IEICE TRANS. INF. & SYST., VOL.E93–D, NO.8 AUGUST 2010

PAPER

A New Subband-Weighted MVDR-Based Front-End for Robust Speech Recognition*

Sanaz SEYEDIN[†a)], ***Student Member and*** **Seyed Mohammad AHADI**[†b)], ***Nonmember***

SUMMARY This paper presents a novel noise-robust feature extraction method for speech recognition. It is based on making the Minimum Variance Distortionless Response (MVDR) power spectrum estimation method robust against noise. This robustness is obtained by modifying the distortionless constraint of the MVDR spectral estimation method via weighting the sub-band power spectrum values based on the sub-band signal to noise ratios. The optimum weighting is obtained by employing the experimental findings of psychoacoustics. According to our experiments, this technique is successful in modifying the power spectrum of speech signals and making it robust against noise. The above method, when evaluated on Aurora 2 task for recognition purposes, outperformed both the MFCC features as the baseline and the MVDR-based features in different noisy conditions.

key words: *feature extraction, robust MVDR power spectral estimation, speech recognition*

1. Introduction

Speech recognizer systems are normally trained in certain conditions and tested in different environments, e.g. clean vs. noisy. This causes a mismatch in the training and test conditions. Therefore, the performance of automatic speech recognition systems degrades drastically in noisy environments. The type of the noise encountered in test conditions is usually not predictable. This makes robust speech recognition one of the most challenging areas in speech processing technology. Robust speech recognition methods may be classified into four main categories [1]:

1. Robust speech feature extraction.
2. Speech enhancement for improved recognition.
3. Model-based compensation for noise.
4. Model-based feature enhancement.

The main purpose of the first method is to find a set of parameters that are robust against the variations made by different noises on speech signals. This category, itself, can be further classified into two main divisions:

1. Extracting more robust features
2. Post-processing of the features for robustness

Among the robust speech feature extraction methods of the former type, modifying the power spectrum of the speech signal to make it robust against additive or convolutional distortions is more widely used, since the speech features are mostly derived either directly from the power spectrum of the speech signal or from a modified version of it. Mel Frequency Cepstral Coefficients (MFCC) [2], Perceptual Linear Prediction (PLP) [3], Differential Power Spectrum (DPS) [4], Autocorrelation Mel Frequency Cepstral Coefficients (AMFCC) [5], DCT and MVDR features [6], [7] are good examples in this category. Furthermore, feature normalization techniques are considered as one of the most significant group of methods in the post-processing of features. Histogram Equalization (HEQ) [8] and cepstral moment normalization methods [9] are the best examples in this division.

In the group of methods based on speech enhancement, the aim of noise reduction or increasing the signal to noise ratio is followed. Therefore, some initial information about the noise and speech signals are required. Spectral Subtraction (SS) [10] and Wiener filtering [11] are the most well-known approaches among speech enhancement methods.

In the case of model-based compensation for noise, speech models such as Hidden Markov Models (HMMs) are considered as the main framework, and the model parameters are improved during the recognition process in order to have better speech representation in noisy environments. Parallel Model Combination (PMC) [12] is known as one of the most important approaches in this category.

Model-based feature enhancement methods aim to extract clean features given noisy speech coefficients by considering different models for speech and noise. Vector Taylor Series (VTS) [13], and switching Linear Dynamic Model (LDM) [14] methods are some examples of this approach.

In this paper, we focus on the first division of robust feature extraction methods, i.e. modifying the power spectrum of the noisy speech signal in order to obtain a noise-robust power spectrum. Spectral estimation methods are either non-parametric or parametric [15]. The FFT-based periodogram is the most popular method of the former strategy, especially in speech recognition areas, while model identification and MVDR methods are among the most well-known approaches of the latter [15]. The model identification methods are classified into three divisions, namely Auto Regressive (AR), Moving Average (MA) and ARMA [15].

The FFT-based periodogram, as the fundamental step in extracting the traditional speech features, Mel-Frequency Cepstral Coefficients (MFCC), suffers from large bias and variance in estimating the power spectrum [7]. Bias is

Manuscript received January 19, 2010.
Manuscript revised April 1, 2010.

[†]The authors are with the Electrical Engineering Department, Amirkabir University of Technology, 424 Hafez Avenue, Tehran 15914, Iran.

[*]A part of this work was published in ICICS'09.

a) E-mail: sseyedin@aut.ac.ir
b) E-mail: sma@aut.ac.ir

DOI: 10.1587/transinf.E93.D.2252

mainly caused by the leakage of power from surrounding frequencies of the band-pass filter used to measure the power [7]. This problem can be solved by using DCT instead of FFT [6]. Moreover, large variance is due to employing a single sample in the power estimation process [7]. Both of these shortcomings have been addressed by the MVDR spectrum estimation method [6], [7]. During the use of stochastic models for acoustic modeling, such as hidden Markov models (HMMs), the state parameter distributions are usually modeled by mixtures of Gaussians. The parameters of these Gaussians should be estimated using features extracted from the training data. The bias and variance of the spectrum estimate affect such features used to extract Gaussian parameters that model the speech classes. Furthermore, increasing the level of noise in noisy signals enlarges the variance of the power spectrum and therefore deteriorates the recognition accuracy. For this reason, incorporating MVDR spectral estimation method as one of the well-known strategies in reducing the bias and variance of the spectrum estimates would be effective in extracting robust speech features. On the other hand, AR or linear prediction (LP) methods as other well-known approaches in extracting widely used features in ASR systems, namely LP and PLP, are ill-suited for accurate estimation of the power spectrum of voiced speech, especially high-pitch voices. This is due to the inaccurate spectrum matching that happens when the number of harmonics decreases. Therefore, since the LP-based envelope tends to follow the fine structure of speech spectrum in such voices, LP-based spectrum may also be sensitive to noise [7]. Thus, MVDR-based speech feature extraction may be considered as an appropriate approach for making ASR systems more robust against noise.

However, it has been shown that the MVDR method, by itself, is not as efficient as expected in low signal to noise ratios [6]. Therefore, an improvement in the feature extraction process based on MVDR is sought in this paper. Here, we present a new method to make the MVDR power spectrum more robust against additive noise. This robustness is achieved by modifying the distortionless constraint of the MVDR spectral estimation method by weighting the sub-band power spectrum values based on the sub-band signal to noise ratios. The recognition results show that this strategy is very helpful in extracting more robust features. The paper is organized as follows. In Sect. 2, we describe the new robust MVDR power spectrum estimation method. In Sect. 3, our proposed robust front-end is introduced. The experimental results are presented in Sect. 4. Finally, discussion and conclusions are given in Sects. 5 and 6, respectively.

