Performance of Spectral Amplitude Warp based WDFTC in a Noisy Phoneme and Word Recognition Tasks

R. Muralishankar
PES Centre for Intelligent Systems
Dept. of Telecommunication Engineering,
PES Institute of Technology, Bangalore, India.
muralishankar@pes.edu

H. N. Shankar
PES Centre for Intelligent Systems
Dept. of Telecommunication Engineering,
PES Institute of Technology, Bangalore, India.
hnshankar@pes.edu

Abstract

In this paper, we investigate the noise robustness of three features, namely, the warped discrete Fourier transform cepstrum (WDFTC, [1]), perceptual minimum variance distortionless response (PMVDR) and Mel-frequency cepstral coefficients (MFCC). We generate WDFTC and PMVDR features by all-pass based warping; we use spectral warping for MFCC. PMVDR and WDFTC use warped-LP and warped discrete Fourier transforms, respectively. We employ WDFTC, PMVDR and MFCC features in continuous noisy monophone and word recognition tasks using the TIMIT corpus. We also test these features on gender-specific monophone and word recognition tasks. Further, we employ spectral amplitude warping (SAW) in WDFTC feature extraction (WDFTC_SAW) and demonstrate enhanced robustness of this feature. We observe that SAW does not improve robustness for the MFCC and PMVDR features. Finally, we report the recognition performance and discuss many interesting properties of these features. Our study shows that the PMVDR and WDFTC_SAW achieve recognition performance superior to the MFCC and WDFTC in noisy conditions.

Index Terms:Robustness, Speech recognition, Warped Discrete Fourier Transform, Cepstrum, WDFTC, PMVDR, Spectral Amplitude Warping, WDFTC_SAW.

1. Introduction

Contemporary automatic speech recognition (ASR) systems perform satisfactorily when the test and training conditions are close. However, ASR performance degrades rapidly in various conditions such as background acoustical noise, stressed speech (*e.g.*, lombard), channel conditions, and speaker variability [2, 3]. With additive acoustical noise the problem is as pragmatic as it is challenging. Interestingly, humans do a far better job than ASR systems in noise, thus pointing to scope for further improvement [2]. Moreover, rapid proliferation of mobile phones concomitant with the large number of interactive voice applications being developed around them continues to present increasingly varied and complex acoustical backgrounds to the ASR systems [4].

In general, there are three approaches towards addressing the problem of noise robustness in ASR. The first incorporates a speech enhancement unit as a part of the feature extraction process, thereby presenting clean features to the ASR unit. Missing features approach [5], vector Taylor series approach [6] and spectral subtraction techniques [7] are instances in point. The second approach involves compensating the trained ASR models using techniques such as Parallel Model Combination (PMC) and dynamic Hidden Markov Model (HMM) variance compensation [8]. The third aims to develop new feature analysis methods which are relatively robust to distortion such as relative spectra (RASTA) [9] or cepstral mean subtraction (CMS). It is within the domain of robust features that in a companion paper we developed and introduced in the warped discrete cosine transform cepstrum (WDCTC) [10]. There, we benchmarked the new feature against the popular Mel-frequency cepstral coefficients (MFCC) in terms of its statistical properties and performance in simple recognition tasks [11]. A new feature representation called the Perceptual-MVDR (PMVDR) [12] has been proposed by Yapanel et al. They compute cepstral coefficients from the speech signal. Warping is incorporated directly into the DFT power spectrum. A variant of this feature, proposed in [13], uses warped-LP coefficients in generating warped-MVDR spectrum. The building block for all these features is the warping adopted to transforms or to the LP model.

In this paper, we propose a variant of MFCC, the warped discrete Fourier transform cepstrum (WDFTC); spectral warping is achieved using the warped discrete Fourier transform (WDFT). We then employ spectral amplitude warping (SAW) in addition to the frequency warping to generate WDFTC, i.e., WDFTC_SAW. We perform a comparative analysis of the features derived from the warping with the MFCC. The four features that we examine in our com-

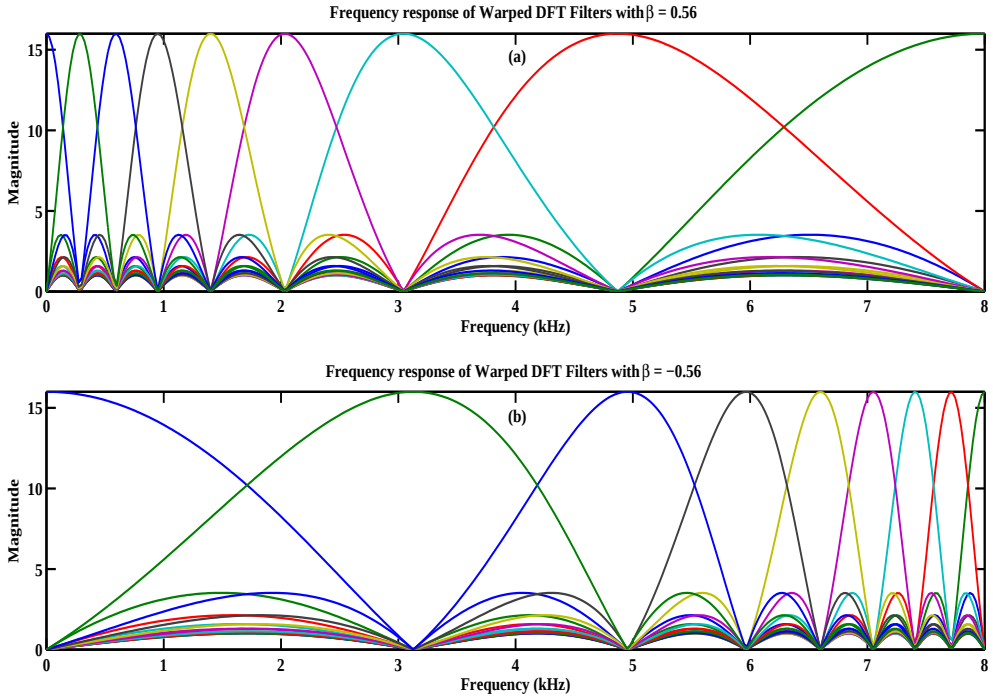

Figure 1. 16-Band Warped Filter bank shown for $f_s/2$. (a) $\beta = 0.56$ (b) $\beta = -0.56$

parative analysis are WDFTC_SAW, WDFTC, PMVDR and MFCC. We focus on studying their noise robustness properties. Particularly, we benchmark WDFTC_SAW, WDFTC and PMVDR against the MFCC in monophone and word recognition using the TIMIT corpus without using language models in monophone recognition and postprocessing of features like CMS. Thus the recognition performance is due to the features alone. We test the features in six different noise conditions – babble, car, fan, factory, tank and F-16 cockpit noise – at signal-to-noise ratio (SNR) from 0 dB through 20 dB. We report simulations reflecting phoneme recognition rates and recognition accuracies for monophone recognition.

In the second part of this work, we use Sphinx-III recognizer for word recognition task using TIMIT corpus. We report word error rate (WER) of the four features under clean and babble noise conditions. Finally, we highlight several interesting observations on noise robustness properties of warp-based features.

2. Warped-DFT Cepstrum

In this section, we briefly review WDFTC with a dyad of purposes in mind. First, to serve a didactic cause and second, to facilitate unfolding of the notations used in the sequel.

