ROBUST SPEECH RECOGNITION UNDER NOISY ENVIRONMENTS USING ASYMMETRIC TAPERS

Md Jahangir Alam, Patrick Kenny, Douglas O'Shaughnessy

INRS-EMT, University of Quebec, Montreal, Canada
CRIM, Montreal Canada

ABSTRACT

This paper presents asymmetric taper (or window)-based robust Mel frequency cepstral coefficient (MFCC) feature extraction for automatic speech recognition (ASR). Commonly, MFCC features are computed from a symmetric Hamming-tapered direct-spectrum estimate. Symmetric tapers have linear phase and also imply longer time delay. In ASR systems, phase information is usually discarded as human speech perception is relatively insensitive to short-time phase distortion. So, any linearity constraint on phase can be removed without adverse effects. Use of asymmetric tapers, having better frequency response and shorter time delay, for MFCC feature extraction in speech recognition can lead to better recognition performance. Using our proposed method it is possible to introduce asymmetry in any symmetric taper by adjusting only one additional parameter, which controls the degree of asymmetry. Experimental results on the AURORA-2 corpus show that the proposed asymmetric tapers outperform the symmetric Hamming taper in terms of word accuracy both in clean and noisy environments.

Index Terms—Asymmetric taper, double dynamic range, speech recognition, Hilbert transform.

1. INTRODUCTION

Mel-frequency cepstral coefficient (MFCC) features are the most dominantly used in speech recognition systems. MFCC processing of speech signal begins with pre-processing (includes DC removal and pre-emphasis, typically using a first-order high-pass filter). Short-time Fourier Transform (STFT) analysis is performed using a finite duration (20-30 ms) symmetric-shaped single taper (e.g., Hamming) technique to estimate the power spectrum of the signal, and triangular Mel-frequency integration is performed for auditory spectral analysis. The logarithmic nonlinearity stage follows, and the final static features are obtained through the use of a Discrete Cosine Transform (DCT). Therefore, the accuracy of the MFCC features depends on the accuracy of the power spectral estimate. Under matched conditions, MFCC features perform well but under mismatched environments (i.e., different training and testing environments due to channel, handset, additive background noise and reverberation), the performance severely deteriorates. The reason for this is that the direct spectral estimate used in MFCC feature computation gets affected by the factors (additive distortion, reverberation etc.) causing mismatched environments. In this paper, for robust estimation of the signal power spectrum, and hence robust MFCC features, we replace the symmetric Hamming taper by an asymmetric taper.

Various tapers have been proposed in the literature for better spectral estimation of the signal [1]. Most of the speech recognition systems use symmetric tapers, such as Hamming, Hann, etc., because of their ease of implementation and linear phase property. Symmetry implies potential drawbacks such as longer time delay and frequency response limitations [2]. It is common belief in speech community that in human perception tasks as well as in automatic speech recognition systems the short-time phase spectrum plays very little (or, no) role. Phase information is usually completely disregarded in recognition systems; so there is no apparent reason for using symmetric tapers. Removal of symmetry constraints therefore give asymmetric tapers, having some better properties like shorter time delay (important for coding but less important for recognition) and robust frequency response. Some low delay speech coders, e.g., ITU-T G.729 [4], use an asymmetric analysis taper, which is formed by combining two symmetric tapers, Hamming and cosine tapers. Nobody has attempted this asymmetric taper in speech recognition. The popularity of asymmetric tapers in speech coding suggests that their advantages can be applied in speech and speaker recognition systems as well. Asymmetric tapers, designed to solve a more complex minimax approximation problem, have been applied in a speech recognition task [2]. In [6], asymmetric double dynamic range (DDR) Hamming tapers were proposed by shifting the peak position of the symmetric DDR Hamming taper [3].

In this paper we use a new method for the construction of an asymmetric taper based on an existing symmetric taper. Compared to the symmetric Hamming taper, only one additional parameter is required, which controls the degree of asymmetry, for the proposed asymmetric tapers. Proposed tapers have rapidly decaying side lobes and the width of the main lobe is larger than that of the Hamming taper. By robust estimation of a signal's power spectrum and consequently MFCC features, we hope that the proposed asymmetric tapers achieve better speech recognition performance on the AURORA-2 corpus than the symmetric Hamming taper.

