Robust Speech Recognition Based on Mixed Histogram Transform and Asymmetric Noise Suppression

Hassan Farsi[1], Samane Koohi moghadam [2]
1- Department of Electronics and Communications Eng., University of Birjand, Birjand, Iran.
Email: hfarsi@birjand.ac.ir
2- Department of Engineering, University of payam noor, Mashhaad, Iran.
Email: kuhimoghaddam@yahoo.com

Received: August 2012 Revised: October 2012 Accepted: December 2012

ABSTRACT:
This paper proposes a new feature extraction algorithm which is robust against noise using histogram compensation and asymmetric filter. Temporal masking is used to improve Automatic Speech Recognition (ASR) systems specifically in matched and multi-style training conditions. Nonlinear filtering and temporal masking are used in the proposed algorithm. By matching the power histograms of the input in each frequency band to those obtained over clean training data, and then mixing the processed and unprocessed spectrums together, speech recognition accuracy can be appropriately increased. The obtained results show that recognition accuracy in comparison with Mel Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction (PLP) and Power Normalized Cepstral Coefficients (PNCC), improves in various training conditions and different SNRs.

KEYWORDS: Robust speech recognition, Temporal masking, Asymmetric nonlinear filter, Averaging weighted spectral.

1. INTRODUCTION

Accuracy of Automatic Speech Recognition (ASR) system decreases when it is used in noisy environment. The reason is the difference between training data and test data. Many robust ASR algorithms have been presented so far. Nevertheless, obtaining good performance in noisy environments still remains a challenging task. The problem is that recognition accuracy degrades significantly if training conditions are not matched to the corresponding test conditions.

The state-of-the-art ASR systems show excellent performance in a controlled environment. These systems have designed for a certain noise, but yet there is not algorithm with acceptable accuracy in different noise environments. Cepstral Mean Normalized (CMN) [1] and Mean and Variance Normalization (MVN) [2] are the simplest forms of these techniques[3], in which it is assumed the mean or the mean and variance of the cepstral vectors should be equal for all utterances. Histogram Equalization (HEQ) [4] is another strong method that assumes all cepstral vectors have the same probability density function. In [5], Kim described PNCC, which is more robust against noise and reverberation than MFCC and PLP features.

Nevertheless, most of these algorithms are based on this issue that the ASR system is trained by clean speech, only to be exposed to noisy data in the testing stage. In recent years, it has been seen that it is useful to train very large systems using noisy data. Such multi-style training is common and requires robustness algorithms with good performance (when noisy data is used for training). This paper is based on histogram transform that is performed on the power of the speech signal in Equivalent Rectangular Bandwidth (ERB)-warped sub-bands. This transformation is followed by averaging a weighted spectral on the processed and unprocessed power spectrums.

For the first time, use of histogram matching was described to compensate the effects of nonlinear channel distortion in speaker identification systems [6]. The first direct application of histogram matching for ASR was recommended in [7] that could be considered as a computationally complicated form of unsupervised speaker adaptation. This method is comparable in performance to Maximum Likelihood Linear Regression (MLLR). In this method using Mel Frequency Cepstral Coefficients (MFCC), histogram matching is used in the feature level. During the recent years, researchers have explored the application of histogram-based methods on robust speech recognition [8]. In [9] it is focused on a parametric implementation, by quintile-based histogram

equalization. This equalization is considered for capstral features and compensates the effect of additive noise.

The proposed method modifies and extends these current approaches by changing the procedure of histogram computation, normalization and through the subsequent averaging weighted spectral. Using nonlinear warping function based on power in each frequency band, noise effects are minimized because related energy to noise signal is matched by clean speech with more accuracy.

In this paper, we also utilize asymmetric noise suppression to minimize noise effects. Since speech power changes more rapidly than background noise in each frequency channel, we can expose this kind of noise compensation for discussion. On the other hand, since speech has higher modulation frequency spectrum than noise, many algorithms have been raised by band-pass filtering or high-pass filtering in modulation spectrum domain [10], [11]. The simplest way is high-pass filtering in each channel that removes low frequency components [12], [13]. A significant issue in the application of conventional linear high-pass filtering in power domain is that the output power can become negative which is mathematically impossible. Also it results in some problems in speech synthesis unless we use appropriate floor value for power coefficients [13]. Therefore, filtering could be performed after applying log nonlinearity (as MFCC method) but this is not suitable for environment containing additive noise. Spectral subtraction is another way for decreasing the effects of noise whose power changes slowly [14]. The noise level is estimated in spectral subtraction techniques from the power of speech parts [14] or through using a continuous- update approach [12]. We introduce a method that results in time variable estimation of noise floor by using asymmetric filter, and then it is subtracted from instantaneous power.