Let the N-point DFT of the input vector $[x(0), x(1), ..., x(N-1)]^T$ be given by $\{X(0), X(1), ..., X(N-1)\}$, where the frequency samples of the $z-$transform of the sequence evaluated at uniformly-spaced points $z = e^{\frac{j2\pi k}{N}}$, $0 \leqslant k \leqslant N-1$, on the unit circle are

$$X(k) = X(z)|_{z=e^{\frac{j2\pi k}{N}}} = \sum_{n=0}^{N-1} x(n)\, e^{\frac{j2\pi kn}{N}} \qquad (1)$$

for $k = 0, 1, ..., N-1$. For spectral analysis applications, DFT provides a fixed frequency resolution given by $2\pi/N$ over $[0, 2\pi]$. WDFT proposed in [14] is the most general form of DFT that can be employed to evaluate the frequency samples arbitrarily at distinct points in the z-plane. If z_k, $0 \leqslant k \leqslant N-1$, denote distinct frequency points in the z-plane, the N-point WDFT of the length-N sequence is given by

$$X_{WDFT}(k) = X(z_k) = \sum_{n=0}^{N-1} x(n) {z_k}^{-n}, 0 \leqslant k \leqslant N-1. \qquad (2)$$

Now, incorporating a nonlinear frequency resolution closely following the psychoacoustic Bark scale, yields enhanced representation for speech. Thus we warp DFT by an all-pass transformation $z^{-1} = A(z)$:

$$A(z) = \frac{-\beta + z^{-1}}{1 - \beta z^{-1}}, \qquad (3)$$

where β controls warping. It may be useful to note that Smith and Abel [15] have shown that for $\beta = 0.56$ warping closely resembles psychoacoustic Bark scale for sampling at $16kHz$. We also use $\beta = 0.56$ in computing the perceptually motivated speech spectrum. WDFT filters for length-16 is in Fig-

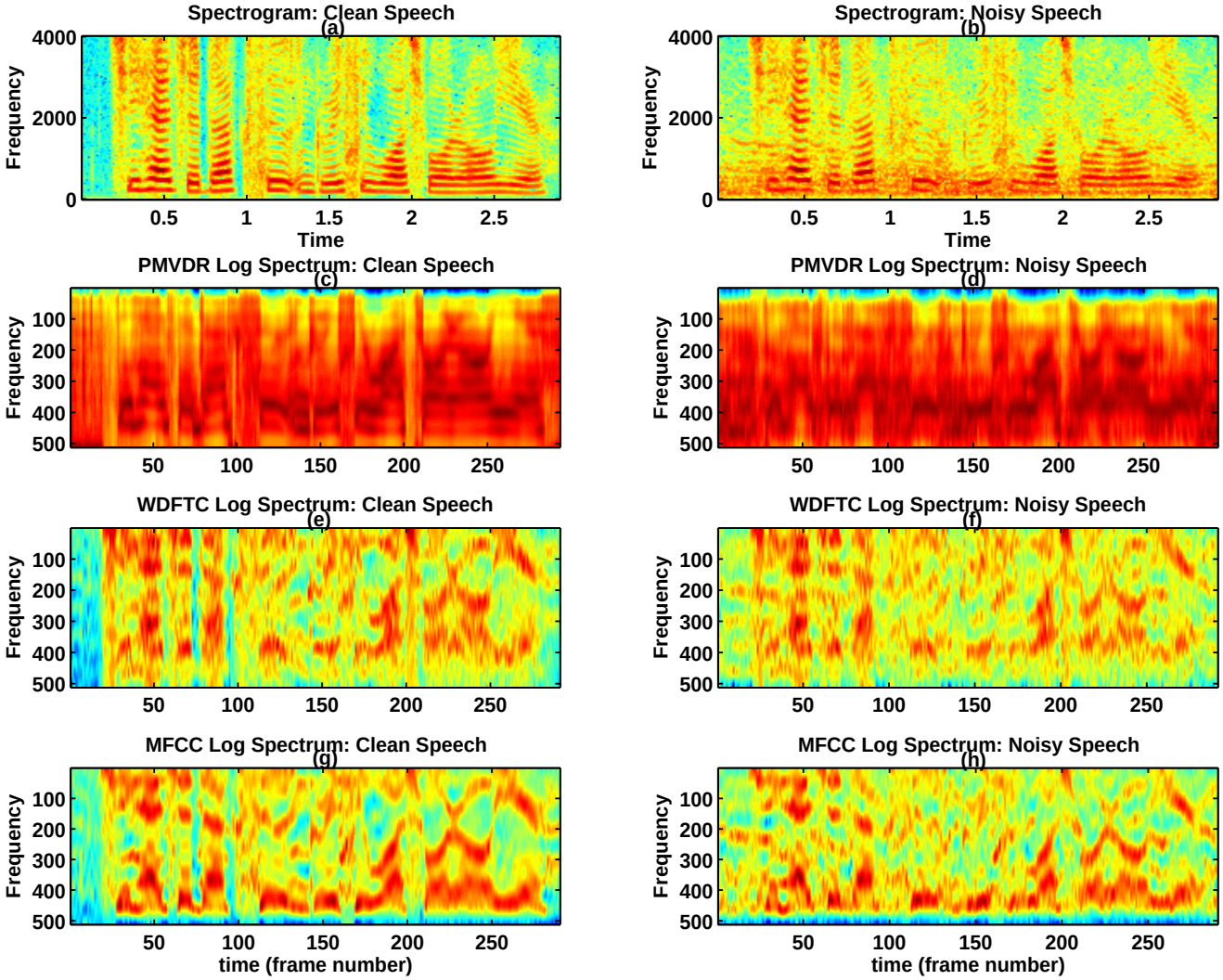

Figure 2. Illustrating the impact of 5 dB babble noise on PMVDR, WDFTC and MFCC Log spectra: (a) Spectrogram of clean speech, (b) Spectrogram of noise corrupted speech (a), (c) PMVDR log spectrum of clean speech (a), (d) PMVDR log spectrum of noise corrupted speech (a), (e) WDFTC log spectrum of clean speech (a), and (f) WDFTC log spectrum of noise corrupted speech (a, (g) MFCC log spectrum of clean speech (a), and (h) MFCC log spectrum of noise corrupted speech (a).

ure 1. Details of the implementation of WDFT are in [14]. The WDFTC algorithm is outlined in Algorithm 1.

3 MVDR Spectral Envelope Estimation

MVDR spectral estimation has been explored for speech parameterization [16, 17, 18]. Here, we present only the computational algorithm and general properties of MVDR and perceptual MVDR (PMVDR). In the MVDR spectrum estimation method, the power spectral density at ω_l is determined by filtering the signal by a distortionless FIR filter, $h(n)$, designed to minimize the output power while constraining the filter gain to unity at the frequency of interest, ω_l. This provides a lower bias with a smaller filter length. The parametric

Algorithm 1 Algorithm to compute the WDFTC

1: Obtain an N-point WDFT, $X_{WDFT}(k)$, $0 \leq k \leq N-1$ for a finite duration, real sequence $x(n)$, $0 \leq n \leq N-1$.
2: Compute $\zeta(k)$ of the WDFT coefficients:

$$\zeta(k) = |X_{WDFT}(k)|, \tag{4}$$

where $|.|$ evaluates the absolute value.
3: Compute the WDFTC $\widehat{x}(n)$ as