2. SYMMETRIC TAPERS IN ASR

Most speaker/speech recognition systems for short-time analysis of a speech signal use standard symmetric-shaped tapers such as Hamming, Hann, etc. These tapers have the linear phase property and have a particular shape of magnitude response [2]. Symmetric tapers have a closed-form expression and are easily computable, but these tapers provide poor magnitude response under mismatched conditions. Also these tapers have larger time delay [2]. Relaxation of linear phase constraints can therefore lead to asymmetric tapers with better magnitude response both in matched and mismatched environments and with a shorter time delay. Since a Hamming taper is widely used in speaker and speech recognition systems, in this paper we use this taper for performance comparison with the proposed asymmetric tapers.

3. PROPOSED ASYMMETRIC TAPERS

The proposed method for the construction of an asymmetric taper from a symmetric taper is shown in figure 1. From a symmetric taper $w_s(n)$ of length N, the instantaneous phase $\theta(n)$ is computed by applying a Hilbert transform to the symmetric taper. Then the asymmetric taper $w_{at}(n)$ is obtained as:

$$w_{at}(n) = c w_s(n) e^{\kappa\theta(n)},\ 0 \le n \le N\text{-}1$$

where n is the time index, $w_s(n)$ is the symmetric taper of length N, $e^{\kappa\theta(n)}$ is an asymmetric function, κ is a parameter that controls the degree of asymmetry, and c is the normalizing constant given by

$$c = \frac{\max(w_s(n))}{\max(w_s(n)e^{\kappa\theta(n)})},\ 0 \le n \le N-1.$$

For positive values of κ the asymmetric function $e^{\kappa\theta(n)}$ acts as a high-pass filter (HPF), it acts as a low-pass filter (LPF) for negative values of κ and $w_{at}(n) = w_s(n)$ when $\kappa = 0$.

In this paper we propose two asymmetric tapers, denoted in this paper as Proposed II and Proposed I, based on the symmetric Hamming and double dynamic range (DDR) hamming tapers [3], respectively, for different values of κ. The DDR Hamming taper was proposed in [3] for use in higher-lag autocorrelation spectrum estimation (HASE) method. The DDR Hamming window is computed from an *N/2*-length Hamming window in the following way [3]:

- Calculate a biased autocorrelation sequence of length *N-1* having a maximum at the zero-th lag in the centre from a Hamming window of length *N/2*.
- The desired DDR window of length *N* is found by padding one zero value at the end of the autocorrelation sequence.

Since the DDR window is constructed from a Hamming window and has dynamic range (86 dB) twice the dynamic range of a Hamming window (43 dB), it is called a DDR Hamming window.

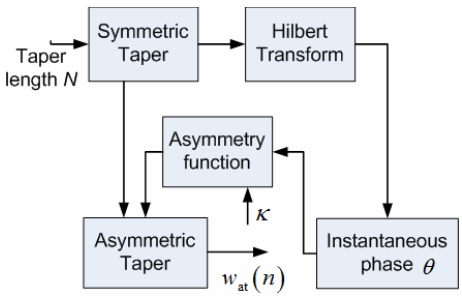


Fig. 1. Block diagram of the proposed asymmetric taper. Parameter κ controls the degree of asymmetry of the asymmetric taper.

According to [6], window centered on the pitch helps to have a good characterization of the formants, because the information regarding the spectral envelope (and the formants) is located on the zero lag area as well as on the pitch and on lags multiple of the pitch. The optimal values for κ are chosen using above mentioned information and by tuning on the development data.

Fig. 2 presents a time and frequency domain comparison of the Hamming and the proposed asymmetric tapers for frame length N = 200 samples. It is observed from fig. 2 (b) that all the asymmetric tapers have wider mainlobe widths and higher attenuation in the sidelobes than the Hamming taper.

Asymmetric tapers also result in shorter time delay [2], which is important for coding but less important for the recognition task alone.

Figs. 3 (a) and (b) show a three-tone signal in the time domain and a comparison of the taper influence on the estimated power spectrum of a signal consisting of three pure tones, respectively. Figs. 4 (a) and (b) present one

frame of a speech signal, degraded by car noise with SNR = 5 dB, in the time domain and a comparison of taper influence on the estimated power spectrum of that of a speech signal using symmetric Hamming and proposed asymmetric tapers. Larger suppression in the sidelobes (can be obtained by widening the mainlobe width) and rapidly decaying height of sidelobes are important for speech recognition performance [2]. It is also observed from fig. 4 (b) that the asymmetric tapers also result in a reduction of variance of the spectrum ordinates compared to the Hamming window. Fig. 5 presents speech spectrograms of a clean signal (MRK_213Z1ZZA.08) obtained using the symmetric Hamming window and the proposed asymmetric tapers. It is observed from fig. 5 that the proposed asymmetric tapers do not distort the clean signal. Fig. 6 presents speech spectrograms of a noisy signal (MRK_213Z1ZZA.08, subway noise, SNR = 15 dB) obtained using the various windowing techniques. The proposed asymmetric tapers show substantially lower noise in the spectrograms.