This paper is organized as follows: in section 2 the overall structure of the proposed method is presented. Next, in section 3, general characteristics of asymmetric nonlinear filter are explained. In sections 4 and 5 temporal masking and weight smoothing applied in the proposed method are presented, respectively. In section 6, histogram based transformation is detailed. Then, in section 7, speech re-synthesis is introduced and it is followed by the experiments in section 8. Finally, conclusions are drawn in section 9.

2. OVERALL REVIEW OF THE PRPOSED STRUCTURE

Fig. 1 shows the flowchart of the proposed system. As shown, the first stage is similar to PLP and MFCC methods, except for using frequency analysis by gammatone filter. This is followed in the next stage by non-linear and time varying operations which are performed using longer duration temporal analysis with noise suppression. In third stage, histogram of input power for each frequency band is matched to those obtained over clean training data. Then, the processed and the unprocessed spectrums are combined together such that the recognition accuracy increases.

In speech processing the length of analysis window is normally between 20-30 ms. However, it has been shown that use of longer window length corresponds to better performance modeling and normalization of background noise [5]. Since power of background noise changes slowly than speech power, using "Medium-Time" process with 50-120 ms in length is more appropriate for analysis of background noise. The input signal is passed through a high-pass pre-emphasis filter and the short-time Fourier transform (STFT) is calculated using a window with length of 52 ms. Note that this window length is in interval 50-150 ms which is used for reducing variance of noise estimation. Experimentally, we found that this window length creates better performance. Then, squared spectrum is integrated using the squared gammatone frequency response. Using this procedure we can get channel by channel power P[m,l] where m and l are frame and channel indices, respectively. In equation form, it is represented as:

$$P[m,l]=\sum_{k=0}^{N-1}\left|X(m,e^{j\omega_k})G_l(e^{j\omega_k})\right|^2 \qquad (1)$$

Where N_a indicates size of FFT. We use 16 KHz sampling rate and $N_a=2048$. After weighting the frequency, the power has been normalized with peak power. $G_l[k]$ is gammatone filter bank for lth channel and $X[m,e^{jw_k}]$ shows short-time spectrum of speech signal for mth frame. The center frequencies are linearly spaced from 200 Hz to 8000 Hz in Equivalent Rectangular Bandwidth (ERB). It has been shown that using gammatone frequency weighting instead of traditional triangle weighting improves the performance of ASR system in noisy environment [5].