$$\widehat{x}(n) = (IDCT(\ln(\zeta(k))). \tag{5}$$

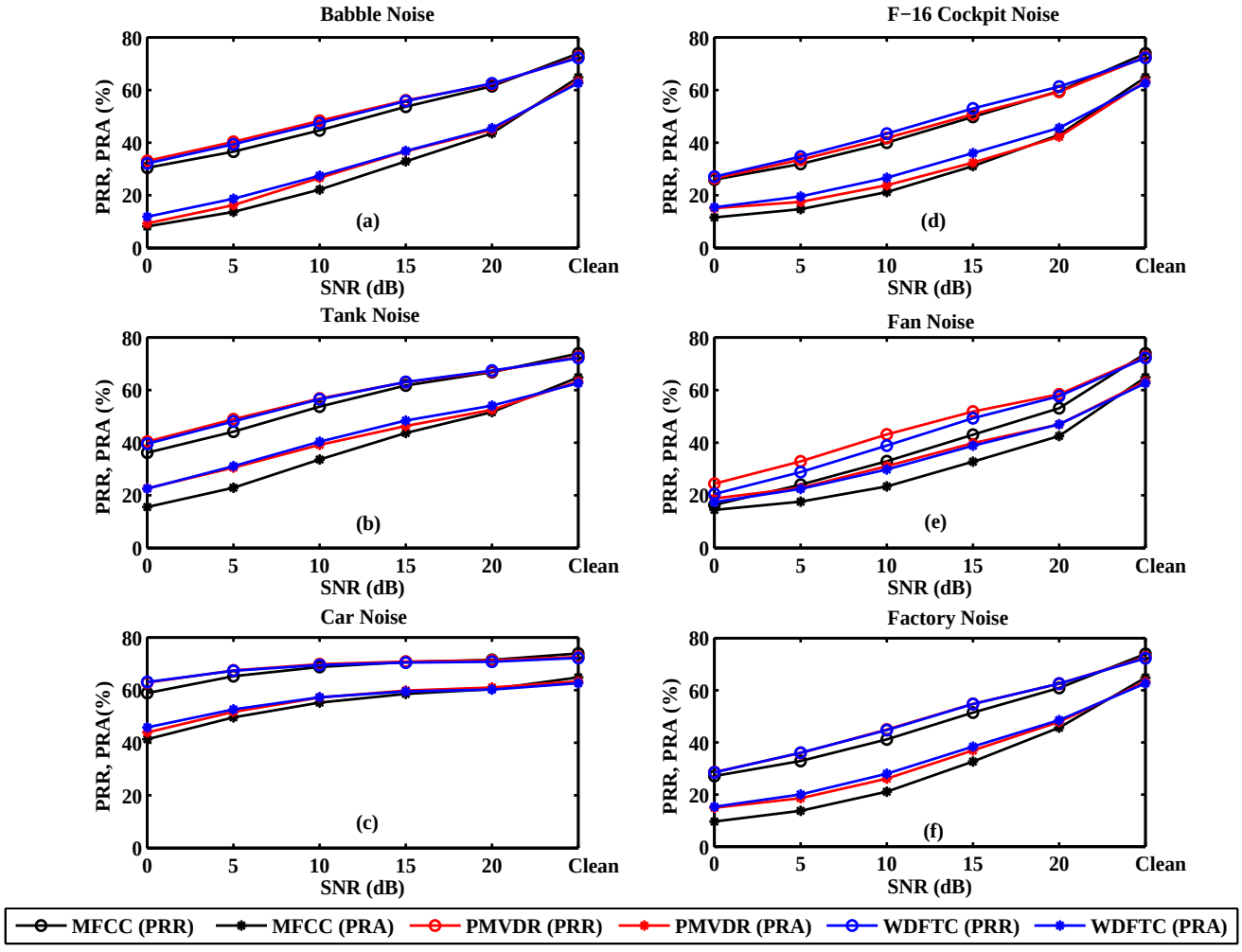

Figure 3. Illustrating the impact of different noises at various SNRs on MFCC, WDFTC and PMVDR phoneme recognition rate (PRR) and phoneme recognition accuracy (PRA): (a) babble, (b) tank, (c) car, (d) F-16 cockpit, (e) fan, and (f) factory noises. Acoustic Models trained on the entire TIMIT corpus using diagonal (DC) covariance.

form of the Mth order MVDR spectrum is given by

$$P_{MV}(\omega) = \frac{1}{\sum_{k=-M}^{M} \mu(k) e^{-j\omega k}} = \frac{1}{|B(e^{j\omega})|^2}. \quad (6)$$

The MVDR coefficients, $\mu(k)$, are obtained from a non-iterative computation using the LP coefficients a_k and prediction error variance, P_e.

$$\mu(k) = \begin{cases} \frac{1}{P_e} \sum_{i=0}^{M-k} L a_i a_{i+k}^*, & k = 0, ..., M \\ \mu^*(-k), & k = -M, ..., -1 \end{cases} \quad (7)$$

where $L = (M + 1 - k - 2i)$. We compute the MVDR envelope using LP coefficients of order M and the prediction error power, ϵ_M, as

$$S_{MVDR}(e^{j\omega}) = \frac{\epsilon_M}{\sum_{k=-M}^{M} \mu(k) e^{-j\omega k}}. \quad (8)$$

4. PMVDR Feature Extraction

MVDR spectrum exhibits useful properties such as low variance, low distortion and good spectral envelope matching across a wide range of pitch frequencies. Therefore it is widely considered as a robust speech parameterization technique in speech recognition. MVDR has been used in spectral estimation [17] and in envelope estimation [18]. A natural extension to the MVDR scheme is the incorporation of the perceptually motivated mel frequency into the otherwise linear frequency scale. In [18], perceptual information was incor-

