Odyssey 2012
The Speaker and Language Recognition Workshop
25-28 June 2012, Singapore

On the use of Asymmetric-shaped Tapers for Speaker Verification using I-vectors

Md Jahangir Alam[1], *Patrick Kenny*[2], *Douglas O'Shaughnessy*[3]

[1,3] INRS-EMT, University of Quebec, Montreal, Canada
[1,2] CRIM, Montreal, Canada
{Jahangir.Alam, Patrick.Kenny}@crim.ca, dougo@emt.inrs.ca

Abstract

This paper presents asymmetric-shaped tapers (or windows) for speaker recognition. Symmetric tapers (e.g., hamming), having the linear phase property and longer time delay, are widely used for short-time analysis of speech signals. Since human speech perception is relatively insensitive to short-time phase distortion, the linearity constraint on phase can be removed without any adverse effects. Use of asymmetric tapers, having better magnitude response and shorter time delay, in speaker recognition can lead to a better recognition performance. Speaker verification results on the telephone and microphone speech of the latest NIST 2010 SRE corpus show that the asymmetric-shaped tapers perform better than the symmetric hamming window.

1. Introduction

The Mel frequency cepstral coefficients (MFCC) features are the most dominantly used in speaker recognition systems. MFCC processing of a speech signal begins with pre-processing (including 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)/multi-taper 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, 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 factors (additive distortion, reverberation etc.) causing mismatched environments. In this paper, for better and robust (to noise distortions) estimation of the signal power spectrum, and hence better and robust MFCC features, we replace the symmetric Hamming taper by an asymmetric-shaped taper assuming that this will bring improvement in speaker verification performance.

Various tapers have been proposed in the literature for better spectral estimation of the signal [1]. Most speaker recognition systems use symmetric tapers, such as Hamming or Hann, because of their ease of implementation and linear phase property. Symmetry implies potential drawbacks like longer time delay and frequency response limitations [2]. Phase information is completely disregarded in recognition systems, so, there is no apparent reason for using symmetric tapers. Removal of the symmetry constraint therefore allows asymmetric tapers to have some better properties such as shorter time delay (important for coding but less important for recognition) and better and robust frequency response. Some low delay speech coders, e.g., ITU-T G.729 [4], use an asymmetric analysis taper. Asymmetric tapers, designed by solving a more complex minimax approximation problem, have also successfully been applied in speech recognition [2], but not in speaker recognition. In this paper we use two asymmetric-shaped tapers, the ITU-T G.729 Hamming Cosine window [4] and the asymmetric form of the double dynamic range (DDR) Hamming window [3], for speaker recognition. The DDR Hamming window was proposed in [3] for higher lag autocorrelation spectrum estimation. For performance evaluation, we use the latest NIST 2010 SRE benchmark data with a state-of-the-art i-vector configuration [5-7].

2. Standard Tapers

For short-time analysis of speech signals, most speaker/speech recognition systems use standard symmetric-shaped tapers such as Hamming or Hann. These tapers have a linear phase property and 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. Relaxation of the linear phase constraint can therefore lead to asymmetric tapers with better magnitude response, both in matched and mismatched environments, and a shorter time delay. Since the Hamming taper is the most popular in speaker/speech recognition, in this paper we will use this taper for performance comparison, with the asymmetric tapers to be discussed in the next section.

3. Asymmetric-shaped Tapers

We present two asymmetric tapers in this section, one used in [4] for low delay speech coding and the other based on a DDR Hamming taper proposed in [3], for performance evaluation and comparison with the symmetric Hamming taper, in the context of speaker verification. The asymmetric-shaped taper used in the ITU-T G.729 coder [4] is given by:

$$w_{\text{asym1}}(n)=\begin{cases}\dfrac{1+\alpha}{2}-\dfrac{1-\alpha}{2}\cos\left(\dfrac{2\pi n}{2N_L-1}\right), & 0\le n\le N_L-1\\ \cos\left(\dfrac{2\pi(n-N_L)}{4N_R-1}\right), & N_L\le n\le N-1\end{cases} \quad (1)$$

where N is the window length, n is the time index, $N_L=\dfrac{5L}{6}$ and $N_R=N-N_L$. With $\alpha=0.08$, the asymmetrical taper, given by (1), consists of the first half of a traditional Hamming window taking up N_L samples, followed by a cosine window of length N_R. We denote this taper in this paper as *asymwind1*.