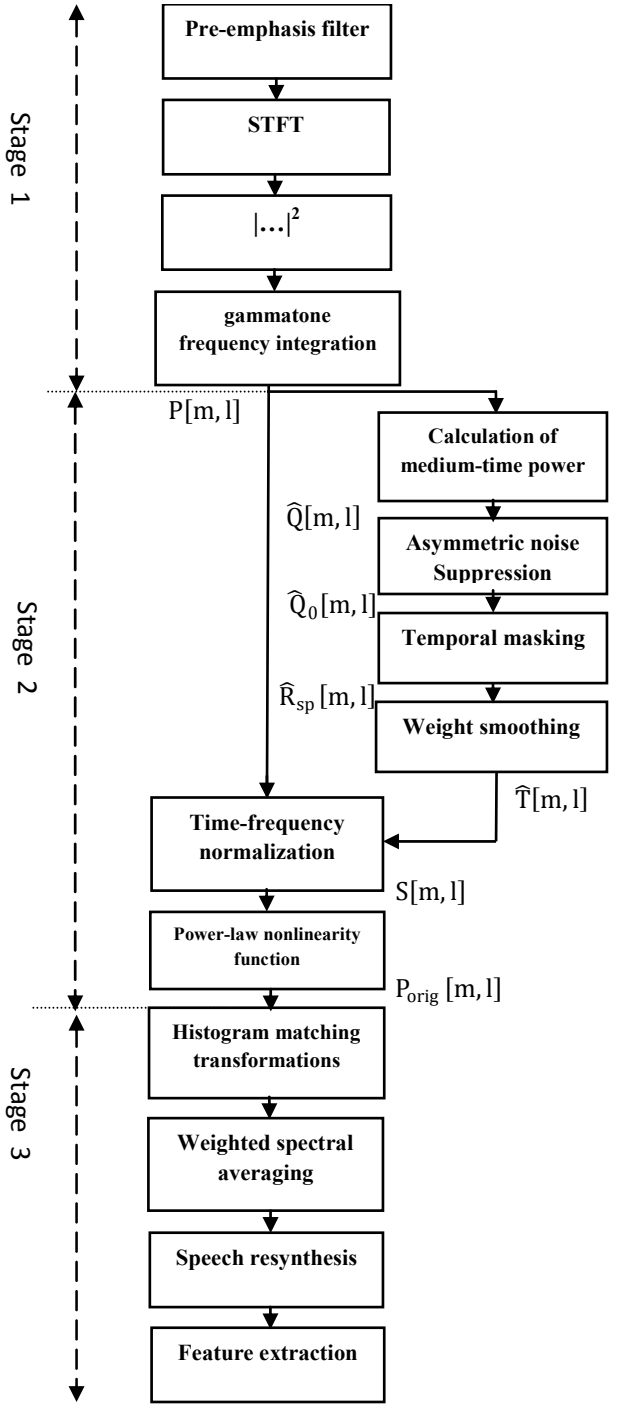


Fig. 1. Block diagram of the proposed structure

We estimate a quantity described as "medium-time power", $\hat{Q}[m,l]$ which is calculated by using the running average of P[m,l] , the power observed in a single analysis frame, according to the equation:

$$\hat{Q}[m,l] = \frac{1}{2M+1} \sum_{m'=m-M}^{m+M} p[m',l] \qquad (2)$$

Selection of factor M has significant effect on performance (especially in case of white noise). It is empirically found that if we choose the value of 2 for M then the recognition accuracy is optimum. However, if we use the features based on $\hat{Q}[m,l]$, then the recognition accuracy decreases. Because onset and offset of the frequency components are undetermined. Therefore, in the proposed feature extraction method, we use $\hat{Q}[m,l]$ only for estimation and compensation of the noise and then we apply both asymmetric nonlinear filter and temporal masking for compensation of environmental noise and therefore, we can improve the features.

3. GENERAL CHARACTERISTICS OF ASYMMETRIC NONLINEAR FILTER

In the proposed system, we use asymmetric non-linear filter for estimation of background noise level in each frequency band and for each time frame. This approach is able to remove slow-varying components regardless to considering many artifacts associated with over-correction techniques such as spectral subtraction [14]. Fig. 2 indicates Asymmetric Noise Suppression (ANS) process and temporal masking. First, we explain general characteristics of asymmetric nonlinear filter.