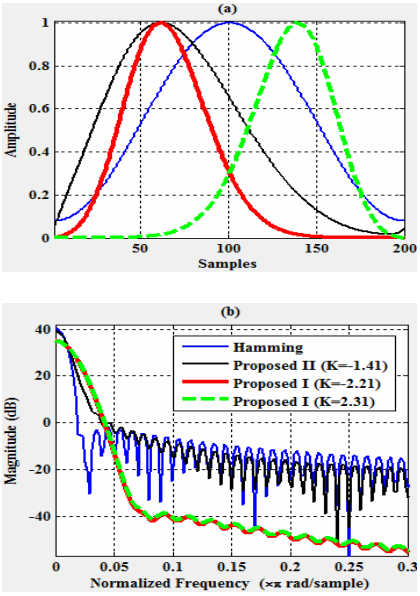
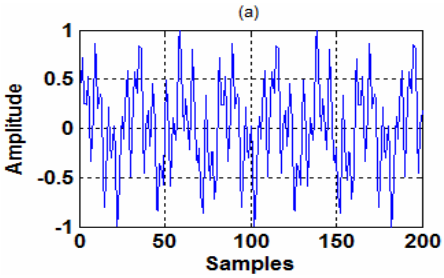


Fig. 2. Comparison of symmetric Hamming and proposed asymmetric tapers in the (a) time domain, (b) frequency domain (magnitude response in dB).



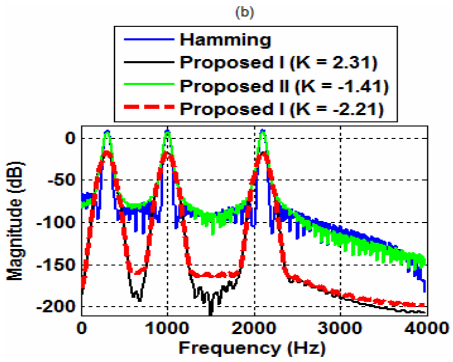


Fig. 3. (a) Time domain signal comprise of three–tones, (b) comparison of taper influence on the estimated power spectrum of a simple three-tone signal using symmetric Hamming and proposed asymmetric tapers.

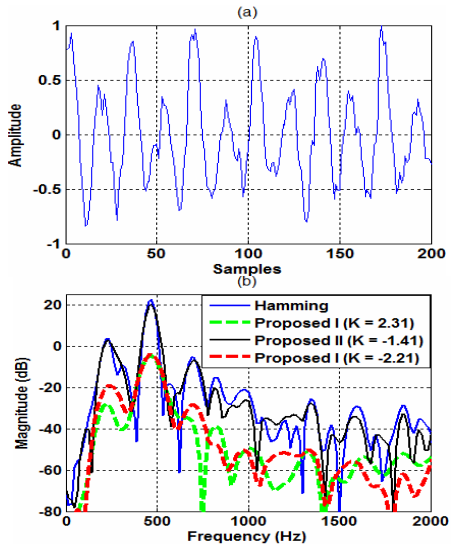


Fig. 4. (a) One frame of speech signal in the time domain, car noise, SNR = 5 dB, (b) comparison of taper influence on the estimated power spectrum of a frame of speech signal, corrupted with 5 dB car noise, using symmetric Hamming and proposed asymmetric tapers.

4. EXPERIMENTAL RESULTS

4.1. Experimental Setup

The AURORA-2 [5] (connected digit, small vocabulary) database is used for comparing the performances of the proposed asymmetric-shaped tapers to the conventional Hamming window, in the context of speech recognition. There are two training sets (clean training set and multi-condition training set) and three test sets (test sets A, B and C). The clean training set consists of clean speech recordings only from 55 male and 55 female adults. The multi-condition training consists of both clean and noisy

speech split into 20 subsets. The 20 subsets represent 4 different noise scenarios (subway, babble, car and exhibition hall) at six different SNRs (20 dB, 15 dB, 10 dB, 5 dB, 0dB and -5 dB). Test set A is composed of speech with conditions matched to the multi-condition training set, test set B is composed of speech with non-matched background noise (restaurant, street, airport and train-station) and test set C is composed of speech with partly matched (with multi-condition training) background noise and non-matched convolutional noise (MIRS (modified intermediate reference system) filtered subway and street noise). The clean training set constitutes mismatched training/testing conditions whereas the multi-condition training set constitutes much more matched training/testing conditions. In this paper we train the recognizer on clean utterances only and perform recognition on all three test sets.