2. Robust MVDR Spectral Estimation

Reducing the bias and variance of the estimated spectrum is the main purpose of MVDR spectral estimation. This goal is accomplished by designing an FIR filter, $h(n)$, which minimizes its output power subject to the constraint that its response at the frequency of interest, ω_l, has unity gain. This constraint, called the distortionless constraint, certifies passing the components of the input signal with the frequency of interest without any distortion through the filter. Moreover, the output power minimization precludes the leakage of power from surrounding frequencies, which reduces bias. The power of signal at the frequency of interest will be equal to the power of the filtered signal [7], [15]. Hence, computing power using all of the output samples decreases the variance. The MVDR filter is designed by solving the following constrained optimization problem [7]:

$$\min_h \mathbf{h}^H \mathbf{R}_{L+1} \mathbf{h} \quad \text{subject to} \quad \mathbf{v}^H(w_l)\mathbf{h} = 1 \tag{1}$$

which results in:

$$\mathbf{h}_l = \frac{\mathbf{R}_{L+1}^{-1}\mathbf{v}(\omega_l)}{\mathbf{v}^H(\omega_l)\mathbf{R}_{L+1}^{-1}\mathbf{v}(\omega_l)} \tag{2}$$

where $\mathbf{v}(\omega) = [1, e^{j\omega}, e^{j2\omega}, \cdots, e^{jL\omega}]$, $\mathbf{R}_{L+1}$ is the $(L+1)\times(L+1)$ Toeplitz autocorrelation matrix of the data, and $\mathbf{h} = [h_0, h_1, \cdots, h_L]^T$. The MVDR spectrum for all of the frequencies is then computed by [7]:

$$P_{\mathrm{MV}}(\omega) = \frac{1}{\mathbf{v}^H(\omega)\mathbf{R}_{L+1}^{-1}\mathbf{v}(\omega)} \tag{3}$$

According to the distortionless constraint in (1), the filter responses at all frequencies have unity gain, and therefore, they contribute to the final result with the same weighting. Consequently, if some of the frequencies are corrupted by noise, the resulting MVDR power spectrum at those frequencies will also be deteriorated. For this reason, in order to make the MVDR spectrum robust against noise, we proposed to modify this constraint such that the response of the filter at the frequency of interest has a gain which is determined by the signal to noise ratio at that frequency, instead of a unity gain. In other words, the higher the SNR at a certain frequency, the larger the gain we assign to that frequency. The robust distortionless constraint, explained above, assures that the components of the input signal at the frequencies least affected by noise, pass through the filter with larger weights, while the others get smaller weights. We assign:

$$\mathbf{v}^H(\omega_l)\mathbf{h} = w(\omega_l) \tag{4}$$

where

$$w(\omega_l) = \frac{S(\omega_l)}{N(\omega_l)} \tag{5}$$

where $S(\omega_l)$ and $N(\omega_l)$ are the clean signal and noise at the frequency of interest, ω_l, respectively. Therefore, the robust MVDR spectrum for all frequencies will be computed by:

$$P_{\mathrm{RMVDR}}(\omega) = \frac{w(\omega)^2}{\mathbf{v}^H(\omega)\mathbf{R}_{L+1}^{-1}\mathbf{v}(\omega)} \tag{6}$$

This process is the same as weighting the power spectrum value at the frequency of interest based on the ratio of the energy of the signal to the energy of noise at that frequency.

3. The Proposed New Front-End: RPMCC Features

It has been shown that extracting MVDR features from the warped power spectrum, i.e incorporating the PLP structure in extracting Perceptual MVDR-based Cepstral Coefficients (PMCC), gives better recognition results [7]. This is due to the fact that exploiting the perceptual information always improves the speech recognition systems. The flow diagram for PMCC parameter extraction is given in Fig. 1 (a). The equal loudness curve and power law of hearing blocks are according to [3]. In this paper, the warped power spectrum is obtained by applying the conventional triangular Mel-based filter-bank to the FFT-based periodogram. Then the warped MVDR power spectrum is computed from the so-called Mel-warped spectrum similar to the way parameters are calculated in PMCC, after applying weighting to sub-bands. Then, the cepstral features are calculated by applying IFFT to the Mel-scale MVDR log-spectrum [6], [7]. The Mel-warped spectrum is also known as sub-band spectrum in the area of speech recognition.

For the same reason, the proposed robust MVDR features are extracted from the robust sub-band MVDR power spectrum, calculated by weighting the sub-band power spec-