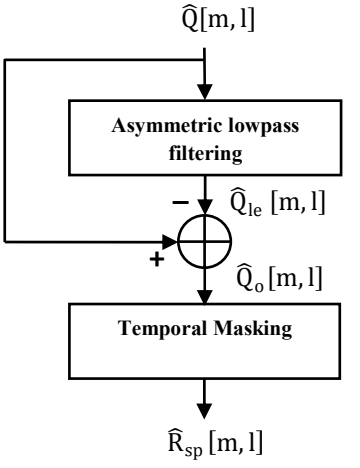


Fig. 2. Block modeled diagram for asymmetric noise removal

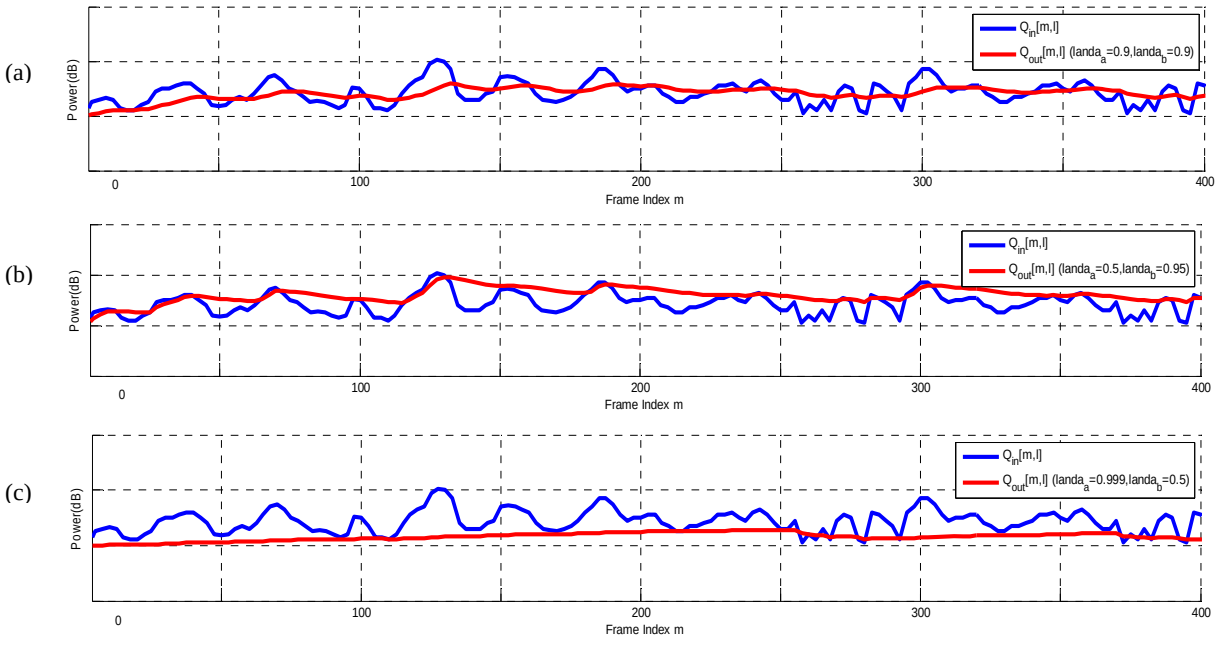


Fig. 3. Sample input (solid curves) and output (dashed line curves) of the filter defined in Eq. (1) for different conditions when:

(a) $\lambda_a = \lambda_b$, (b) $\lambda_a < \lambda_b$, (c) $\lambda_a > \lambda_b$.

This filter is described for the arbitrary input, $\hat{Q}_{in}[m,l]$ and output, $\hat{Q}_{out}[m,l]$, as:

$$\hat{Q}_{out}[m,l] = \begin{cases} \lambda_a \hat{Q}_{out}[m-1,l] + (1-\lambda_a)\hat{Q}_{in}[m,l] \\ \quad if \quad \hat{Q}_{in}[m,l] \geq \hat{Q}_{out}[m-1,l] \\ \lambda_b \hat{Q}_{out}[m-1,l] + (1-\lambda_b)\hat{Q}_{in}[m,l] \\ \quad if \quad \hat{Q}_{in}[m,l] < \hat{Q}_{out}[m-1,l] \end{cases} \quad (3)$$

where m and l are indices of frame and channel, respectively. λ_a and λ_b are constants with values between 0 and 1. If $\lambda_a = \lambda_b$, reviewing Eq. (3) will be easy, and since λ is positive, it will become a low-pass IIR filter as observed in Figure 3-a.

If $1 > \lambda_b > \lambda_a > 0$, then the nonlinear filter functions will become upper envelope detectors (Figure 3-b), and finally, as shown in Figure 3-c, if $1 > \lambda_a > \lambda_b > 0$, the filter output, $\hat{Q}_{out}$ will tend to follow the lower envelope of the input, $\hat{Q}_{in}[m,l]$. For better estimation of modeling the medium-time noise, lower envelope with changes is applied. Therefore, as this envelope reduces in the main input, $\hat{Q}_{in}[m,l]$, slow changes of non-speech components are deleted. We use Eq. (4) to represent the nonlinear filter described by Eq. (3).

$$\hat{Q}_{out}[m,1] = AF_{\lambda_a,\lambda_b}\left[\hat{Q}_{in}[m,1]\right] \quad (4)$$

This equality will be established only for index m in each channel l.

Regarding to asymmetric nonlinear filter features mentioned above, the lower envelope, $\hat{Q}_{le}[m,l]$, indicating noise average power is obtained by Asymmetric Noise Suppression (ANS) processing related to the following equation as observed in Figure 3-c:

$$\hat{Q}_{le}[m,l] = AF_{0.999,\,0.5}\left[\hat{Q}[m,l]\right] \quad (5)$$

Where $\hat{Q}[m,l]$ is the medium-time power obtained from Eq. (2) and A is a constant value. Then $\hat{Q}_{le}[m,l]$ is subtracted from $\hat{Q}_{in}[m,l]$. As shown in Fig. 4, we can observe that the obtained results for recognition accuracy by using asymmetric nonlinear filter, after implementing this structure for different values of λ_a

and λ_b. We add three kinds of noise, white noise, background music, and reverberation (with a delay of about 0.3 seconds). As observed in Fig. 4 the values of λ_b from 0.25 to 0.75 result in good recognition accuracy. According to this figure the best value for λ_a is 0.9. Therefore, in practice, we consider $\lambda_a = 0.999$ and $\lambda_b = 0.5$ because the recognition accuracy for speech is maximum in the presence of noise.

4. TEMPORAL MASKING

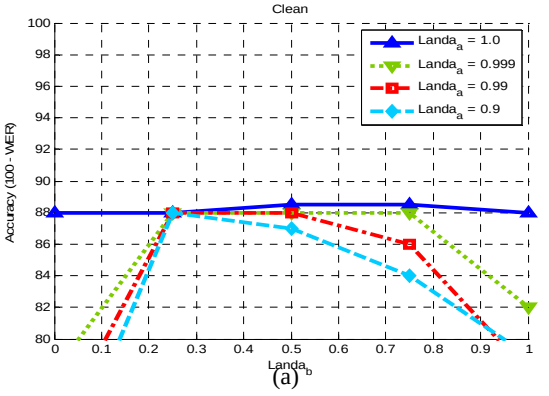
Many researchers have found the human auditory system focuses more on the onset of an incoming power envelope in comparison with falling edge of the same power envelope. In this regard, several algorithms have been proposed to improve the onset. In this section, we propose a simple procedure to incorporate this effect in processing the extracted feature vectors. It could be applied by using a moving peak for each frequency channel, l, and omitting instantaneous power if it is under this envelope. This process is shown in block diagram of Fig. 5.