Another form of asymmetric taper can be derived from the DDR (double dynamic range) Hamming taper, used in the HASE (higher lag autocorrelation spectrum estimation) method [3], as follows:

$$w_{\text{asym2}}(N-n+1)=\begin{cases}w_{\text{ddr}}(n), & (c-N/2)<n\le(c+N/2)\\ 0, & \text{otherwise}\end{cases}$$

where c is the parameter used to shift the peak position of $w_{ddr}(n)$, N is the frame length, $w_{ddr}(n)$ is the DDR hamming window computed from a *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 at the end of the autocorrelation sequence.

Various steps for construction of *N*-length DDR hamming taper from a hamming window is also presented in Fig. 1. 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.

Here we use $c = 55$. This asymmetric taper will be denoted as *asymwind2*.

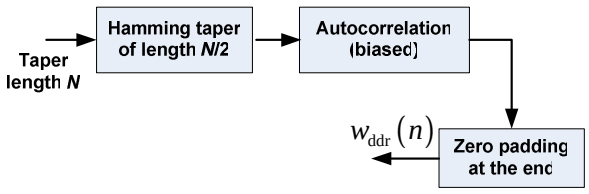
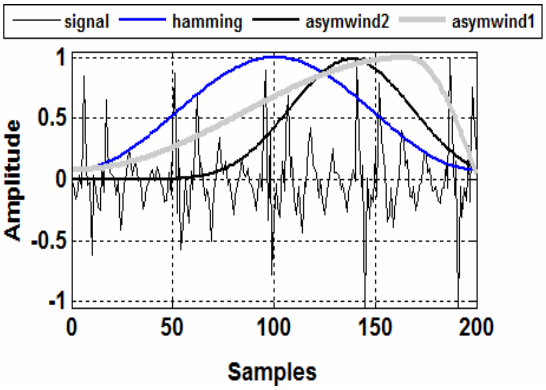


Figure 1: Block diagram showing various steps for the construction of DDR hamming taper.

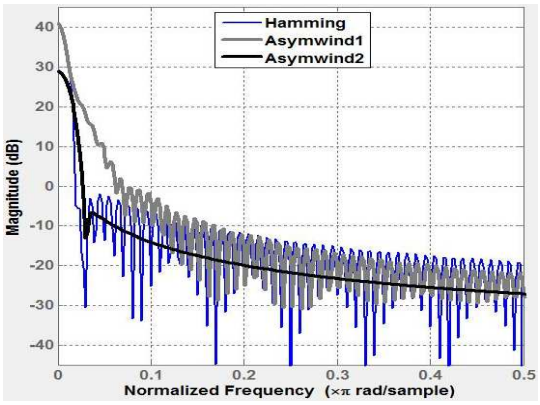
Fig. 2 presents a time and frequency domain comparison of the Hamming and asymmetric tapers asymwind1 & asymwind2 for frame length N = 200 samples. It is observed from fig. 2(b) that both the asymwind1 and asymwind2 have wider mainlobe widths and higher attenuation in the sidelobes than the Hamming taper.

Asymmetric tapers also result in shorter time delay [2], which is important for coding but not important for the recognition task alone.

Figs. 3 (a) & (b) show a comparison of the taper influence on the estimated power spectrum of a signal consisting of two equal and unequal pure tones, respectively. 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]. Since both the speaker and speech recognition share the same front-end, the same will be true for speaker recognition as well.