For our experiments, we use 13 MFCC features (including the 0^{th} 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 were calculated using 5-frame and 3-frame windows, respectively. For all the feature extractors, the features, after appending delta and double delta features, were normalized using the conventional mean and variance (MVN) normalization technique over the whole utterance. For the recognition task we use the HTK speech recognizer. In the experiments we choose a simple HMM-based system with 16 states per word model, 3 Gaussian components per state.

4.2. Results and Discussion

The AURORA-2 corpus [5] is used for comparing the performances of the proposed asymmetric tapers (Proposed I for $\kappa = 2.31, 2.21, -2.21$ and Proposed II for $\kappa = -1.41$) to the conventional Hamming taper, in the context of speech recognition. Most of the Hamming taper-based feature extractors used in speech recognition perform well in controlled environments where speech data is collected from reasonably clean environments.

Real-life environments are far less controlled. Acoustic mismatch due to different training and testing environments degrades the performance of speech recognition systems. For the performance evaluation of the asymmetric taper-based MFCC feature extractors, we have chosen mismatched conditions. Features extracted from the clean training data are used for training the recognizer. For testing we have used all ten noise scenarios of the AURORA-2 corpus at six different SNRs (clean (SNR > 40 dB), 20 dB, 15 dB, 10 dB, 5 dB, 0 dB).

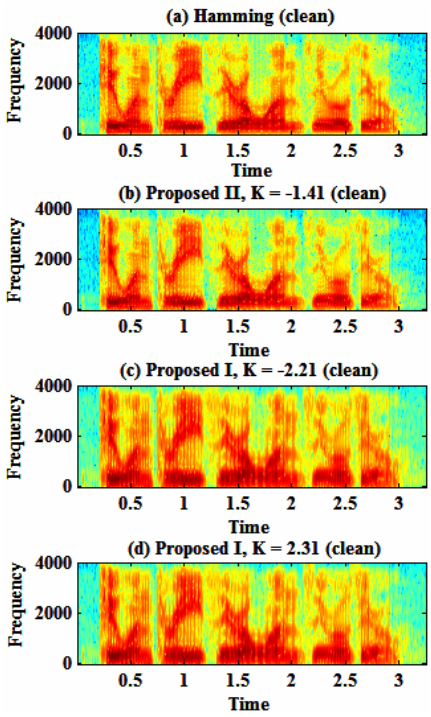


Fig. 5. Spectrograms of clean speech signal obtained using the symmetric Hamming and the proposed asymmetric tapers. X-axis represents time in seconds and Y-axis represents the frequency in Hz.

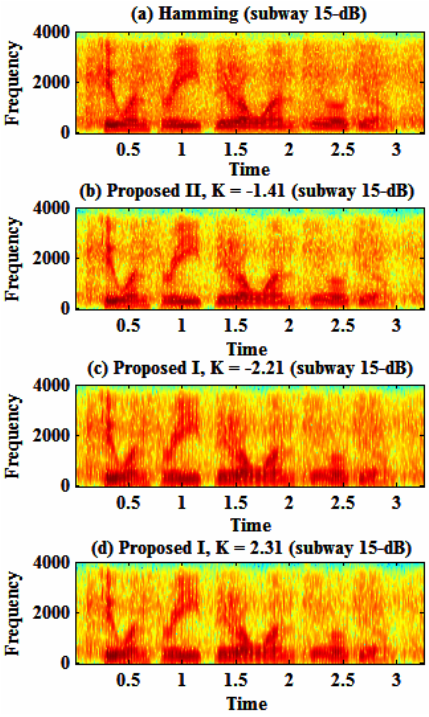


Fig. 6. Spectrograms of noisy speech signal (subway noise, SNR = 15 dB) obtained using the symmetric Hamming and the proposed asymmetric tapers. X-axis represents time in seconds and Y-axis represents the frequency in Hertz (Hz).