In first stage, power of on-line peak, $\hat{Q}_p[m,l]$, is calculated for each channel by following equation.

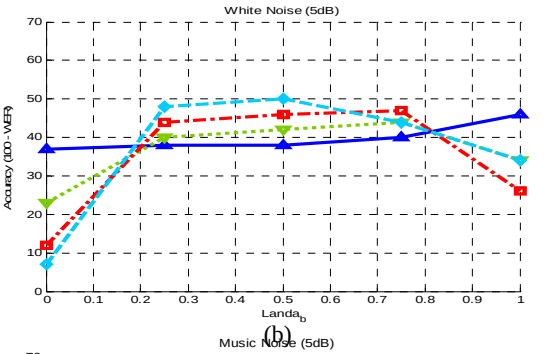
$$\hat{Q}_p[m,l] = Max(\lambda_t \hat{Q}_p[m-1,l], \hat{Q}_0[m,l]) \qquad (6)$$

Where λ_t is forgetting factor for calculation of on-line peak, m and l are frame and channel indices, respectively, and $\hat{Q}_0[m,l]$ is output power after ANS process. Temporal masking for speech parts is obtained through the following equation:

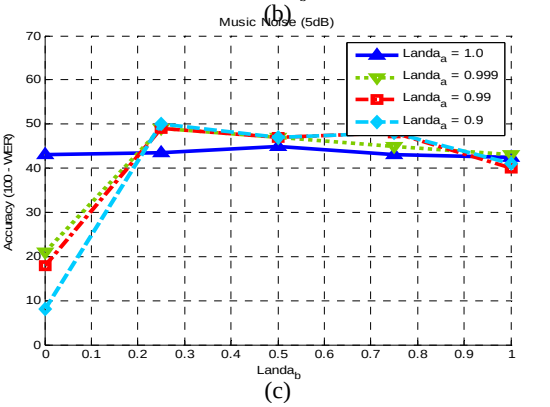
$$\hat{R}_{sp}[m,l] = \begin{cases} \hat{Q}_0[m,l] \\ if \quad \hat{Q}_0[m,l] \geq \lambda_t \hat{Q}_p[m-1,l] \\ \mu_t \hat{Q}_p[m-1,l] \\ if \quad \hat{Q}_0[m,l] < \lambda_t \hat{Q}_p[m-1,l] \end{cases} \qquad (7)$$



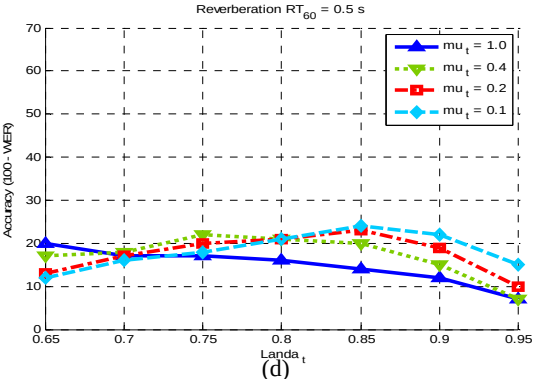
(a)



(b)



(c)



(d)

Fig. 4. Relationship among forgetting factors (λ_a , λ_b) and recognition accuracy for speech: (a) clean, (b) 5-dB Gaussian white noise, (c) 5-dB music noise and (d) Reverberation with RT_{60} = 0.5.