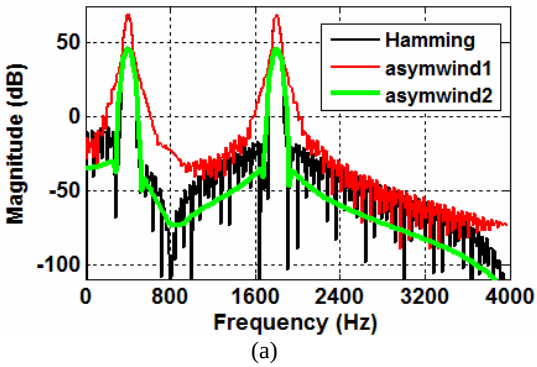


(a)

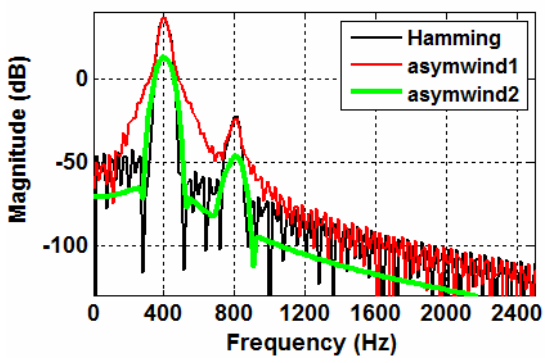


(b)

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



(a)



(b)

Figure 3: Comparison of taper influence on the estimated power spectrum of a simple two-tone signal when both tones have (a) equal amplitude, (b) unequal amplitude.

4. Experiments and Results

4.1. Experimental setup

We conducted experiments on the *extended core-core* condition of the NIST 2010 SRE extended list. The performance of the asymmetric tapers was evaluated using following the evaluation metrics: the Equal Error Rate (EER), the old normalized minimum detection cost function (DCF_{Old}) and the new normalized minimum detection cost function (DCF_{New}). DCF_{Old} and DCF_{New} correspond to the evaluation metric for the NIST SRE in 2008 and 2010, respectively.

4.1.1 Feature Extraction & UBM training

We use 20-MFCC features (including log-energy) augmented with their delta and double delta coefficients, making 60-dimensional MFCC feature vectors. The analysis frame length is 25 ms with a frame shift of 10 ms. Silence frames are removed using the VAD labels. After that we apply short-time Gaussianization, which uses a 300-frame sliding window, to normalize the features. We train a gender-independent, full-covariance Universal Background Model (UBM) with 2048-component Gaussian Mixture Models (GMMs). NIST SRE 2004 and 2005 telephone data were used for training the UBM.

4.1.2 Training and extraction of i-vectors

Our 800-dimensional gender-independent i-vector extractor was trained using the following data: LDC release of Switchboard II - phase 2 and phase 3, Switchboard Cellular - part 1 and part 2, Fisher data, NIST SRE 2004 and 2005 telephone data, NIST SRE 2005 and 2005 microphone data and NIST SRE 2008 interview development microphone data. Linear Discriminant Analysis (LDA) is used to reduce the dimension of the i-vectors (from 800 to 200) to handle telephone speech as well microphone speech. An optimal reduced dimension of 200 is determined empirically [7]. The length of the i-vectors is normalized to gaussianize the distribution of the i-vectors. For more details about the i-vector extractor, see [5-7].

4.1.3 Training the PLDA model

We train two *Probabilistic Linear Discriminant Analysis (*PLDA) models, one for the males and another for the females. These models were trained using all the telephone and microphone training i-vectors; then we combine these PLDA models to form a mixture of PLDA models in i-vector space [7].

4.2 Results

Speaker verification results are reported for five evaluation conditions corresponding to det conditions 1-5 in the evaluation plan [8]. Figures 3 and 4 depict the speech spectrograms for Hamming, asymwind1 and asymwind2 tapers under clean and additive noise conditions (white Gaussian noise, SNR = 5 dB), respectively. It is observed from the plotted spectrograms that, in clean condition, the asymmetric tapers do not distort the speech signal and under additive noise condition, compared to the hamming taper, asymmetric tapers show substantially lower noise in the spectrograms.