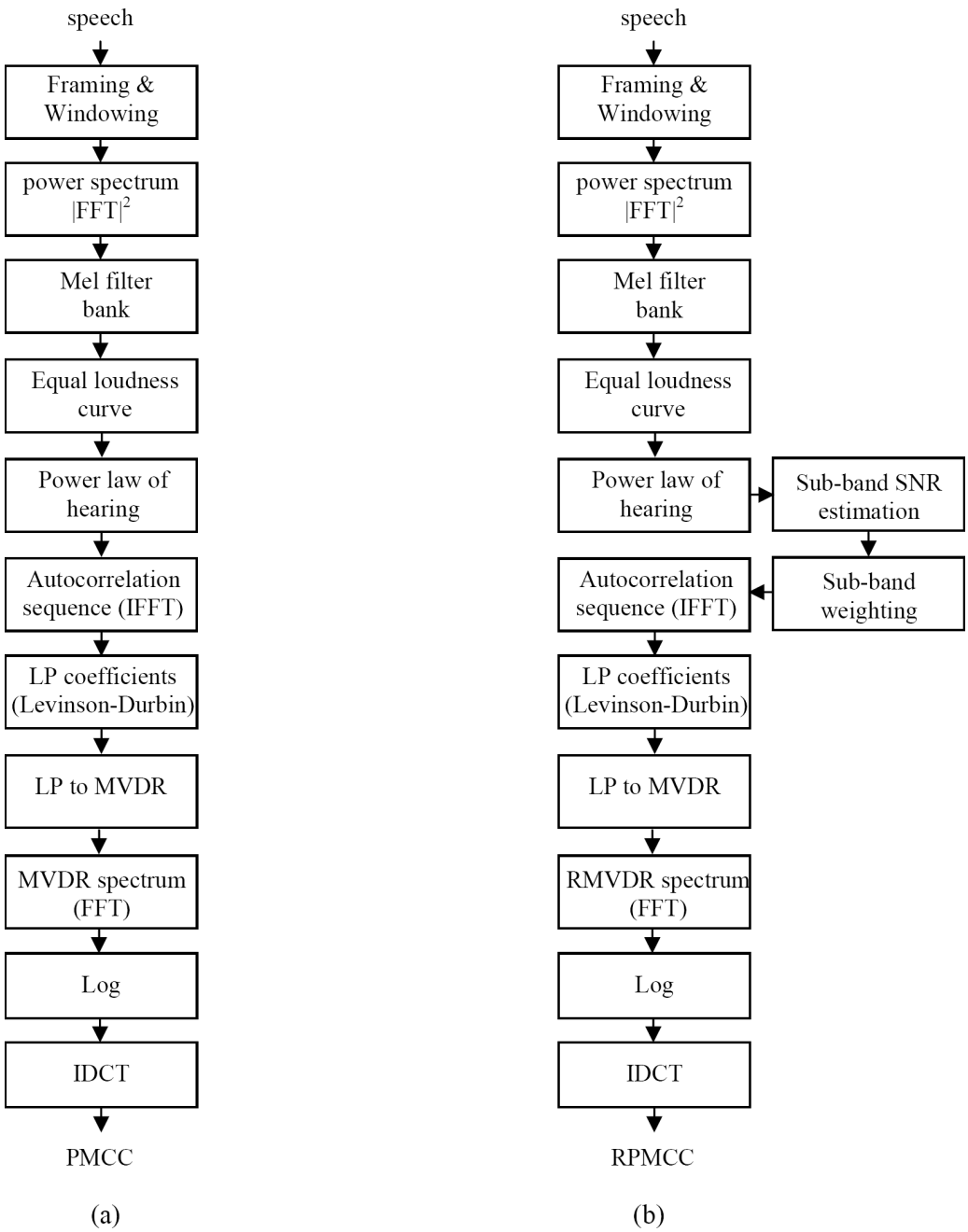
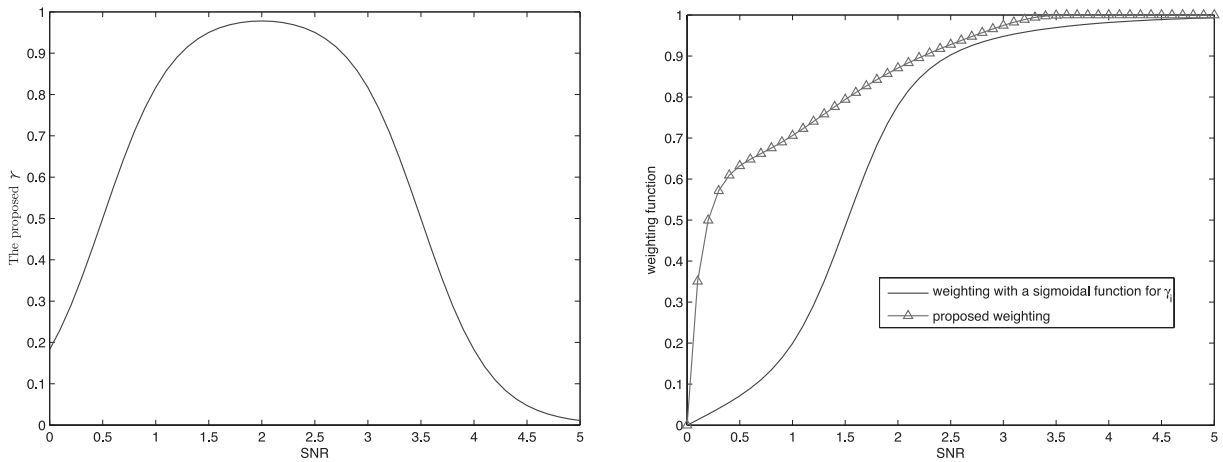


Fig. 1 (a) PMCC front-end [7], (b) The proposed front-end.



(a) (b)

Fig. 2 (a) The proposed γ as the gain controlling the steepness of the weighting function, (b) Comparison of the proposed weighting function with the one obtained by a sigmoidal function for γ.

trum values based on the sub-band signal to noise ratios. Our experiments showed that using the raw sub-band signal to noise ratios as the weighting factors does not lead to sufficient recognition accuracies in low SNRs. Therefore, we decided to employ the experimental findings of psychoacoustics in defining a suitable weighting function with values between zero and one [16], [17]:

$$w_i^2 = 1 - \exp\left(-\frac{\mathrm{SNR_i}}{\gamma_i}\right) \tag{7}$$

where $\mathrm{SNR_i}$ is the signal to noise ratio computed from the ratio of the energy of noisy signal to noise in the i^{th} mel frequency sub-band and γ_i is the gain that controls the steepness of the weighting function. The use of this weighting function may be justified by the experimental findings of psychoacoustics on the masking effects of background noise on the perceived loudness [16], [17]. In [17], it has experimentally been shown that the masking sound (noise) not only shifts the audible loudness, known as the masked threshold, but also produces a masked loudness function that has steeper slope than the unmasked one for low sound pressure levels. Nevertheless, in large sound pressure levels, the masked loudness functions almost reach the unmasked one. Therefore, we suggested using an exponential weighting function which follows this property in low and high signal to noise ratios. Different functions have been tested in order to find the optimum γ_i which gives the best recognition accuracy. According to the experimental results given in Sect. 4, using a high γ, which is close to one, gives better recognition results in medium SNRs, while a lower γ is preferred in low and high SNRs. Thus, a function that corresponds to this property can work better than the sigmoid function used to define the sub-band compression and weighting functions in [16], [18] respectively. Although a Gaussian function is among those that assure these characteristics, our experimental results to find an optimum function for γ_i, showed that a wider function with flat peak performed better than a Gaussian regarding recognition accuracies. Therefore, we applied a function which is made up of the difference between two sigmoidal functions, i.e.

$$\gamma_i = \frac{1}{1 + \exp(-3(\mathrm{SNR_i} - 0.5))} - \frac{1}{1 + \exp(-3(\mathrm{SNR_i} - 3.5))}. \tag{8}$$

This function has been shown in Fig. 2 (a). Figure 2 (b) compares the weighting function obtained by applying our proposed γ_i with that of the sigmoidal function suggested in [16], [18]. According to this figure, a larger weight is always assigned in our proposed algorithm in comparison with using a sigmoidal function for γ_i. Furthermore, it is worth mentioning that assigning a smaller γ in lower SNRs means setting larger weights for the mentioned signal to noise ratios in comparison with choosing a larger γ. Therefore, we will save more information in low SNRs by assigning smaller γ compared to medium SNRs. This is due to the fact that the probability of having error in SNR estimation for low SNRs is more than the medium ones, and therefore, if we assign smaller weights, we will lose more information in case of estimating the SNRs inaccurately. Moreover, since the information in high SNRs is more reliable, we also choose smaller γ for high SNRs in order to assign larger weights. In addition, this γ_i makes the sub-band weights smoother in comparison with the sigmoid function in [16], [18]. This fact has been shown in Fig. 3. The smoother variations of the weights with SNR assure the robustness of our proposed method. This robustness is achieved because our weighting function follows the SNR values more smoothly; and therefore, is not as susceptible as the previous weighting function suggested in [16], [18] to errors in SNR estimation.