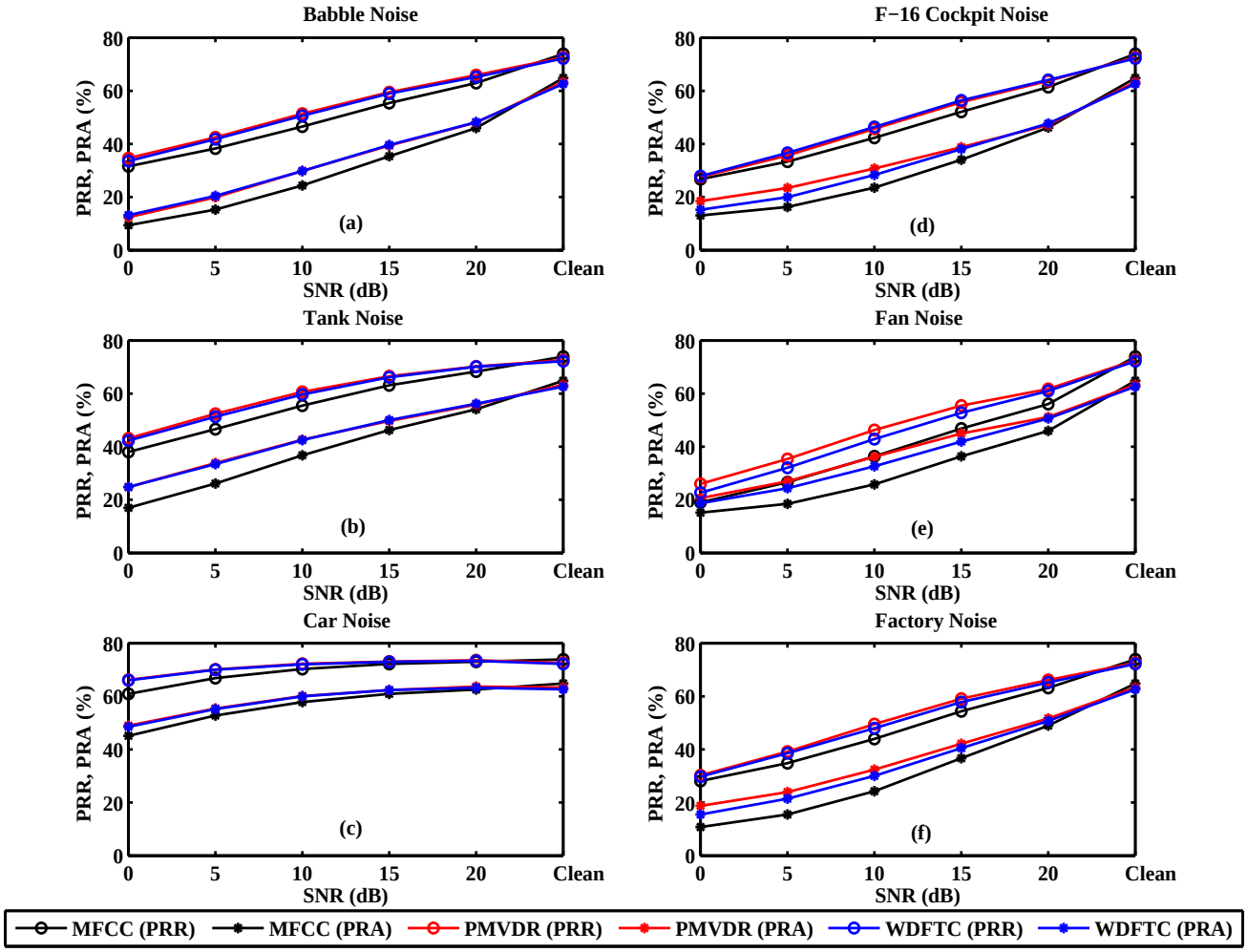

Figure 4. Illustrating the impact of different noises at various SNRs on MFCC, WDFTC and PMVDR phoneme recognition rate (PRR) and phoneme recognition accuracy (PRA): (a) babble, (b) tank, (c) car, (d) F-16 cockpit, (e) fan, and (f) factory noises. Acoustic Models trained on speech samples from a male speaker in the TIMIT corpus using diagonal (DC) covariance.

porated directly into spectral estimation by using mel-scaled filter banks. It was easily seen that the filter bank structure is only a rough approximation to the perceptual scale since it samples the perceptual spectrum at the center frequencies of the filter bank. Furthermore, the filter bank is less effective in completely removing the harmonic excitation information from the spectrum. Alternatively, the use of warping techniques is also popular in contemporary literature, e.g., incorporating warping directly into the DFT power spectrum [12], or the use of warped-LP coefficients in generating the warped-MVDR spectrum [13]. Presently, we generate the PMVDR features using the warped-LP coefficients.

5. Spectral Amplitude Warping (SAW)

Features discussed so far employ frequency warping with improved resolution in the lower frequency band of the power spectrum. SNR in this band is often higher than at high frequencies. The result of nonuniform resolution provided by the frequency warping in turn helps in improved phoneme recognition performance specifically under noisy conditions. It is well known that the acoustic models generated by clean speech performs poorly under mismatched conditions and in a simulation the same models trained on mismatched conditions perform superior to the former case. Generating such an acoustic model is possible only for a few simulated mismatched conditions. It is useful here to incorporate the general effect of noise (such as reduced peak to valley differ-

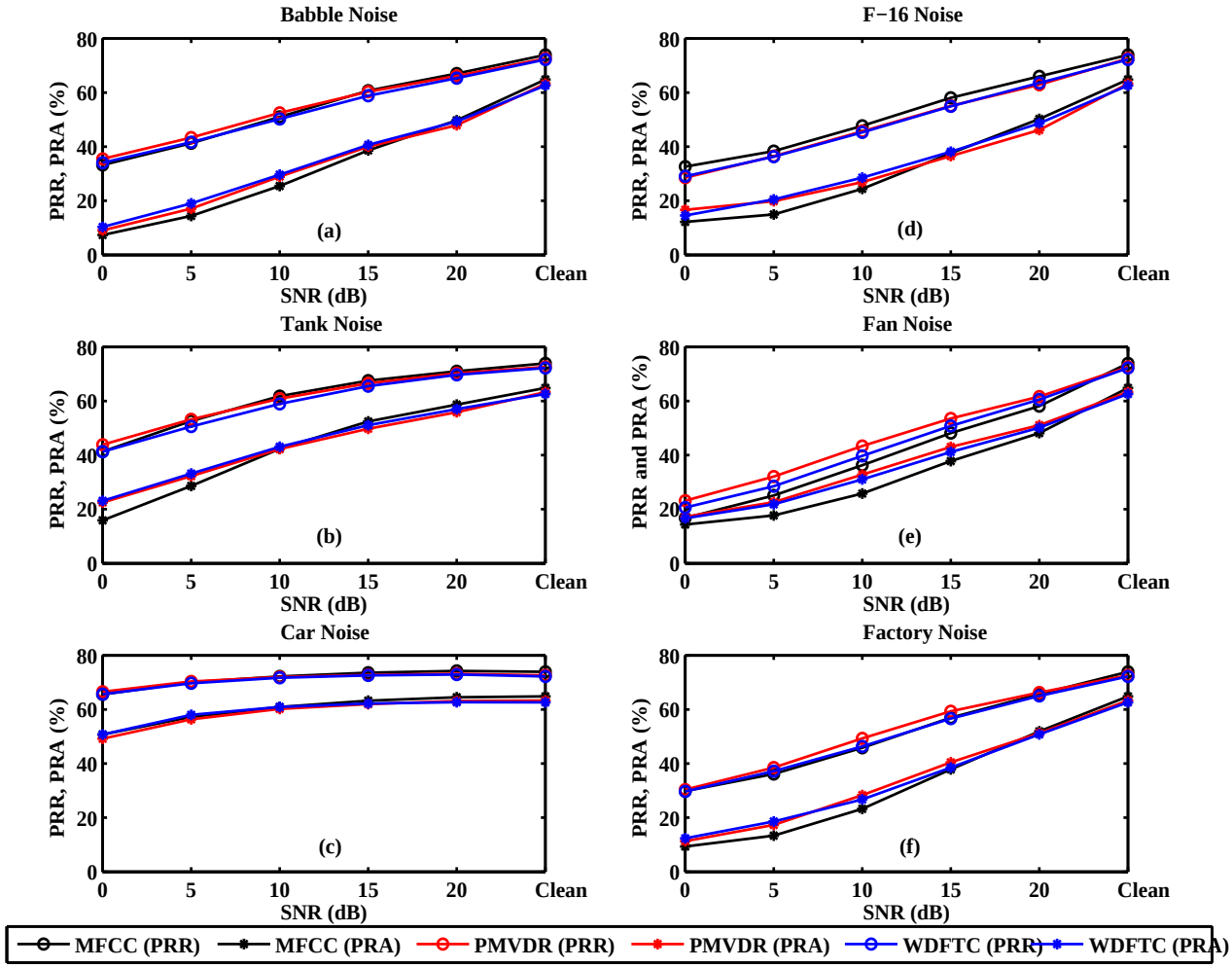

Figure 5. Illustrating the impact of different noises at various SNRs on MFCC, WDFTC and PMVDR phoneme recognition rate (PRR) and phoneme recognition accuracy (PRA): (a) babble, (b) tank, (c) car, (d) F-16 cockpit, (e) fan, and (f) factory noises. Acoustic Models trained on speech samples from female speaker in the TIMIT corpus using diagonal (DC) covariance.

ences and change in the formant positions) on speech spectrum used in front-end feature extraction.

In this paper, we model reduction in peak-to-valley difference using Spectral amplitude warping (SAW). SAW has been employed to shape coding noise in speech and audio coders [19] in pre- and post-processing blocks to provide a non-linear transformations to the signal short-time spectrum before and after encoding. Adopting SAW improves the noise shaping capability of an existing coder without modifying the coder itself. It is reported [19] that the output quality of G.722 wideband speech coder operating at 48 kbps is close to the same coder operating at 64 kbps.