Tables 1-3 present the EERs, the $minDCF_{Old}$ and the $minDCF_{New}$, respectively, for the Hamming and asymmetric tapers. Fig. 6 presents a comparison of speaker verification accuracy of the symmetric Hamming and asymmetric tapers using DET (detection error trade-off) curves. It is observed from fig. 6 and from tables 1-3 that the asymmetric-shaped tapers performed better than the symmetric Hamming taper in the most of the det conditions, except in det1, det3 and det4 conditions for the male trials. Compared to the baseline Hamming taper, asymwind1 provides an average relative improvement (female-male & det1-det5) of 9.78%, 14.8%, and 5.32% in EER, $minDCF_{Old}$ and $minDCF_{New}$, respectively, whereas asymwind2 provides an average relative improvement of 9.6%, 8.86%, and 2.86% in EER, $minDCF_{Old}$ and $minDCF_{New}$, respectively, compared to the baseline.

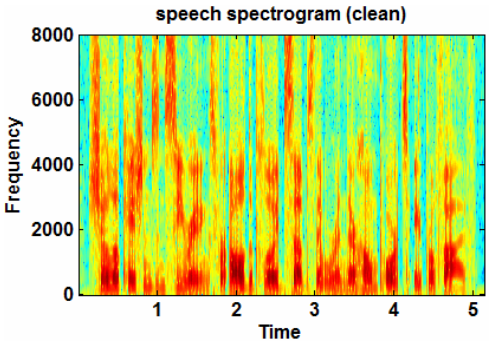
Robustness of asymmetric tapers, under additive noise environments, has been shown in [2], in the context of speech recognition. We did speech recognition experiments under additive noise distortion on AURORA-2 corpus and verified that asymmetric tapers provide better word accuracy than the Hamming taper, specifically in low SNR conditions. We expect that this will be true for speaker recognition as well.

4.2.1 Performance evaluation under additive noise

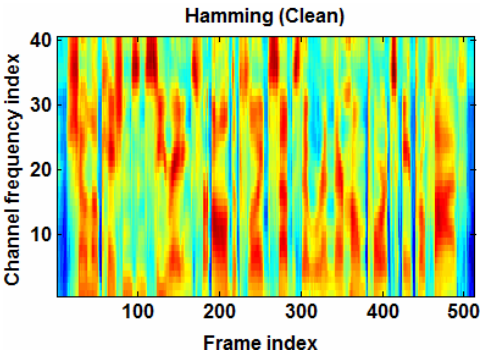
In order to evaluate the performance of the asymmetric tapers under additive noise environments, in the context of i-vector speaker verification, speech signals (test data only) are degraded with babble noise with a signal-to-noise ratio of 5 dB. I-vectors from the noise test data are extracted using the i-vector extractor, mentioned in section 4.1, that is trained with the clean training data. We train the Gaussian PLDA models using the clean training i-vectors and verification task is performed with the noisy i-vectors.

Tables 4-6 present the EERs, the $minDCF_{Old}$ and the $minDCF_{New}$, respectively, for the Hamming and asymmetric tapers, when the test signals are degraded with the babble

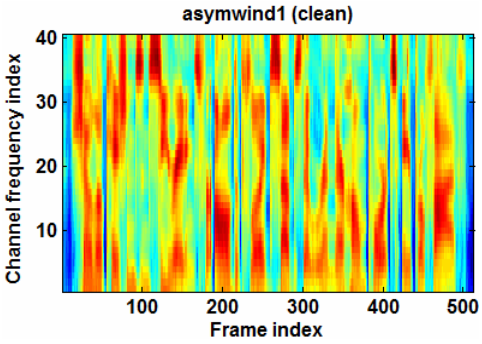
noise (SNR = 5 dB). Experimental results in babble noise condition show that asymmetric tapers are more robust to additive noise than the symmetric hamming window.