In order to improve the performance of our algorithm against non-stationary noises, the noise power spectrum is estimated by a simple updating algorithm where the first few non-speech frames are considered as the initial noise values [19]:

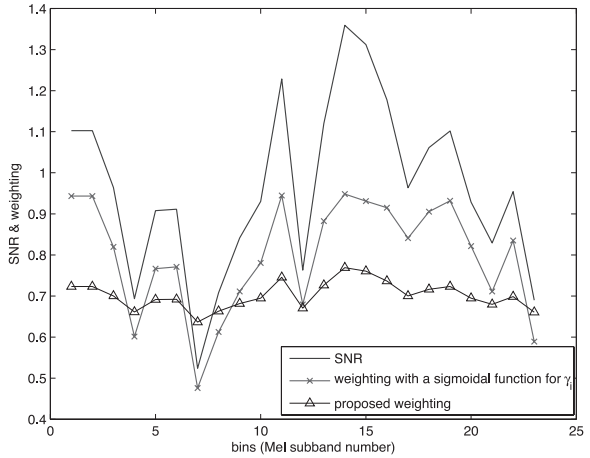


Fig. 3 Comparison of the variations of the proposed weighting function with SNR and the weighting function used in [16], [18], for all 23 Mel subbands in a file with SNR = 0 chosen from Aurora 2 task.

$$\begin{aligned} &\text{if} && \mathrm{E}[y_l(i)] \le \beta \mathrm{E}[N_l(i-1)] \\ &\text{then} && \mathrm{E}[N_l(i)] = \alpha \mathrm{E}[N_l(i-1)] + (1-\alpha)\mathrm{E}[y_l(i)] \\ &\text{else} && \mathrm{E}[N_l(i)] = \mathrm{E}[N_l(i-1)] \end{aligned} \tag{9}$$

where $\mathrm{E}[y_l(i)]$ and $\mathrm{E}[N_l(i)]$ are the estimated energies of the noisy signal and the noise of the l^{th} sub-band in frame i, respectively. In addition, α has been set to 0.99 and β to 2. Furthermore, $\mathrm{SNR}_l(i)$, which is the signal to noise ratio of the l^{th} sub-band in frame i is calculated as:

$$\mathrm{SNR}_l(i) = \frac{\mathrm{E}[y_l(i)]}{\mathrm{E}[N_l(i)]} \tag{10}$$

For computational purposes, the L^{th} order MVDR spectrum is computed using LP coefficients a_k and prediction error variance P_e [7], [15]:

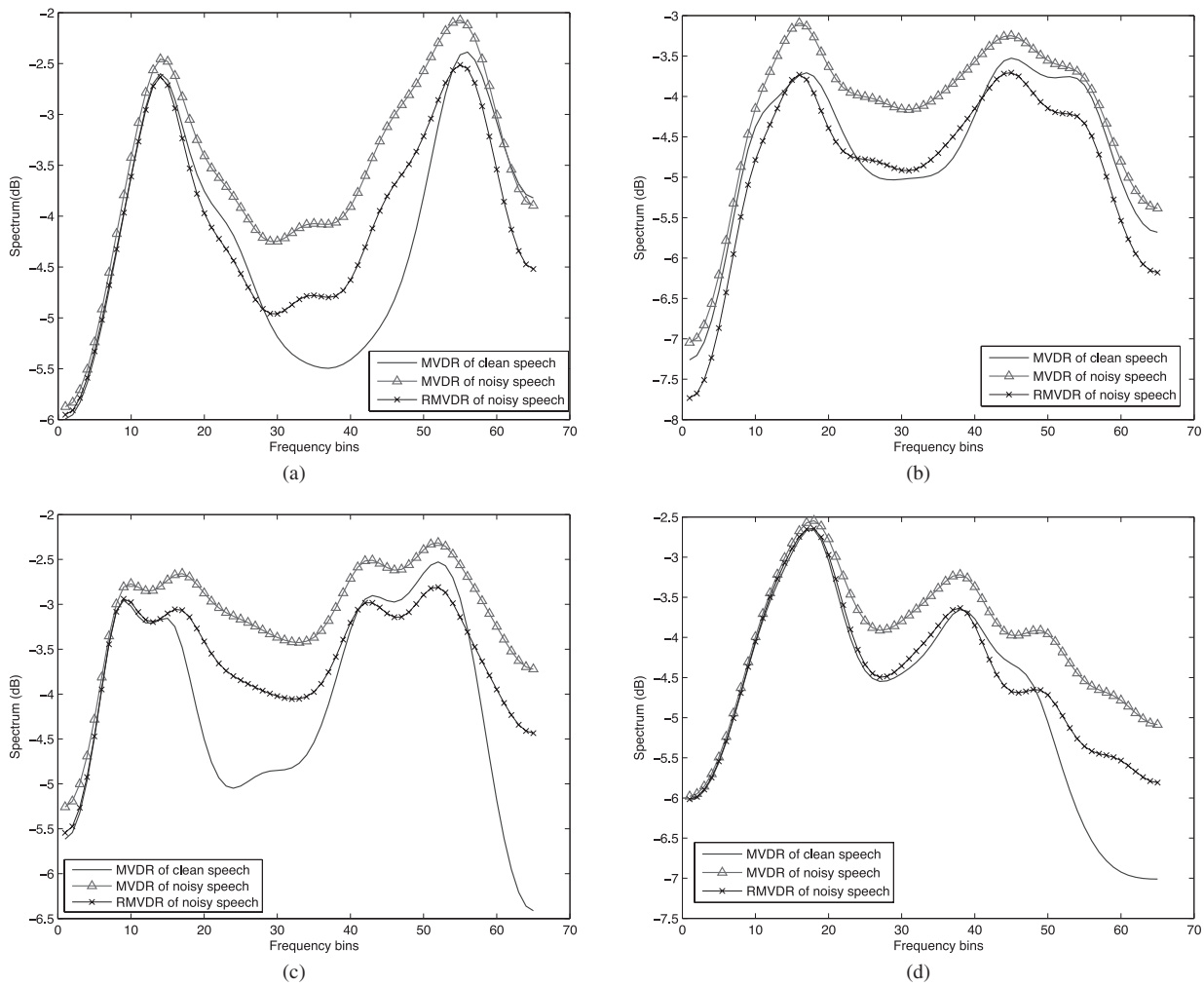


Fig. 4 Comparison of warped MVDR and RMVDR spectral estimates for 25 ms long voiced speech frames chosen from four different noisy speech files of Test set A in Aurora 2 task. These utterances include connected digits pronounced by females and contaminated by four different noises (a) /ey/ sound in subway noise, (b) /v/ sound in babble noise, (c) /n/ sound in car noise and (d) /o/ sound in exhibition noise.