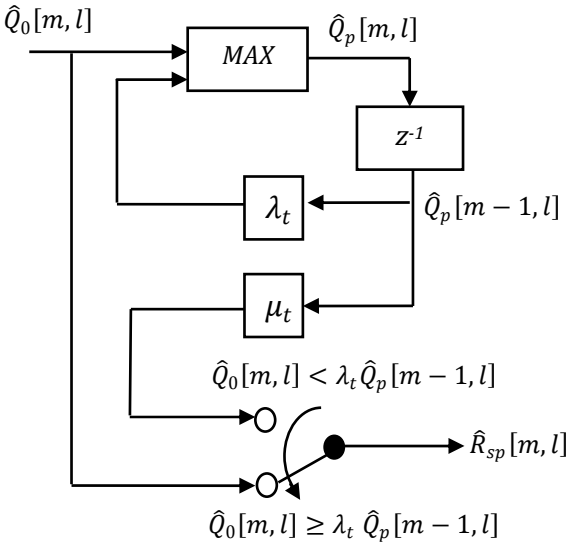


Fig. 5. Block diagram model for temporal masking

Fig. 6 indicates the relationship between recognition accuracy and forgetting factor (λ_t) and also coefficient of elimination (μ_t). We represent results of recognition system by using complete structure of Fig. 1 and just we change the coefficients of the forgetting and elimination factors (λ_t , μ_t). In a clean environment, as observed in Figure 6-a, if $\lambda_t \leq 0.85$ and $\mu_t \leq 0.2$, the recognition accuracy will almost remain constant. However, if $\lambda_t > 0.85$, performance will be degraded. In an additive noise environment such as weight or music noise as shown in Figures 6-b and 6-c, the performance is the same. However, for the reverberation, as shown in Figure 6-d, the application of temporal masking scheme provides considerable improvement.

5. WEIGHT SMOOTHING

It has been shown that weight smoothing plays an important role in speech enhancement and noise suppression [15], especially in non-linear processing. Therefore, in order to reduce side effects of non-linear operations, we use weight smoothing. According to our discussion so far, the effect of combination of asymmetric noise suppression and temporal masking for a frequency band and specific time frame is presented by:

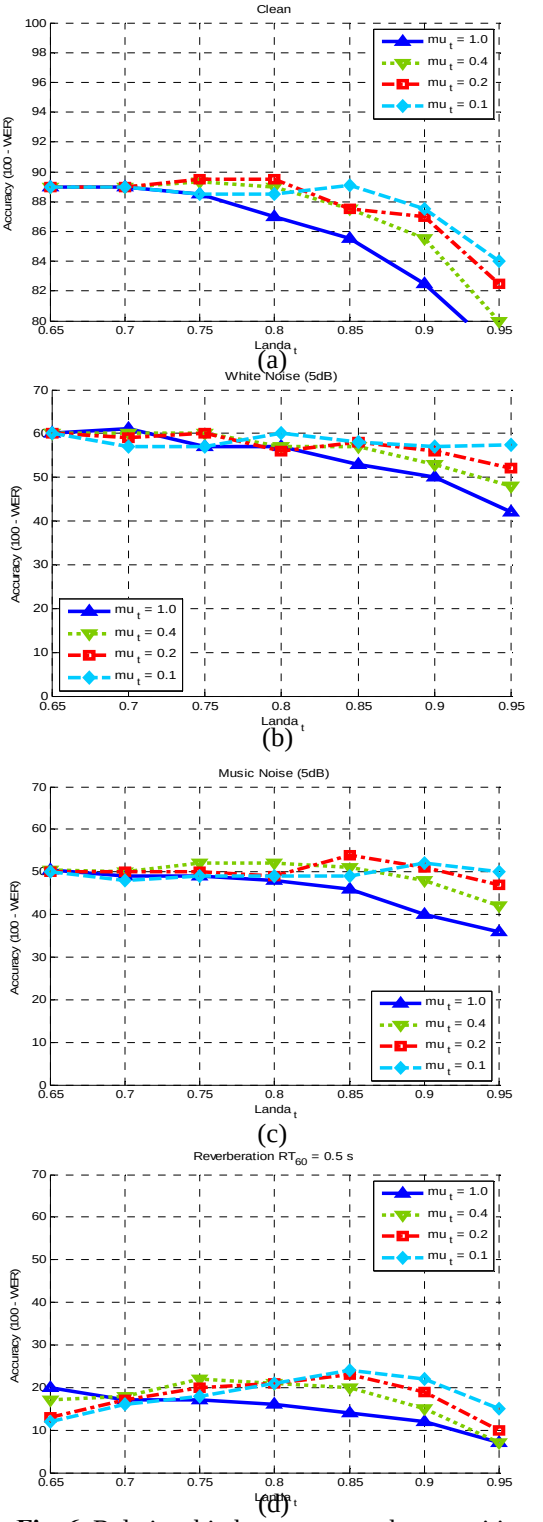


Fig. 6. Relationship between speech recognition accuracy and forgetting factor (λ_t) and elimination factor (μ_t): (a) clean, (b) 5-dB Gaussian white noise(c) 5-dB music noise and (d) Reverberation with RT_{60} = 0.5.