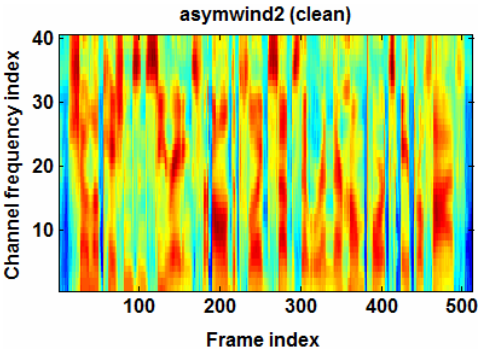
(a)



(b)

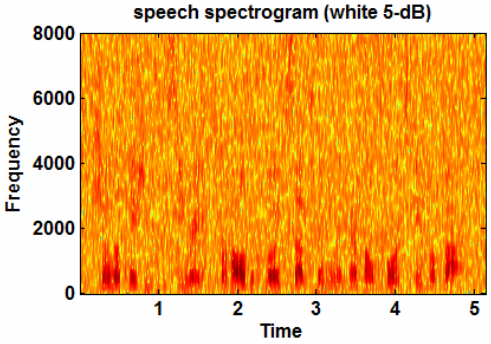


(c)

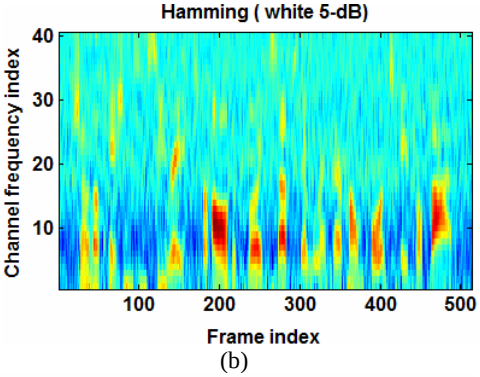


(d)

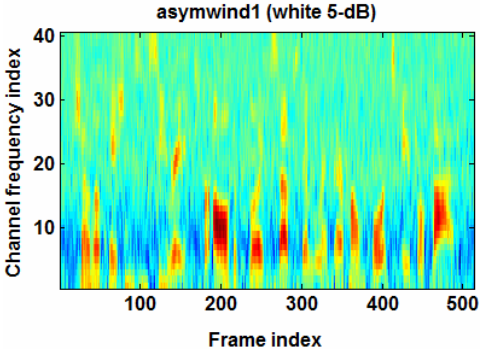
Figure 4: Comparison of speech spectrograms: (a) clean speech, (b) symmetric Hamming taper, (c) asymmetric taper asymwind1, and (d) asymmetric taper asymwind2.



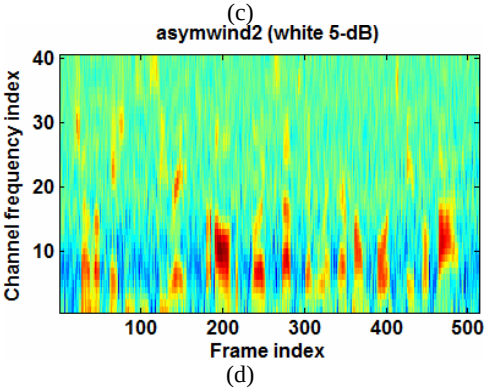
(a)



(b)

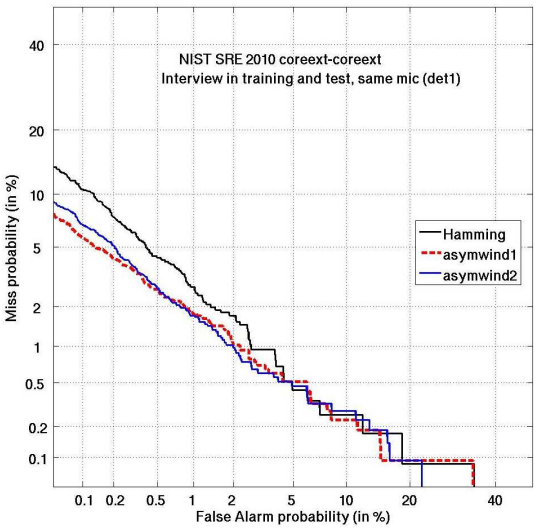


(c)