Consider

$$X_w(k) = f_{nl}(X(k)), \tag{9}$$

where $X(k)$ and $X_w(k)$ are the DFT spectra of signals $x(n)$ and $x_w(n)$ respectively, and $f_{nl}(\cdot)$ is nonlinear. In [19],

$$X_w(k) = X(k)\frac{|X(k)|^{\alpha(k)}}{|X(k)|}, \tag{10}$$

where $\alpha(k)$. There, $\alpha(k) = 0.5\ \forall k$. This reduces the dynamic range of $|X_w(k)|$. The attenuation is relatively higher for larger $|X(k)|$. In a nut shell, the effects of this transformation are (a) attenuation of the formants; and (b) attenuation of the harmonics. The disparity between peak and valleys is reduced both at the formant and harmonic levels. We use this transformation to generate front-end features from the transformed spectra so as to distort the speech spectra equivalent to the effect of noise on it. Consequently acoustic models generated from these spectra perform better than those from clean speech spectra. In our monophone and con-

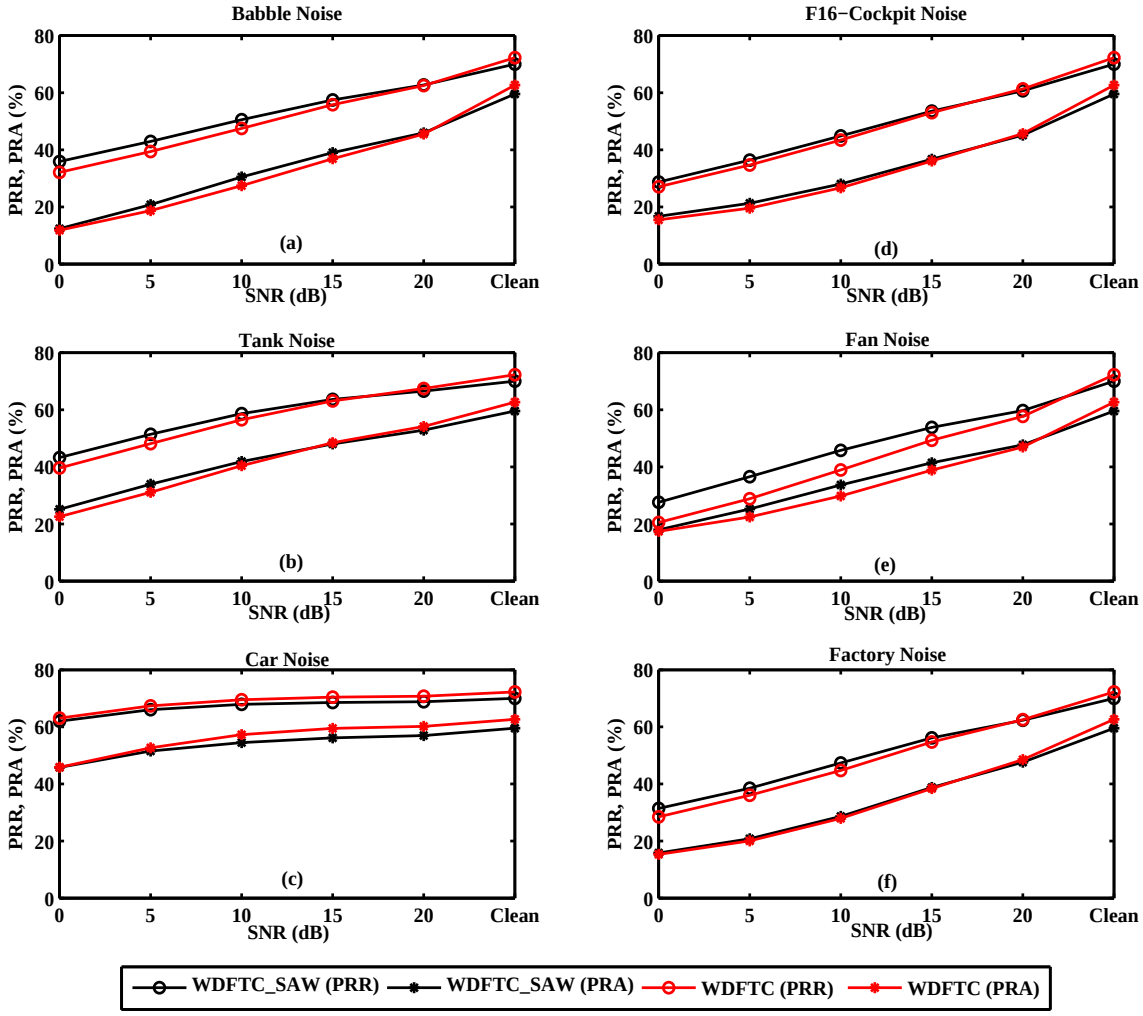

Figure 6. Illustrating the impact of different noises at various SNRs on WDFTC and WDFTC_SAW phoneme recognition rate (PRR) and phoneme recognition accuracy (PRA): (a) babble, (b) tank, (c) car, (d) F-16 cockpit, (e) fan, and (f) factory noises. Acoustic Models trained on the entire TIMIT corpus using diagonal (DC) covariance.

tinuous speech recognition we adopt this transformation with $\alpha(k) = 0.5$.

6. The Monophone Recognition Setup

We use GMM-HMM where the front-end features are the WDFTC or PMVDR or MFCC. We train and test the HMM recognizer using the HTK Toolkit [20], and the entire TIMIT corpus of 6300 sentences recorded from 630 speakers. We train phonetic HMMs using speech from the TIMIT train set with 462 speakers and test on the TIMIT test set with 168 speakers. It is common that the 61 TIMIT phonemes are mapped to a reduced set of 39 phonemes after training and testing, and the results are reported on this reduced set [21]. We train the HMMs using diagonal covariances (DCs). We use a 3-state HMM model for each of the 39 phonemes and for each state, a mixture splitting procedure under the DC setup [20]. The mixture splitting procedure starts with one mixture per state and it goes up to 8 mixtures in four steps with a re-estimation algorithm in each step. Finally, we present monophone performance under a DC setup for clean and noisy cases.

All the speech files are sampled at 16 kHz and pre-emphasized with $1 - 0.97z^{-1}$. They are then Hamming windowed. Speech signal is analyzed every 10 ms with a frame width of 25 ms. We generate 13-dimensional WDFTC, PMVDR and MFCC features from each speech frame including the zeroth coefficient. The WDFTC and MFCC are generated using the Algorithm 1 and a mel-scale triangular filter bank with 24 filter bank channels respectively. We append a