$$T[m,l]=\frac{\hat{R}_{sp}[m,l]}{\hat{Q}[m,l]} \quad (8)$$

Where $\hat{R}_{sp}[m,l]$ and $\hat{Q}[m,l]$ indicate output and input powers of asymmetric noise suppression filter with temporal masking, respectively. Frequency transfer function smoothing is obtained by calculation of average of the function $T[m,l]$ for lth channel. Therefore, the frequency averaged weighting function, $\hat{T}[m,l]$, is given by:

$$\hat{T}[m,l]=\frac{1}{l_2-l_1+1}\sum_{l=l_1}^{l_2}T[m,l] \quad (9)$$

Where $l_1=\max(l-N,L)$, $l_2=\min(l+N,L)$ and L is the number of channels. In the proposed method, N=4 and L=40 are considered. Note that if the number of channels differs then the value of N will be changed. Time-averaged with frequency-averaged transfer function, $S[m,l]$, for modulation of short-time power, $P[m,l]$, is given by:

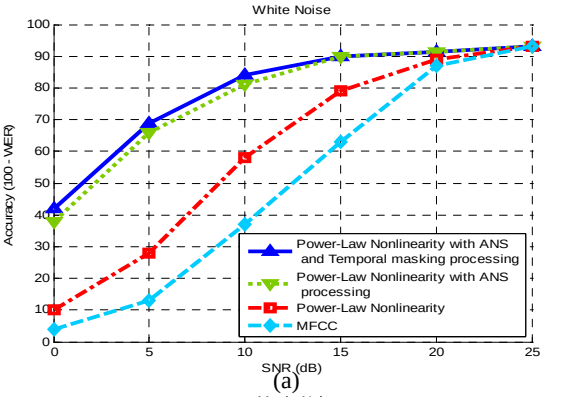
$$S[m,l]=P[m,l].\hat{T}[m,l] \quad (10)$$

As shown in Fig. 1, we then apply time-frequency normalization. Next, power-law nonlinearity function is used. This function is given by:

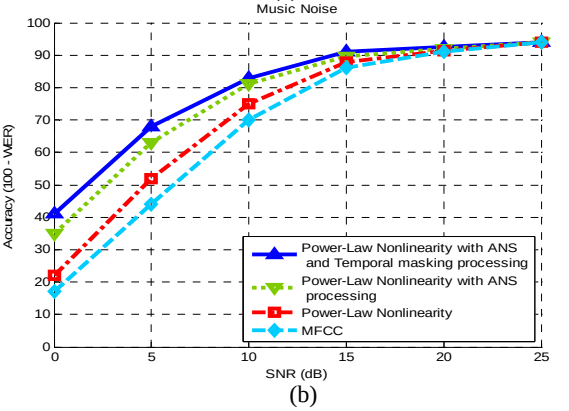
$$P_{orig}[m,l]=S[m,l]^{1/15} \quad (11)$$

Where $S[m,l]$ is out power after normalization. The effect of using ANS process and temporal masking has been shown in Fig. 7. As shown, the recognition accuracy increases by using power-law nonlinearity function instead of non-linear logarithm function applied in MFCC method. It has to be mentioned that the power-law nonlinearity function has been already used in PNCC method, but in combination with ANS process in the proposed system results in more accuracy in recognition especially for white noise and background music (see Fig. 7(a) and Fig. 7.(b)).

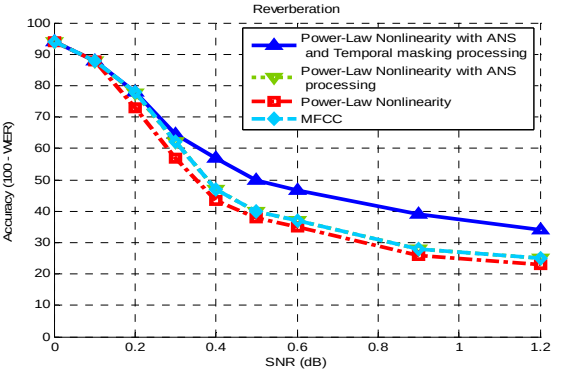
In addition it is observed that the recognition accuracy increases when the temporal masking is used. This is considerably observed in reverberated speech, especially for longer reverberations.



(a)



(b)



(c)

Fig. 7. The contribution of the power-law nonlinearity, asymmetric noise suppression, and temporal masking in the presence of (a) white noise, (b) music noise, (c) reverberation.

6. HISTOGRAM BASED TRANSFORMATIONS

In this step, we minimize the effect of noise, using a non-linear power-warping function in each frequency band. In other words, the robustness of the recognition system can be more improved by matching between the power of input histograms and those obtained over clean training data in each frequency band, and then combination of the processed and the unprocessed spectrums together [16]. Before calculation of the histograms over training data, the power signals are normalized by its local minimum and maximum values.

On the other hand, the calculated histograms over testing data are normalized by global minimum and maximum values of relative spectrum. This normalization method considerably reduces the effect of noise. In following section, we show that using the averaged-weighting between the processed and the unprocessed power spectrums can improve the performance of the recognition system histogram based. We can automatically achieve the proper nonlinear warping using non-parametric histogram matching. We use histogram matching described in [16].