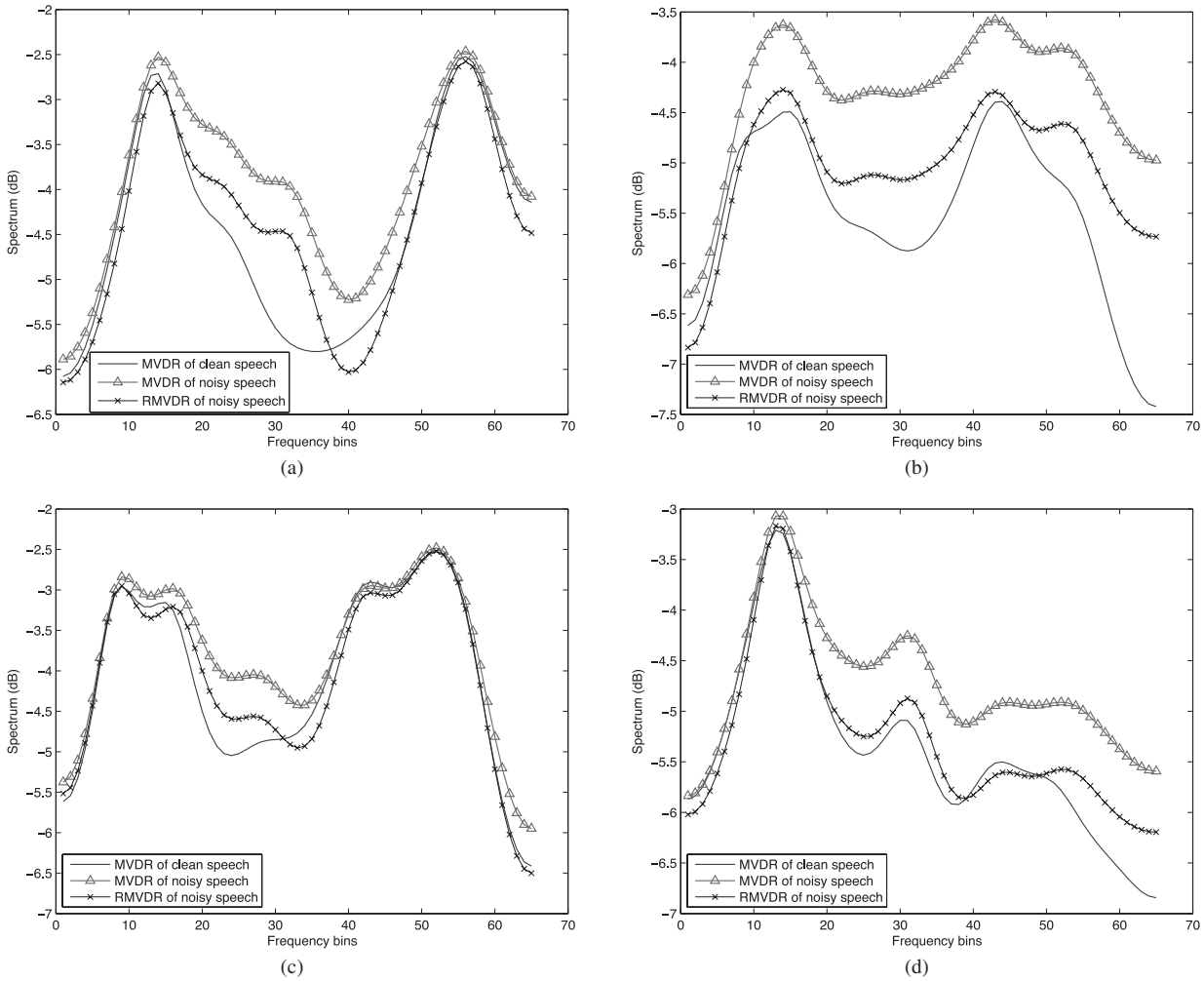


Fig. 5 Comparison of warped MVDR and RMVDR spectral estimates for 25 ms long voiced speech frames chosen from four different noisy speech files of Test set B in Aurora 2 task. These utterances include connected digits pronounced by females and contaminated by four different noises (a) /ey/ sound in restaurant noise, (b) /v/ sound in street noise, (c) /n/ sound in airport noise and (d) /o/ sound in train station noise.

$$P_{MVDR}(\omega) = \frac{1}{\sum_{-L}^{L} \mu(k) e^{-j\omega k}} \tag{11}$$

$$\mu(k) = \begin{cases} \frac{1}{P_e} \sum_{i=0}^{L-k} (L+1-k-2i) \\ \quad \times a_i a_{i+k}^* & k = 0, \cdots, L \\ \mu^*(-k) & k = -L, \cdots, -1 \end{cases} \tag{12}$$

where $(2L+1)$ coefficients of $\mu(k)$ are called the MVDR coefficients, and the MVDR spectrum can easily be calculated by an FFT computation according to (11).

In order to extract Robust Perceptual MVDR-based Cepstral Coefficients (RPMCC), LP coefficients are extracted from the weighted perceptually warped power spectrum as discussed before. Figure 1 (b) shows the flow diagram of our proposed robust front-end in detail. Equal loudness curve and power law of hearing blocks have been used in calculating RPMCC features according to the PLP structure [3]. In addition, according to (6), the weighting should be applied before calculating the MVDR spectrum. Therefore, as Fig. 1 (b) shows, the SNR estimation is performed using the FFT spectrum, and not the MVDR power spectrum.

In order to perform a comparison between the methods in improving the power spectrum of noisy speech signals, we have shown the perceptually warped MVDR log-spectral estimates, called MVDR spectrum, and robust perceptually warped MVDR log-spectral estimates, called RMVDR spectrum, of a frame of some speech files of Aurora 2 task, pronounced by female speakers and contaminated with different noises. Figure 4 compares the MVDR and RMVDR spectral estimates of 25 ms frames of four speech files chosen from Test set A corrupted by subway, babble, car, and exhibition noises, at SNR 10 dB, respectively. In addition, the MVDR and RMVDR spectral estimates of four different files of Test set B corrupted by restaurant, street, airport, and train station noises, at SNR 10 dB, have been represented in Fig. 5 (a)–(d) respectively. According to these figures, the RMVDR spectrum of noisy speech can follow the clean

spectrum better than the MVDR spectrum. It is worth mentioning that the RMVDR method is more successful in finding the fundamental peaks or formants of the power spectrum as figures show. Consequently, the RMVDR spectral estimation method is able to reduce the noise and estimate a more robust power spectrum in comparison with MVDR spectrum, which itself, was shown to be more robust than MFCC in [7].