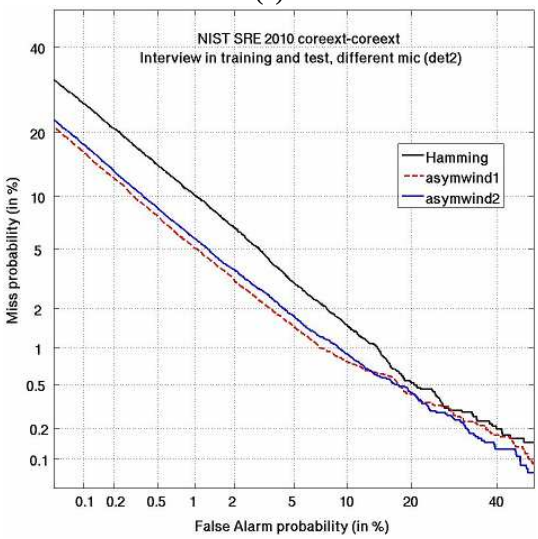


(d)

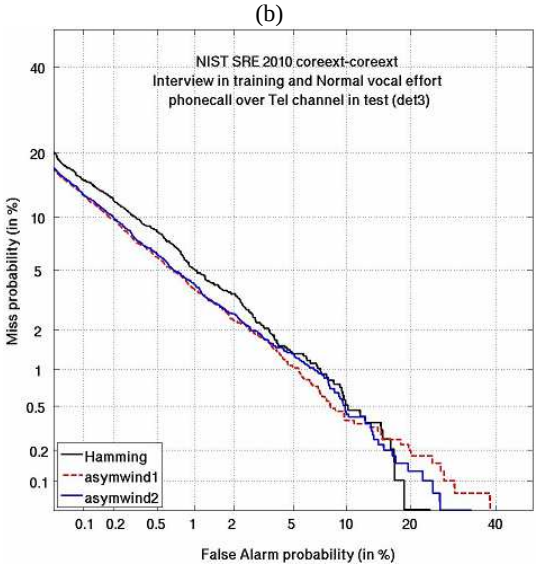
Figure 5: Comparison of speech spectrograms: (a) noisy speech (white 5 dB), (b) symmetric Hamming taper, (c) asymmetric taper asymwind1, and (d) asymmetric taper asymwind2.



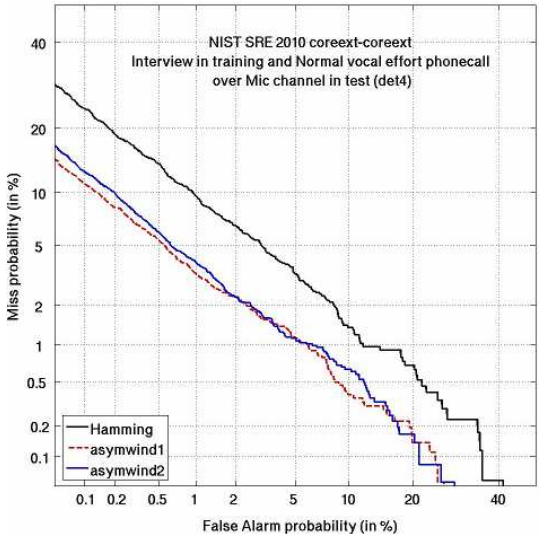
(a)