Tables 1-3 present the average word accuracy (in %) of the proposed asymmetric tapers and symmetric Hamming taper-based feature MFCC extractors, considered in this paper, on test sets A, B, and C, respectively. In all noise scenarios except in two cases at 20 dB SNR, the proposed asymmetric tapers provide better word accuracy than the widely used symmetric Hamming window. It is observed from the experimental results that the improvements obtained by the asymmetric tapers compared to the symmetric taper, in terms of word accuracy, are much higher in low SNR conditions, specifically at 0 dB and 5 dB, than the high SNR cases. These results indicate that the asymmetric tapers provide better magnitude response of the speech signal than the symmetric one, both in matched and mismatched (due to additive and channel effects) environments.

Table 1: Average word accuracy in percentage (averaged over all noise conditions of test set A) obtained using a Hamming window and the proposed asymmetric tapers. The best results are in bold face.

Set A	**Clean**	**20 dB**	**15 dB**	**10 dB**	**5 dB**	**0 dB**
Hamming	99.03	96.52	92.95	84.55	65.09	31.76
Proposed I (κ =2.31)	99.09	96.52	93.22	85.21	**68.14**	**35.55**
Proposed I (κ =2.21)	99.09	96.52	93.13	85.14	67.68	33.96
Proposed I (κ =-2.21)	99.11	**96.80**	**93.48**	**85.32**	67.11	32.22
Proposed II (κ =-1.41)	**99.18**	96.67	93.09	84.62	66.38	33.89

Table 2: Average word accuracy in percentage (averaged over all noise conditions of test set B) obtained using the Hamming window and proposed asymmetric tapers. The best results are in bold face.

Set B	**Clean**	**20dB**	**15dB**	**10dB**	**5dB**	**0dB**
Hamming	99.03	**96.94**	93.68	85.77	66.62	33.85
Proposed I (κ =2.31)	99.09	96.84	93.71	**86.71**	**68.69**	**35.64**
Proposed I (κ =2.21)	99.09	96.80	93.78	86.65	68.12	34.04
Proposed I (κ =-2.21)	99.11	96.92	**93.89**	86.43	67.61	33.72
Proposed II (κ =-1.41)	**99.18**	96.83	93.81	86.12	67.63	34.88

Table 3: Average word accuracy in percentage (averaged over all noise conditions of test set C) obtained using the Hamming window and proposed asymmetric tapers. The best results are in bold face.

Set C	**Clean**	**20dB**	**15dB**	**10dB**	**5dB**	**0dB**
Hamming	99.04	**96.87**	93.27	85.42	67.05	33.82
Proposed I (κ =2.31)	99.06	96.86	93.18	86.09	**69.80**	**39.14**
Proposed I (κ =2.21)	99.07	96.71	93.17	85.89	69.32	37.04
Proposed I (κ =-2.21)	99.00	96.84	93.19	**86.40**	69.06	35.49
Proposed II (κ =-1.41)	**99.12**	96.81	**93.33**	85.77	68.24	36.39

5. CONCLUSION

In this paper we proposed a generalized method for the construction of asymmetric tapers from a symmetric taper. We incorporated those tapers in the MFCC feature extraction process and compared their performances in the context of speech recognition on the AURORA-2 corpus. Experimental results indicate that the asymmetric tapers outperformed the symmetric Hamming taper. The largest improvements in % word accuracy over the baseline were observed for SNRs 0 dB and 5 dB on all test sets. Here, we tuned the values of the parameter κ experimentally on the development data. Our future work would be to make κ adaptive so that no manual tuning is needed.

6. REFERENCES

[1] J.G. Proakis, D.G. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications*, 3rd edition, Prentice Hall, New York, 2000.

[2] R. Rozman, D.M. Kodek, "Using asymmetric windows in automatic speech recognition." Speech Comm., vol. 49, pp. 268-276, Jan 2007.

[3] B. Shannon, K.K. Paliwal, "Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition," Speech Comm., vol. 48, pp. 1458–1485, August 2006.

[4] ITU-T, Geneva, Recommendation G.729, Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP), Mar. 1996.

[5] H. G. Hirsch and D. Pearce: The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Condition, In: ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium, France (2000).

[6] Juan A. Morales-Cordovilla, Victoria Sánchez, Antonio M. Peinado and Ángel M. Gómez. "On the use of asymmetric windows for robust speech recognition". *Circuits, Systems and Signal Processing (Springer)*, September, 2011.