4. Experimental Results

Recognition experiments were conducted on Aurora 2 task [20] with clean training scenario. The Aurora 2 corpus is known as one of the most popular tasks in evaluating the robust speech recognition methods in speaker-independent systems. It is derived from the TIDigits database, consisting of connected digits spoken by American English talkers, and is downsampled to 8 kHz. It includes two training modes: clean-condition training (training on clean-data only) and multi-condition training (training on clean and noisy data). 8440 utterances, containing the recordings of 55 male and 55 female adults, have been selected from the training part of the TIDigits for the former mode. All of these signals have been filtered with G.712 characteristic. For the latter mode, the same 8440 utterances are equally split into 20 subsets with 422 utterances in each one. Four different noises, namely suburban train, babble, car and exhibition hall, have been added to the so-called subsets at SNRs of 20 dB, 15 dB, 10 dB, and 5 dB. The filtering process is exactly like the clean training mode.

The test data of Aurora 2 task have been divided into three sets, namely, Test set A, B and C. 4004 utterances from TIDigits test set data are split into four subsets with 1001 utterances in each. Besides the clean speech signals, one noise type is added to each subset at SNRs of 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, and −5 dB, in order to constitute the test set. The noises of test set A are suburban train, babble, car and exhibition hall. In addition, restaurant, street, airport, and train station are used to make test set B. Test set C consists of 2 of the 4 subsets. Consequently, while each of the test sets of A and B contains 28028 utterances, test set C is made up of 14014 utterances. Suburban train and street are used as the additive noises in test set C. In this set, speech and noises are filtered with an MIRS characteristic before adding the so called noises. Test set C is used to evaluate the performance of ASR systems in presence of a convolutional noise.

In this paper, we have only used the clean condition training. It is advantageous to multi-condition scenario, since the speech is modeled without distortion by any type of noise. We used hidden Markov models (HMM) to model the digits and pauses using the same topology in [20]. The robustness of the obtained features was evaluated on Aurora 2 task using HTK software [21].

The baseline uses the well-known MFCC features. For extracting all the features, speech was segmented into 25 ms frames with a frame-shift of 10 ms. The Mel filter-bank con-

Table 1 Average recognition accuracies over different noise types for various SNR values and three values of γ_i.

SNR	γ		
	0.25	**0.5**	**1**
clean	98.42	98.56	98.11
20	95.00	94.28	93.52
15	91.40	90.91	88.97
10	78.73	80.32	79.76
5	48.91	53.37	55.61
0	24.82	25.07	24.19
−5	13.37	12.69	12.20

Table 2 Average recognition accuracies over different noise types and SNRs for test sets A, B and C and different features.

Feature	Set A	Set B	Set C
MFCC	63.90	66.15	58.21
WMFCC	66.59	68.89	60.93
PMCC	66.25	68.73	60.87
RPMCC	72.36	72.99	67.17

Table 3 Recognition accuracies for different features in various noise types of test set A of Aurora 2 task.

Noise type	Feature	SNR							
		clean	**20**	**15**	**10**	**5**	**0**	**−5**	**Average**
Subway	**MFCC**	97.88	95.70	90.88	74.33	41.91	23.95	16.15	65.35
	WMFCC	97.97	95.52	91.25	77.10	49.59	25.76	16.86	67.84
	PMCC	97.97	96.38	92.17	77.34	45.66	24.99	17.65	67.31
	RPMCC	98.19	96.90	94.11	86.21	65.70	31.84	18.30	74.95
Babble	**MFCC**	97.97	93.86	89.84	77.27	51.39	24.06	11.15	67.28
	WMFCC	97.91	94.01	89.57	78.14	56.95	26.33	11.19	69.00
	PMCC	97.94	94.89	91.44	81.47	56.35	26.03	11.88	70.04
	RPMCC	98.25	94.92	91.48	82.56	62.00	28.84	12.70	71.96
Car	**MFCC**	97.88	96.12	90.25	67.49	32.36	20.61	11.90	61.37
	WMFCC	97.73	96.45	91.59	74.59	39.55	21.12	11.72	64.66
	PMCC	97.94	96.63	92.51	73.07	35.70	21.77	12.97	63.94
	RPMCC	98.15	97.17	95.17	86.28	57.95	25.80	13.54	72.47
Exhibition	**MFCC**	97.78	94.85	89.02	68.65	35.36	20.18	12.13	61.61
	WMFCC	97.99	96.24	90.37	73.84	41.13	22.80	12.87	64.88
	PMCC	98.12	95.06	90.50	72.29	39.68	21.04	11.60	63.71
	RPMCC	97.96	94.79	92.47	82.51	54.18	26.35	13.67	70.06

Table 4 Recognition accuracies for different features in various noise types of test set B of Aurora 2 task.

Noise type	Feature	SNR							
		clean	20	15	10	5	0	−5	Average
Restaurant	MFCC	97.88	92.51	88.39	77.89	53.27	24.99	11.33	67.41
	WMFCC	97.97	92.42	88.64	78.57	57.84	30.27	12.68	69.55
	PMCC	97.97	93.64	90.42	80.96	56.77	26.87	11.85	69.73
	RPMCC	98.19	92.35	89.87	81.30	61.81	31.44	13.11	71.35
Street	MFCC	97.97	95.86	89.75	69.98	41.44	23.64	13.66	64.13
	WMFCC	97.91	95.86	89.39	73.04	46.70	25.12	13.69	66.02
	PMCC	97.94	96.31	91.54	74.24	44.50	25.39	15.08	66.40
	RPMCC	98.25	96.67	93.20	83.31	57.92	29.72	15.54	72.16
Airport	MFCC	97.88	93.26	89.32	78.59	52.61	27.68	12.62	68.29
	WMFCC	97.73	94.04	90.01	81.06	59.59	31.85	13.57	71.31
	PMCC	97.94	94.21	91.95	82.02	56.93	30.72	14.49	71.17
	RPMCC	98.15	94.78	92.57	85.30	65.64	34.75	15.75	74.61
Train Station	MFCC	97.78	93.80	89.73	74.88	43.23	22.28	12.71	64.78
	WMFCC	97.99	95.22	91.58	79.85	51.77	24.96	12.16	68.68
	PMCC	98.12	95.09	91.82	79.98	47.24	23.94	13.51	67.61
	RPMCC	97.96	95.56	93.77	86.33	63.19	30.33	14.25	73.84

Table 5 Recognition accuracies for different features in various noise types of test set C of Aurora 2 task.