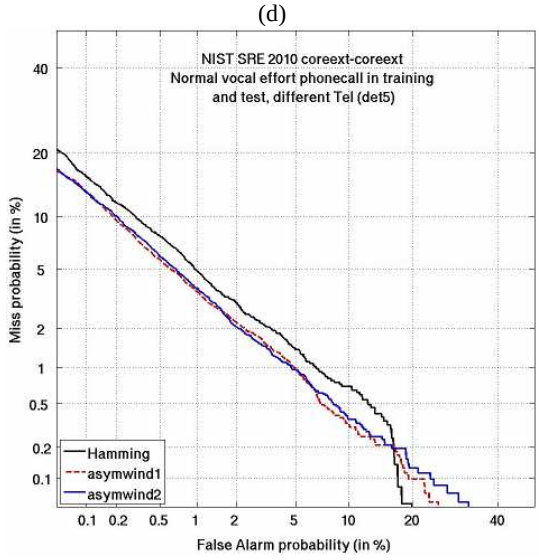
(b)



(c)



(d)



(e)

Figure 6: Effect of different feature tapers to speaker verification accuracy: (a) det1, (b) det2, (c) det3, (d) det4, and (e) det5. All features are extracted using 60-dimensional MFCC features, but the windowing methods vary. Asymmetric tapers seem to perform better in all det conditions. For this experiment we train a UBM with 2048-mixture components and the i-vectors are reduced to a dimension of 200 from 800.

Table 1: Male and female (det1-det5) speaker verification results for symmetric and asymmetric tapers measured by EER. For each row the best EER is in boldface. For this experiment we train a UBM using 2048-mixture components and the i-vectors are reduced to a dimension of 200.

EER (%)				
		Hamm	**Asym-wind1**	**Asym-wind2**
Female	det1	1.8	1.6	1.46
	det2	3.9	**3.06**	3.55

	det4	4.0	2.85	**2.73**
	det3	2.6	**2.04**	2.20
	det5	2.5	**2.15**	2.31
Male	det1	**1.0**	1.18	1.10
	det2	2.0	**1.95**	2.04
	det4	1.8	**1.53**	1.64
	det3	2.5	**2.16**	2.26
	det5	1.8	2.02	**1.67**

Table 2: Male and female (det1-det5) speaker verification results for symmetric and asymmetric tapers measured by $\text{minDCF}_{\text{Old}}$. For each row the best $\text{minDCF}_{\text{Old}}$ is in boldface. For this experiment we train a UBM using 2048-mixture components and the i-vectors are reduced to a dimension of 200.

$\text{minDCF}_{\text{Old}}$				
		Hamm	**Asym-wind1**	**Asym-wind2**
Female	det1	0.088	**0.07**	0.076
	det2	0.19	**0.16**	0.17
	det4	0.18	**0.13**	0.14
	det3	0.13	**0.10**	0.11
	det5	0.12	**0.11**	0.12
Male	det1	**0.045**	**0.045**	0.048
	det2	0.097	**0.085**	0.093
	det4	0.079	**0.062**	0.075
	det3	0.11	**0.095**	0.097
	det5	0.096	**0.088**	0.089

Table 3: Male and female (det1-det5) speaker verification results for symmetric and asymmetric tapers measured by $\text{minDCF}_{\text{New}}$. For each row the best $\text{minDCF}_{\text{New}}$ is in boldface. For this experiment we train a UBM using 2048-mixture components and the i-vectors are reduced to a dimension of 200.

$\text{minDCF}_{\text{New}}$				
		Hamm	**Asym-wind1**	**Asym-wind2**
Female	det1	0.30	0.26	**0.25**
	det2	0.54	**0.49**	0.52
	det4	0.52	**0.39**	0.43
	det3	0.41	**0.37**	0.38
	det5	0.39	**0.35**	0.38
Male	det1	**0.18**	0.20	0.19
	det2	0.37	**0.34**	0.35
	det4	**0.25**	0.26	0.27
	det3	**0.37**	0.40	0.40
	det5	0.32	**0.29**	0.31

Table 4: Male and female (det1-det5) speaker verification results under additive noise condition (**babble noise, SNR = 5 dB**) for symmetric and asymmetric tapers measured by EER. For each row the best EER is in boldface. For this experiment we train a UBM using 2048-mixture components and the i-vectors are reduced to a dimension of 200.

