sondhi, M. M. and Huang, Y. [Eds], pp. 653-679, 2008.

[17] Holmes, N. J. and Sedgwick, N. C., "Noise compensation for speech recognition using probabilistic models," *Proc. of ICASSP*, vol. 11, p. 741-744, 1986.

[18] Q. Huo, C. Chan, and C. H. Lee, "Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition," IEEE Trans. Speech and Audio Processing, vol. 3, pp. 334-345, Sep. 1995.

[19] Gales, M. J. F. and Young, S. J., "On stochastic feature and model compensation approaches to robust speech recognition," Speech Communication, vol. 25, pp. 29-47, 1998.

[20] T. Kinnunen, M. J. Alam, P. Matejka, P. Kenny, J. "Honza" Cernocky, D. O'Shaughnessy, "Frequency Warping and Robust Speaker Verification: A Comparison of Alternative Mel-Scale Representations," Proc. INTERSPEECH, Lyon, France, 2013.

[21] H. Hermansky, Perceptual linear prediction analysis of speech, J. Acoust. Soc. Am., vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[22] N. Parihar, J. Picone, D. Pearce, H.G. Hirsch, "Performance analysis of the Aurora large vocabulary baseline system," Proc. of EUSIPCO, Vienna, Austria, 2004.

[23] S.-K. Au Yeung, M.-H. Siu, "Improved performance of Aurora-4 using HTK and unsupervised MLLR adaptation," Proceedings of the Int. Conference on Spoken Language Processing, Jeju, Korea, 2004.

[24] S. J. Young et al., HTK Book, Entropic Cambridge Research Laboratory Ltd., 3.4 edition, 2006.

[25] A. Makur and S. Mitra, "Warped discrete-Fourier transform: Theory and applications," *IEEE Trans. Circuits and Systems I: Fundamental Theory and Applications*, vol. 48, no. 9, pp. 1086–1093, September 2001.

[26] M.N. Murthi and B.D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," IEEE Trans. Speech Audio Processing, vol. 8, no. 3, pp. 221–239, May 2000.

[27] R. Venkataramanan, K.M.M. Prabhu, "Estimation of frequency offset using warped discrete Fourier transform," Signal Processing journal, vol. 86, pp. 250-256, 2006.

[28] S. Franz, S. K. Mitra, J. C. Schmidt, G. Doblinger, "Warped Discrete Fourier Transform: A New Concept in Digital Signal Processing," Proc. of ICASSP, pp. 1205-1208, 2002.

[29] J. Makhoul "Linear Prediction: a Tutorial Review," Proc. of IEEE, vol. 63, no.4, pp.561-580, 1975.

[30] M. Wolfel, Q. Yang, Q. Jin, T. Schultz, "Speaker Identification using Warped MVDR Cepstral Features," Proc. Interspeech, pp. 912-915, 2009.