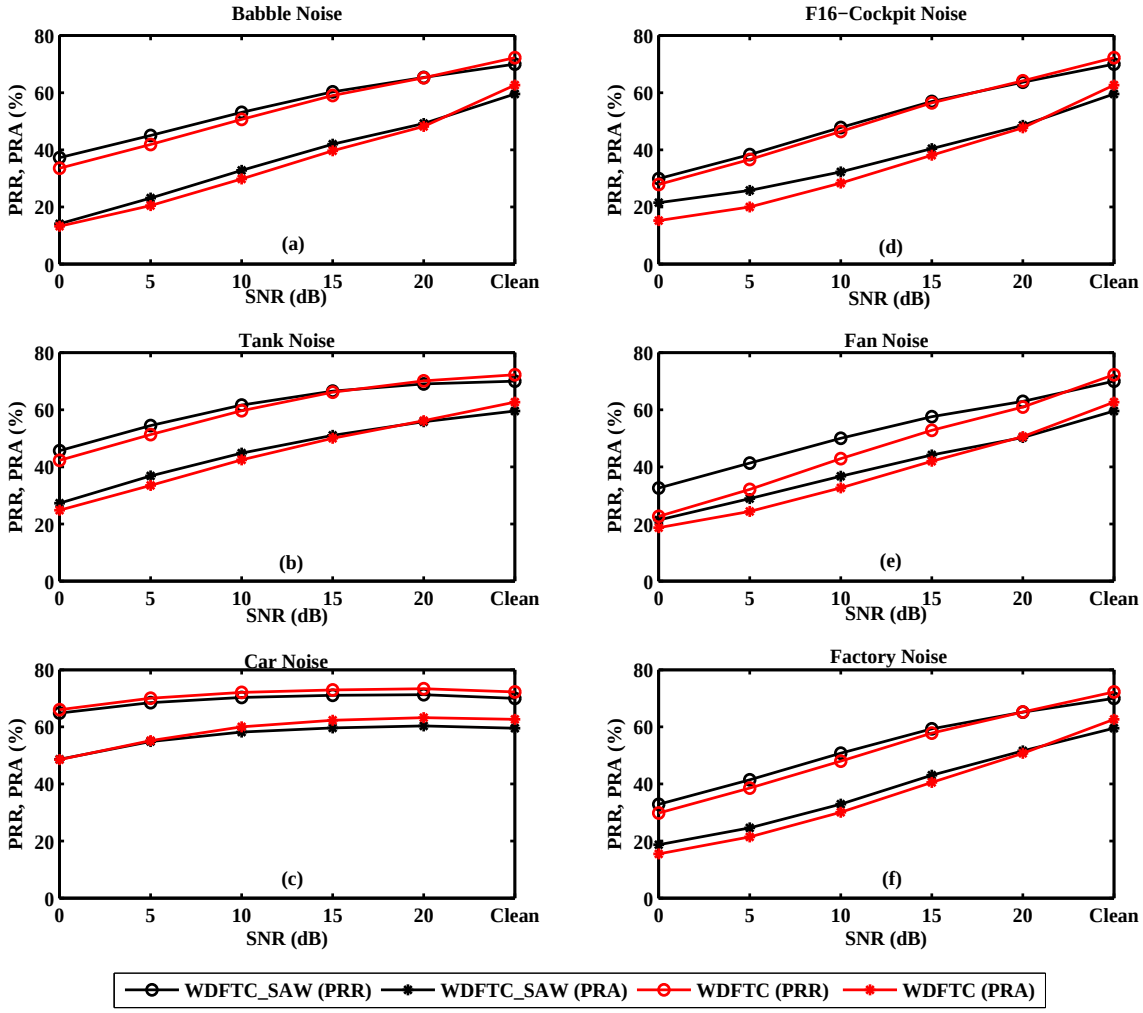

Figure 7. Illustrating the impact of different noises at various SNRs on WDFTC and WDFTC_SAW phoneme recognition rate (PRR) and phoneme recognition accuracy (PRA): (a) babble, (b) tank, (c) car, (d) F-16 cockpit, (e) fan, and (f) factory noises. Acoustic Models trained on speech samples from a male speaker in the TIMIT corpus using diagonal (DC) covariance.

26-dimensional delta and delta-delta cepstral features to 13-dimensional WDFTC and MFCC. We then employ the procedure in [22] to obtain delta and delta-delta for MFCC. We may employ dynamic spectral parameters [23] to compute delta and delta-delta for PMVDR and WDFTC.

7. The Word Recognition Problem

We employ the Sphinx-III speech recognizer [24]and the TIMIT speech database to evaluate the WDFTC_SAW front-end parameters uunlike traditional feature like MFCC, WDFTC and PMVDR. The utterances of the words in the transcriptions of the database are generated by the lexicon provided with the database. We use a phoneme set of 43 symbols including silence.

7.1 Generating Acoustic Models

For each of the above described features we train one acoustic model in two stages – encoding and decoding.

7.1.1 Encoding

Using the transcriptions in the database, the acoustic models are force-aligned against the transcription of the training data; thus pronunciations of the words with multiple phonetic representations are extracted. New acoustic models are synthesized by tieing the existing models following the same procedure. As before, the feature vectors are 13-D. We pre-emphasize the signal in the time domain with a factor of 0.97. The resulting speech waveforms are segmented into 25 ms frames with a step of 10 ms. The first and second deriva-

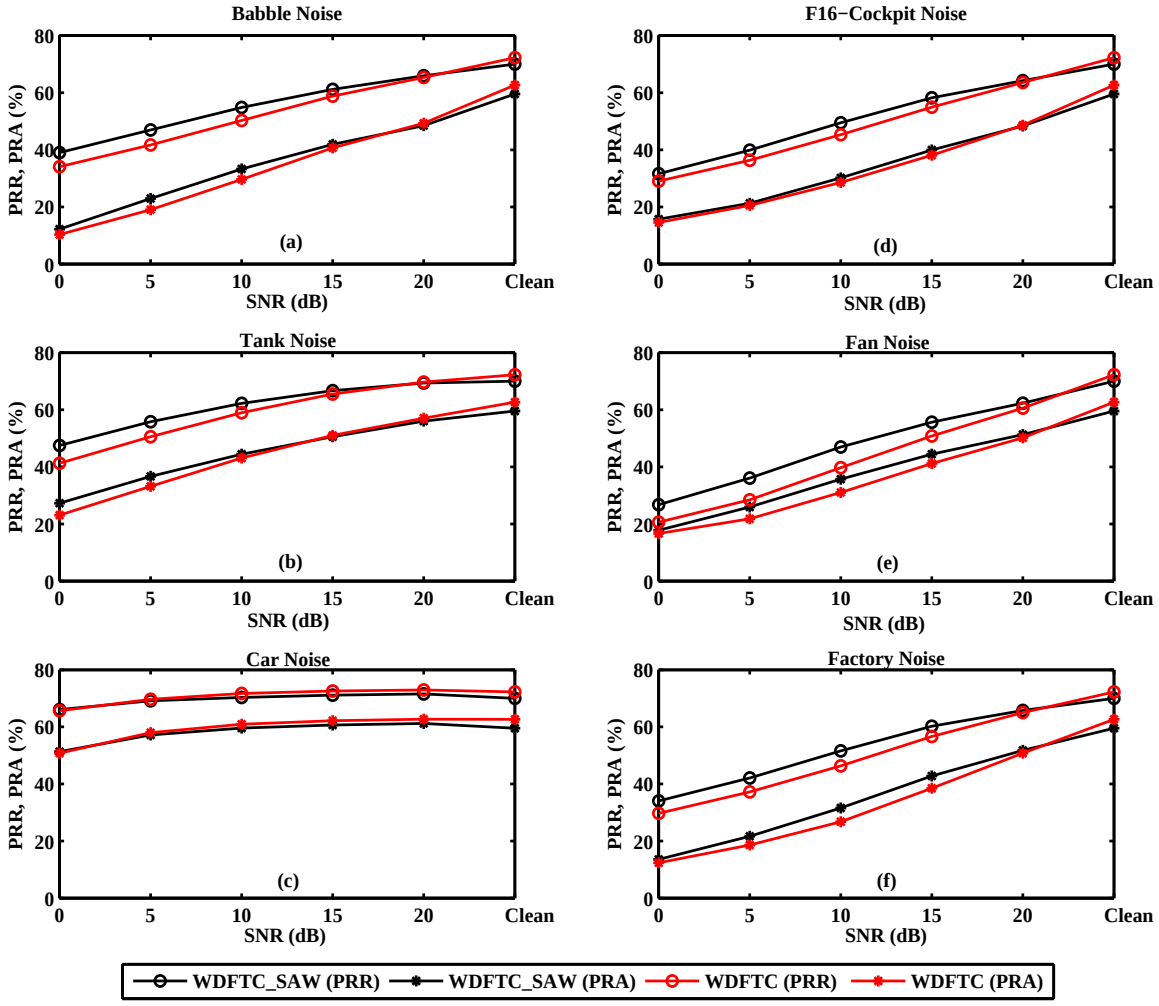

Figure 8. Illustrating the impact of different noises at various SNRs on WDFTC and WDFTC_SAW phoneme recognition rate (PRR) and phoneme recognition accuracy (PRA): (a) babble, (b) tank, (c) car, (d) F-16 cockpit, (e) fan, and (f) factory noises. Acoustic Models trained on speech samples from female speaker in the TIMIT corpus using diagonal (DC) covariance.

tive coefficients are appended to the static features in 13-D, finally resulting in feature vectors in 39-D.