Noise type	Feature	SNR							
		clean	20	15	10	5	0	−5	Average
Subway (MIRS)	MFCC	97.94	91.50	82.68	62.67	33.50	14.43	8.32	56.96
	WMFCC	97.61	94.41	87.14	68.50	37.15	18.39	9.76	61.12
	PMCC	97.94	93.61	85.97	66.75	37.43	17.96	9.09	60.34
	RPMCC	97.97	93.92	89.84	78.14	52.29	24.04	10.25	67.65
Street (MIRS)	MFCC	97.76	94.50	86.91	61.85	34.79	19.26	12.70	59.46
	WMFCC	97.52	94.17	87.00	64.09	37.42	21.04	13.33	60.74
	PMCC	97.85	95.22	88.27	65.24	37.70	20.56	13.36	61.40
	RPMCC	97.76	95.47	90.93	76.30	46.70	24.03	13.51	66.69

sists of 23 triangular filters. To optimize the model order for MVDR-based coefficients, different orders were tested and the optimum model order of 15 which gives the best average recognition accuracy was used. Finally, each frame was represented by a vector consisting of 12 cepstral features augmented by their first and second order derivatives.

A set of preliminary experiments were carried out to find the optimum γ_i as the steepness controller of the weighting function. For this reason, we created a compact corpus consisting of 2110 Aurora 2 training files plus 5600 test files extracted from its test Sets A and B. Different values of γ_i with various types of noises and SNRs were evaluated. Table 1 shows the average recognition accuracies obtained in all kinds of noises for this compact Aurora 2 task. According to this table, best recognition accuracies for very high and low SNRs were obtained by choosing smaller values of γ in comparison with the medium SNRs. Although this is not always the case, we decided to choose a function for γ_i which is in accordance with such characteristic. As mentioned in Sect. 3, our experiments showed that choosing a function which is made up of a difference of two sigmoidal functions gives better recognition results. The parameters of this function were also tuned using the results of these preliminary experiments.

Table 2 gives the average recognition accuracies of MFCC, PMCC and RPMCC features over different noise types and SNRs for Test sets A, B and C. Tables 3, 4 and 5 show word recognition accuracies for the proposed and baseline features in different types of noises for test sets A, B and C, respectively. In addition, the recognition results with Weighted MFCC features (WMFCC), which are extracted by weighting the Mel sub-bands according to (7), have been included in tables for comparison. Furthermore, the average recognition accuracies over different noise types for each of the test sets A, B and C have been shown in Fig. 6 (a)–(c) for better comparison.

5. Discussion

The tables clearly show the robustness of the proposed features on three different test sets of Aurora 2 task in different noisy conditions. It is worth mentioning that even in clean conditions, the proposed features demonstrate better performance in comparison with MFCC, WMFCC and PMCC in most of the noisy conditions. This happens because according to (7), the weights assigned to the sub-bands will almost be equal to one in clean conditions, and zero in very noisy environments. Therefore, our suggested method not only does not deteriorate the characteristics of speech features in clean conditions, but also improves them due to incorporating the findings of psychoacoustics. As tables show, the PMCC features are more robust than MFCC, due to the

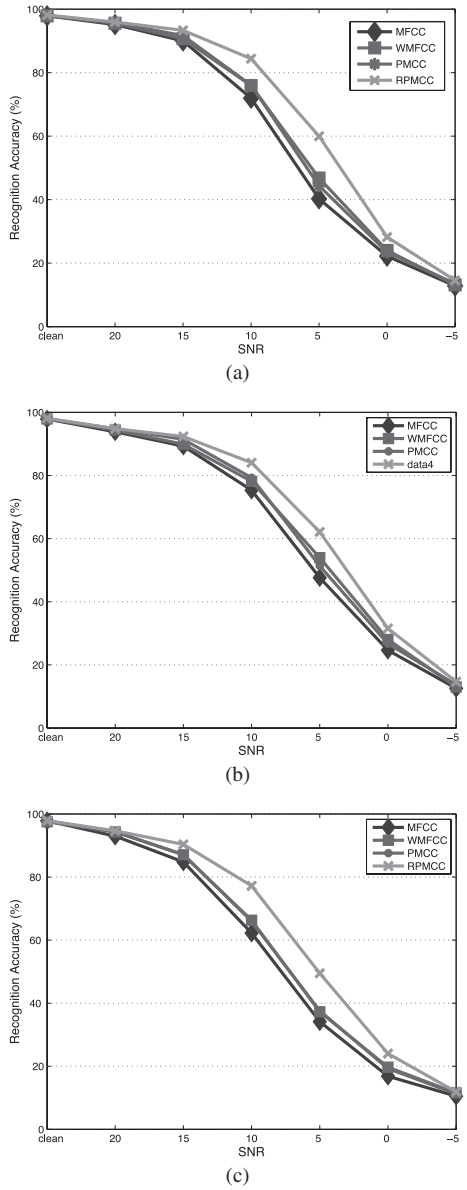


Fig. 6 Average recognition accuracies over various noise types for different features of Aurora 2 task. (a) Test set A. (b) Test set B. (c) Test set C.

smaller bias and variance in estimating the MVDR power spectrum. The low bias is helpful in detecting low level peaks in the presence of a higher one; and therefore, the formants of the signal in low signal to noise ratios can better be preserved. It is also useful for extracting the power spectrum more accurately in clean conditions. Moreover, decreasing the variance and hence smoothing the undesired fine structure in estimating the spectrum in MVDR-based features makes them more robust against different additive noises. It is worth stating that this variance reduction has only been achieved by using MVDR spectral estimation method instead of FFT-based approaches, without using any temporal smoothing.

Furthermore, the recognition results of RPMCC features show that modifying the distortionless constraint of the MVDR spectral estimation method and defining a new robust distortionless constraint based on the sub-band signal to noise ratios is absolutely advantageous in extracting more robust features in comparison with MFCC, WMFCC and PMCC coefficients. It is also worth mentioning that choosing the weighting function in accordance with the findings of psychoacoustics plays an important role in making our robust algorithm successful, especially in low signal to noise ratios. The improvement of recognition accuracies of WMFCC features in comparison with MFCC also proves this claim.

Moreover, comparing the execution time of extracting the proposed and baseline features shows that our suggested method leads to a modest increase in the computational load. To show this, we selected 1000 files out of test set A of Aurora 2 task, and extracted the features on a Pentium 4 computer with a 3.2 GHz dual core processor and 2 GB RAM, running MS Windows XP. The PMCC and RPMCC features of this set were obtained in 50 seconds and 62 seconds, respectively, implemented on MATLAB. Therefore, while our approach improves recognition accuracy in noisy environments, it does not lead to substantial increase in the complexity of the front-end algorithm.