EER (%)				
		Hamm	**Asym-wind1**	**Asym-wind2**
Female	det1	2.73	**2.66**	2.83
	det2	8.15	**7.78**	8.1
	det4	4.12	**3.48**	**3.48**
	det3	3.20	**2.88**	3.20
	det5	2.60	**2.35**	2.4
Male	det1	1.90	1.9	**1.82**
	det2	**6.15**	6.2	**6.15**
	det4	2.37	**2.19**	2.43
	det3	**2.90**	2.91	**2.90**
	det5	2.07	1.97	**1.73**

Table 5: Male and female (det1-det5) speaker verification results under additive noise condition (**babble noise, SNR = 5 dB**) for symmetric and asymmetric tapers measured by $\text{minDCF}_{\text{Old}}$. For each row the best $\text{minDCF}_{\text{Old}}$ is in boldface. For this experiment we train a UBM using 2048-mixture components and the i-vectors are reduced to a dimension of 200.

$\text{minDCF}_{\text{Old}}$				
		Hamm	**Asym-wind1**	**Asym-wind2**
Female	det1	0.13	**0.12**	0.13
	det2	0.37	**0.36**	0.37
	det4	0.19	**0.17**	**0.17**
	det3	0.17	**0.15**	0.17
	det5	0.130	**0.12**	**0.12**
Male	det1	0.088	**0.087**	**0.087**
	det2	**0.26**	0.27	0.27
	det4	0.13	**0.11**	**0.13**
	det3	0.14	**0.14**	0.14
	det5	0.099	**0.094**	0.095

Table 6: Male and female (det1-det5) speaker verification results under additive noise condition (**babble noise, SNR = 5 dB**) for symmetric and asymmetric tapers measured by $\text{minDCF}_{\text{New}}$. For each row the best $\text{minDCF}_{\text{New}}$ is in boldface. For this experiment we train a UBM using 2048-mixture components and the i-vectors are reduced to a dimension of 200.

$\text{minDCF}_{\text{New}}$				
		Hamm	**Asym-wind1**	**Asym-wind2**
Female	det1	0.43	**0.42**	0.39
	det2	0.78	**0.76**	0.77
	det4	0.56	**0.50**	0.51
	det3	0.58	**0.52**	0.55
	det5	0.40	**0.35**	0.36
Male	det1	**0.34**	0.35	0.36
	det2	0.71	**0.69**	**0.69**
	det4	0.45	**0.43**	0.44
	det3	**0.56**	0.60	0.61
	det5	0.33	**0.31**	**0.31**

5. Conclusions

In this paper we incorporated two asymmetric-shaped tapers in the MFCC feature extraction process and compared their performances in the context of i-vector speaker verification. Experimental results under clean and noisy environments indicate that the asymmetric tapers outperformed the symmetric Hamming taper. Asymmetric tapers are found to be robust to additive noise. The largest relative improvements over the baseline were observed for conditions involving female trials.

Our future work includes the development of a generalized method to construct asymmetric taper from the existing symmetric tapers.

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] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, No. 4, pp. 788-798, May, 2011.

[6] P. Kenny, "Bayesian speaker verification with heavy tailed priors," *Proceedings of the Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, Jun. 2010.

[7] M. Senoussaoui, P. Kenny, N. Brummer, E. de Villiers, and P. Dumouchel, ''Mixture of PLDA models in I-vector space for gender independent speaker recognition,'' *Proceedings of INTERSPEECH 2011*, Florence, Italy, August 2011.

[8] National Institute of Standards and Technology, NIST Speaker Recognition Evaluation, http://www.itl.nist.gov/iad/mig/tests/sre/.