Using the 39-D feature vectors so extracted we build context independent (CI) phone models for each of the 42 monophones on a 3-state Bakis-topology HMMs [25] with a non-emitting terminating state. CI phone models are trained by the Baum-Welch algorithm. Next, context-dependent (CD) untied triphone models are trained for every triphone that occurs at least 8 times in the training data. The CI model parameters initialize the parameters of the CD models. The CD models are now trained as above through the Baum-Welch algorithm. Decision trees determining similar HMM states of all untied models are built in order to be merge the common states or senones. In all, 1000 senones are trained. Decision trees were pruned to restrict the number of leaves to within a prespecified number of tied states. Every state of all the HMMs was modelled by a mixture of 16 Gaussians.

7.1.2 Decoding

We employ the Sphinx-III decoder. Throughout the recognition process the most probable sequence of words is considered as the recognized one. This result depends on two factors, namely, the acoustic score that the HMM models provide and the probability of the existence of the sequence of words called language weight. We use a 3-gram language model [26]. The training corpus of the language model included all the transcriptions of the database.

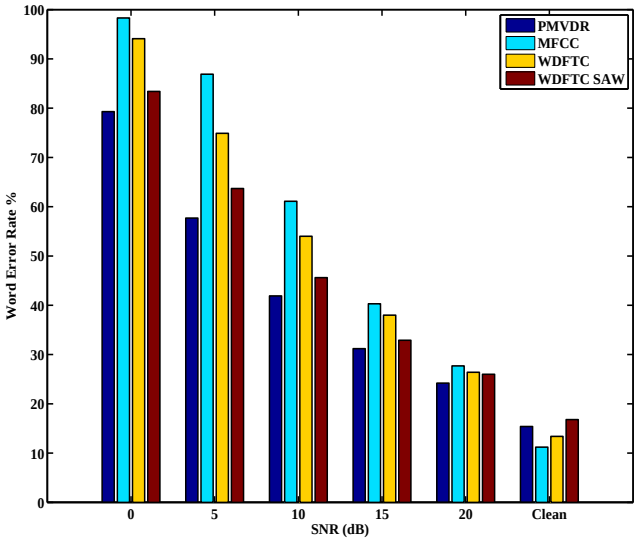

Figure 9. Word Error Rate (WER) of MFCC, PMVDR, WDFTC and WDFTC_SAW features under noise-free and Babble noise conditions

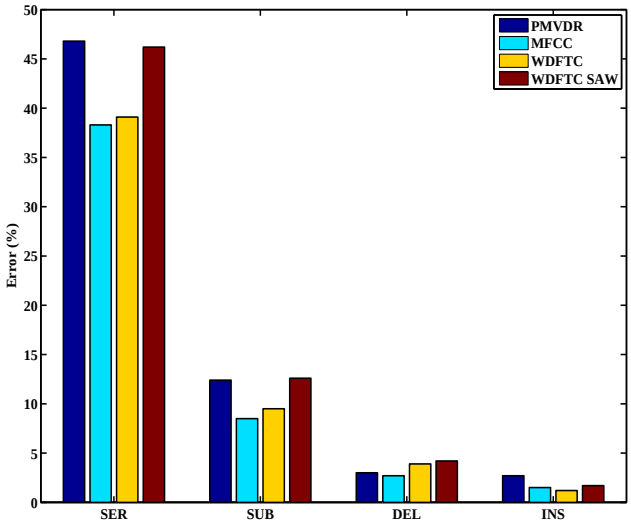

Figure 10. Sentence Error Rate (SER) and Errors in Word Recognition Performance

Table 1. Monophone Performance of MFCC, PMVDR, WDFTC and WDFTC_SAW on gender independent and specific speech samples from the clean TIMIT Corpus

Gender	All		Male		Female	
Feature	PRR	PRA	PRR	PRA	PRR	PRA
MFCC	73.9	64.8	75.8	65.3	76.2	67.8
PMVDR	72.6	63.3	75.0	65.3	74.6	65.6
WDFTC	72.2	62.6	74.9	65.9	74.3	65.3
WDFTC_SAW	70.0	59.5	72.5	62.8	72.5	62.9

8. Results and Discussions

Figure 2 shows the spectrograms of a TIMIT sentence (*She had your dark suit in greasy wash water all year*) and spectrograms of cepstral features from the PMVDR, WDFTC and MFCC in Fig. 2(a),(c),(e) and (g) respectively. Correspondingly, their noisy versions with 5 dB 'babble' noise are in Fig. 2(b),(d),(f) and (h) respectively. Robustness of WDFTC and PMVDR *vis-a-vis* MFCC is evident. Further, PRR and PRA for the PMVDR, WDFTC and MFCC features on the clean TIMIT corpus are outlined in Table 1 for gender independent and specific samples from the entire dataset. It can be seen from the tables that the MFCC performs slightly better than both the PMVDR and WDFTC with a 1-2% margin on the PRR and PRA for the entire dataset.

It may be noted from Figs. 3, 4 and 5 that the car, fan and tank noises are narrowband and stationary relative to the factory, F-16 cockpit and babble noises. From Fig. 3, it is observed that while the PMVDR and WDFTC exhibit a consistent degradation with falling SNR, their PRAs and PRRs performances are better than the MFCC. Further, it can be seen that the performance of PMVDR is marginally better than that of WDFTC and appreciably better than that of MFCC. The sensitivity of the acoustic models trained on MFCC in noisy backgrounds is clearly obviated by this observation. In comparison the PMVDR and WDFTC acoustic models seem to degrade gracefully with falling SNRs. The DC models of WDFTC and PMVDR are closer to each other in performance. MFCC exhibit poor performance under DC condition for most of the noise cases and SNRs. In the case of WDFTC and PMVDR, the DC models seem to give a better PRA for narrowband and broadband noise types, respectively. This is especially true for the low values of SNR. We proceed to test MFCC, PMVDR and WDFTC on gender-specific samples from the TIMIT database. We generate and test gender-specific cases by employing DC models; the phoneme recognition performance is shown in Figs. 4 and 5. We observe that PMVDR and WDFTC outperform MFCC in all noisy conditions except with F-16 cockpit noise.

We adopt WDFTC_SAW as a front-end feature for monophone and word recognition. Figure 6 presents the comparative performance of the WDFTC and WDFTC_SAW features. We observe that WDFTC_SAW outperforms WDFTC, particularly, under low SNR conditions except the case for car noise. Substantial improvement in PRR and PRA performance can also be seen for fan noise. Results from the gender-specific tests is shown in Figs. 7 and 8. Even in these tests SAW has improved the overall performance of WDFTC and the performance difference with the PMVDR is minimal. However, this improved performance is at the cost of marginal fall in noise-free condition.

In word recognition, we rank list word recognition performance of these features with clean and babble noise for SNR from 0 to 20 dB, vide Fig.9. It may be observed that the MFCC has lower WER compared to other features under clean conditions. However, PMVDR achieves lower WER under babble noise conditions for SNRs from 0 to 20 dB. The performance with WDFTC_SAW feature is better than WDFTC and MFCC. The performance difference between WDFTC_SAW and PMVDR is minimal.