It is worth mentioning that in this paper we aimed to propose a new robust front-end for speech recognition based on improving the power spectrum of speech signals. Thus we chose MFCC and PMCC features as the baselines for our experimental comparisons. There also exist more advanced features that perform well in noisy conditions, such as the ETSI standard front-end [22], that involves rather high computational load due to its complex algorithm. Nevertheless, we showed in this paper that our proposed front-end outperforms the MFCC and PMCC features with a comparable computational load. While this approach might also be able to improve the performance of more complicated front-ends, such as the ETSI front-end, such applications are out of the scope of this paper.

6. Conclusions

In this paper, a new front-end for robust speech recognition was proposed. This front-end is based on making the MVDR power spectrum estimation method robust against noise. This robustness is achieved by modifying the distortionless constraint of the MVDR spectral estimation method such that the gain of the filter output at each Mel sub-band is determined by a function dependent on the sub-band signal to noise ratio. Consequently, the signal components which have been affected by noise more than the others are given smaller weights. These weights are carefully chosen by incorporating the findings of psychoacoustics. Extracting more robust power spectrum without excessively increasing the computational load is one of the benefits achieved by implementing this algorithm. In addition, obtaining better recognition accuracy in comparison with MFCC, WMFCC and PMCC features in most cases, even in clean environ-

ments, is another valuable advantage of the proposed robust feature extraction approach.

Acknowledgements

This project has been funded by Iran Telecommunication Research Center (ITRC).

References

[1] M. Su, P. Li, Z. Wang, P. Ding, and B. Xu, "A novel noise robust front-end using first order VTS in construction of Mel-warped wiener filter," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.I-777–779, 2006.

[2] L. Rabiner, B. Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993.

[3] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," J. Acoust. Soc. Am., vol.87, pp.1738–1752, 1990.

[4] J. Chen, K.K. Paliwal, and S. Nakamura, "Cepstrum derived from differential power spectrum for robust speech recognition," Speech Commun., vol.41, no.2-3, pp.469–484, 2003.

[5] J. Shannon and K.K. Paliwal, "Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition," Speech Commun., vol.48, pp.1458–1485, 2006.

[6] S. Seyedin and M. Ahadi, "Feature extraction based on DCT and MVDR spectral estimation for robust speech recognition," Proc. 9th IEEE International Conference on Signal Processing, pp.605–608, Oct. 2008.

[7] S. Dharanipragada, U.H. Yapanel, and B.D. Rao, "Robust feature extraction for continuous speech recognition using the MVDR spectrum estimation method," IEEE Trans. Audio, Speech, Language Process., vol.15, no.1, pp.224–234, Jan. 2007.

[8] S.H. Lin, Y.M. Yeh, and B. Chen, "Exploiting polynomial-fit histogram equalization and temporal average for robust speech recognition," ICSLP 2006, pp.2522–2525, Sept. 2006.

[9] R. Togneri, A.M. Toh, and S. Nordholm, "Evaluation and modification of cepstral moment normalization for speech recognition in additive babble ensemble," 11th Australasian Int. Conf. on Speech Science and Technology (SST), 2006.

[10] J. Beh and H. Ko, "A novel spectral subtraction scheme for robust speech recognition: Spectral subtraction using spectral harmonics of speech," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.648–651, 2003.

[11] C.H. Lee, F.K. Soong, and K.K. Paliwal, Automatic speech and speaker recognition, Kluwer Academic Publishers, Norwell, 1996.

[12] M.J.F. Gales and S.J. Young, "Robust continuous speech recognition using parallel model combination," IEEE Trans. Speech Audio Process., vol.4, no.5, pp.352–359, 1996.

[13] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.869–872, 2000.

[14] J. Deng, M. Bouchard, and T.H. Yeap, "Speech feature estimation under the presence of noise with a swithing linear dynamic model," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.497–500, 2006.

[15] S.L. Marple, Digital Spectral Analysis with applications, Prentice Hall, 1987.

[16] K.K. Chu and S.H. Leung, "SNR-dependent non-uniform spectral compression for noisy speech recognition," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp.I-973–976, 2004.

[17] E. Zwicker and H. Fastl, Psychoacoustics, facts and models, 3rd ed., Chapter 8, Springer-Verlag, 2007.

[18] H. Yeganeh, S.M. Ahadi, S.M. Mirrezaie, and A. Ziaei, "Weighting of mel sub-bands based on SNR/Entropy for robust ASR," Proc. 8th IEEE Int. Symposium on Signal Processing and Information Technology, pp.292–296, Dec. 2008.

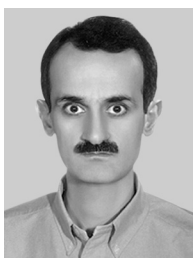
[19] Z. Junhui, K. Jingming, X. Xiang, and H. Shilei, "Noise suppression based on teager energy operator for improving the robustness of ASR front-end," Proc. International Workshop on Acoustic Echo and Noise Control, pp.135–138, Sept. 2003.

[20] H.-G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," ASR-2000, pp.181–188, Sept. 2000.

[21] HTK, 2002. The hidden Markov model toolkit. Available from: http://htk.eng.cam.ac.uk

[22] "ETSI: Speech processing: Transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithm," ETSI ES 202 050 v1.1.3, ETSI standard, 2003.

Sanaz Seyedin received the B.Sc. degree from Amirkabir University of Technology, Tehran, Iran, and the M.Sc. from Iran University of Science and Technology, Tehran, Iran, both in electronics engineering, in 2001, and 2005, respectively. She is currently a Ph.D. student at Electrical Engineering Department, Amirkabir University of Technology. Her research interests are in the areas of robust speech recognition, speech processing, and statistical signal processing.



Seyed Mohammad Ahadi received the B.Sc. and M.Sc. degrees in electronics from the Electrical Engineering Department, Amirkabir University of Technology, Tehran, Iran, in 1984 and 1987, respectively. He was appointed faculty member at the same department in 1988, where he began his teaching profession as well as involvement in research projects. From 1992 to 1996, he pursued his studies toward the Ph.D. degree, working in the field of speech recognition. He received his Ph.D. degree in engineering from the University of Cambridge, Cambridge, UK, in 1996. Since then, he has been again with the Electrical Engineering Department, Amirkabir University of Technology, where he is currently an associate professor, teaching several courses and conducting research in electronics and communications. He is a senior member of IEEE and the secretary of IEEE Iran section.