Figure 10 shows sentence error rate (SER) and word recognition errors such as Substitutions (SUB), deletions (DEL) and Insertions (INS) from these four features under noise-free conditions. We can observe that MFCC has lowest SER, SUB, DEL and INS errors. PMVDR has heighest SER, substitution and insertion errors. Finally, adopting SAW to generate MFCC and PMVDR does not yield good results in both monophone and word recognition tasks.

In general, it can be easily concluded on the basis of all the results presented above that WDFTC_SAW, WDFTC and PMVDR outperform MFCC in different noise types and SNRs. This may be attributed to an all-pass based spectral warping in PMVDR, WDFTC and WDFTC_SAW feature generation. It has been shown [27] in addition that warping reduces the dynamic range of features and we believe that it plays an important role in noise robustness of the features. This is not very surprising as it is well known that the low-frequency spectrum of speech contains more energy than the high-frequency spectrum, and therefore it is more robust to additive noise. The high frequency spectrum in contrast is more susceptible to distortion by additive noise. Naturally, we achieve noise robustness by adopting a warping scheme which boosts at low frequency and bucks at high frequency. In other words, robustness is achieved by retaining reliable spectral bands in speech spectrum and eliminating bands where SNR is poor. The non-unitary nature of the WDFT which amplifies the low frequency spectrum and attenuates the high frequency part of the spectra also helps in highlighting the signal and downplaying the impact of the additive noise. It may be useful to note that the better performance of the PMVDR, WDFTC and WDFTC_SAW are attributed to their noise robustness and low feature variance.

9. Conclusion

In this paper, we presented exhaustive results illustrating the noise robust properties of the WDFTC, WDFTC_SAW and PMVDR features which have been benchmarked with MFCC. We introduced SAW to enhance robustness of WDFTC and demonatrate enhanced performance in noisy conditions. MFCC and PMVDR however failed to match the recognition results of these features alone. It was also shown that unlike MFCC, the WDFTC_SAW, WDFTC and PMVDR models degrade gracefully with falling SNR. It is important to note that we have not employed postprocessing of features like cepstral mean subtraction and variance normalization. Our experiments essentially shed light on some very useful properties of MFCC, WDFTC, WDFTC_SAW and PMVDR features that may be useful in improving the performance of current ASR systems in noise. Therein lies the take-home message.

References

[1] R. Muralishankar and D. O'Shaughnessy, "A comparative analysis of noise robust speech features extracted from all-pass based warping with mfcc in a noisy phoneme recognition," in *Proc. ICDT 2008*, Bucharest, Romania, July 2008, pp. 180–185.

[2] Douglas O'Shaughnessy, "Interacting with computers with voice: automatic speech recognition and synthesis," *Proc. of the IEEE*, vol. 91, no. 9, pp. 1272–1305, Sept. 2003.

[3] John H.L. Hansen and M.A. Clements, "Source generator equalization and enhancement of spectral properties for robust speech recognition in noise and stress," *IEEE Trans. on Speech & Audio Proc.*, vol. 3, no. 5, pp. 415–421, Sep. 1995.

[4] I Varga, S Aalburg, B Andrassy, S Astrov, J.G. Bauer, C Beaugeant, C Geissler, and H Hoge, "ASR in mobile phones - an industrial approach," *IEEE Trans. on Speech & Audio Proc.*, vol. 10, no. 8, pp. 562–569, Nov. 2002.

[5] Bhiksha Raj and R. M. Stern, "Missing feature approaches in speech recognition," *IEEE Signal Proc. Mag.*, vol. 22, no. 5, pp. 101–116, Sept. 2005.

[6] Pedro J. Moreno, Bhiksha Raj, and Richard M. Stern, "A vector taylor series approach for environment-independent speech recognition," in *ICASSP*, May 1996, pp. 733–736.

[7] Guillaume Lathoud, Mathew Magimai.-Doss, Bertrand Mesot, and Herve Bourland, "Unsupervised spectral subtraction for noise-robust ASR," in *ASRU*, Dec. 2005, pp. 343–348.

[8] Li Deng, Jasha Droppo, and Alex Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainity computed from a parametric model of speech distortion," *IEEE Trans. on Speech & Audio Proc.*, vol. 13, no. 3, pp. 412–421, May 2005.

[9] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. on Speech & Audio Proc.*, vol. 2, pp. 578–589, Oct. 1994.

[10] R. Muralishankar, A. Sangwan, and D. O'Shaughnessy, "Warped Discrete Cosine Transform Cepstrum: A new feature for speech processing," in *Proc. EUSIPCO*, Sep. 2005.

[11] A. Sangwan, R. Muralishankar, and D. O'Shaughnessy, "Performance analysis of the Warped Discrete Cosine Transform Cepstrum with MFCC using different classifiers," *MLSP*, pp. 99–104, Sept. 2005.

[12] U. H. Yapanel and John. H. L. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Communication*, vol. 50, no. 2, pp. 142–152, Feb. 2008.

[13] M. Wolfel, John. McDonough, and A. Waibel, "Warping and scaling of the minimum variance distortionless response," in *ASRU*, 2003, pp. 387–392.

[14] S. Bagchi and S. K. Mitra, *Nonuniform Discrete Fourier Transform and its Signal Processing Applications*, Norwell, MA: Kluwer, 1999.

[15] J. O. Smith III and J. S. Abel, "Bark and ERB bilinear transforms," *IEEE Trans. on Speech & Audio Proc.*, vol. 7, pp. 697–708, June 1999.

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

[17] S. Dharanipragada and B. D. Rao, "MVDR-based feature extraction for robust speech recognition," in *ICASSP*, May 2001, vol. 1, pp. 309–312.

[18] U. H. Yapanel and S. Dharanipragada, "Perceptual MVDR-based cepstral coefficients(PMCCs)for noise robust speech recognition," in *ICASSP*, May 2003, vol. 1, pp. 644–647.

[19] R. Lefebvre and C. Laflamme, "Spectral amplitude warping SAW for noise spectrum shaping in audio coding," in *Proc. ICASSP*, Apr. 1997, vol. 1, pp. 335–338.

[20] S. J. Young, *HTK Version 3.3: Reference Manual and User Manual*, Cambridge Univ. Engg. Dept.-Speech Grp., 2005.

[21] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. on Acoust. Speech & Signal Proc.*, vol. 37, no. 11, pp. 1641–1648, Nov. 1989.

[22] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. on Acoust., Speech, Sig. Proc.*, vol. 34, no. 2, pp. 52–59, Feb. 1986.

[23] S. M. Ahadi, H. Sheikhzadeh, R. L. Brennan, G. H. Freeman, and E. Chou, "On the use of dynamic spectral parameters in speech recognition," in *Proc. ISSPIT*, Dec. 2003, pp. 757–760.

[24] K. F. Lee, H. W. Hon, and R. Reddy, "An overview of the SPHINX speech recognition system," *IEEE Trans. on Acoust. Speech & Signal Proc.*, vol. 38, no. 1, pp. 35–45, Jan. 1990.

[25] R. Bakis, "Continuous speech word recognition via centi-second acoustic states," in *Proc. of ASA Meeting (Washington, DC)*, Apr. 1976.

[26] "The CMU-Cambridge statistical language modelling toolkit, v2," Available: http://www.speech.cs.cmu.edu/SLM/toolkit_documentation.html.

[27] R. Muralishankar, A. Sangwan, and D. O'Shaughnessy, "Theoretical complex cepstrum of DCT and warped DCT filters," *Proc. IEEE Signal Proc. Ltrs.*, vol. 14, no. 5, pp. 367–370, May 2007.