6.1. Histogram-based power warping

Given a Cumulative Distribution Function (CDF) as input part, CX and CY are defined for random variables, X and Y, respectively. Therefore, X can be transformed to Y by:

$$f(x) = C_Y^{-1}\left(C_X(X)\right) \tag{12}$$

In proposed method, the sample CDFs are effectively scaled so that the corresponding histograms never have empty bins at the edges. This means that the histograms representing the input signal levels may be have empty bins at their edges, which we can decrease substantial noise.

40 histograms are achieved to display the output of each Gammatone filter bank channel for clean speech. As shown in Fig. 1, the histograms are calculated immediately after nonlinearity power function. Each histogram is obtained by dividing the range of the data in the channel into 100 bins with uniform band width. Real band widths and centers are changed from one channel to other channel. Therefore, there are never empty histogram bins at the edges.

For each utterance, the centers of 100 bins are equally scaled and used for calculation of input histograms. These centers are calculated by dividing the global range of the spectrum over all Gammatone channels. For a given frequency band, input histogram includes empty bins at the edges. Matching between test and sample histograms is nonlinear due to being mismatch in the related frequency band. Power warping causes that the maximum and minimum values of a specific channel are changed to the global maximum and minimum, respectively. Therefore, all data is redistributed, which tends to reduce noise effects.

6.2. Averaging weighted spectral

In some cases, the nonlinear warping is affected by the histogram matching. This causes that the noisy part is amplified compare to the parts without noise. This is related to the conditions that noise is filtered. Due to processing nature, a filtered channel is normally amplified. The histogram matching may partially suppress parts of speech with low energy. In order to balance the compensation of the unprocessed signal with the processed signal, we propose a methodology which calculates linear combination of spectral weighting by using the processed and the unprocessed spectrums. By averaging the post-processed weighted spectral, basic power spectrum linear combination, Porig, is related to process spectrum, Pproc, as:

$$P_{out}[m,l] = w.P_{orig}[m,l] + (1-w)\,P_{proc}[m,l]$$
$$0 \le w \le 1 \tag{13}$$

Where m and l are indices of frame and channel, respectively.

Experimentally, the weighting parameter w is determined based on training style and the type of noise. The optimum weighting parameter, w, depends on training style, SNR and the type of noise. For example, if additive white Gaussian noise (AWGN) with 5 dB SNR is used under multi-style training, the optimum value of w is 0.6 as shown in Fig. 8.

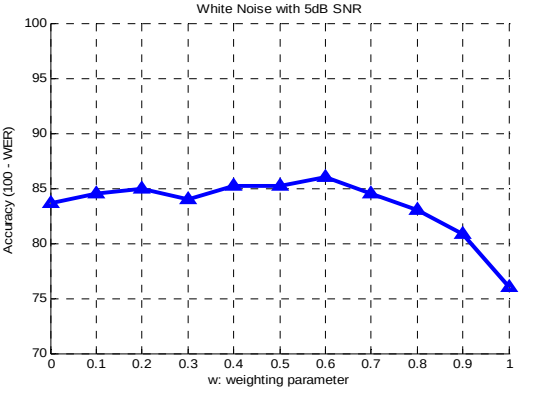


Fig. 8. system performance to various values of w in 5 dB SNR white noise under multi-style training

Table 1: optimum values of w for each SNR and training style for white noise

25dB SNR	20dB SNR	15dB SNR	10dB SNR	5dB SNR	0dB SNR	Training style
0.7	0.1	0.2	0	0	0	Clean
0	0.5	0.5	0.5	0.6	0.6	Multi-style training
0.7	0.1	0.1	0.2	0.3	0.3	Matched training

7. SPEECH RE-SYNTHESIS

Although the proposed system is performed on parametric based on STFT, we can obtain better recognition accuracy by re-synthesis of speech signal and then calculation of conventional cepstral-based features, rather than deriving cepstral parameters directly without re-synthesized speech. The speech re-synthesis causes that the proposed system is easily coordinated with traditional feature extraction algorithms. Speech re-synthesis is performed using overlap-add (OLA) algorithm [17].

8. EXPERIMENTAL RESULTS

Applied speech database is chosen from small FarsDat database [18], that is selected from repetitive and log sentences of speech. Therefore, recognition is continuous and independent from speaker. 200 sentences for training database and 62 sentences for testing database are selected. FarsDat database includes 44 acoustic labels. To prevent recognition model vague, some explosive phoneme packages are combined. Therefore, phoneme units are decreased to 35 units. Since phonic files of FarsDat database are recorded in silence mode, their quality is high and the ratio of signal to noise (SNR) is about 34 dB. Therefore to create noise database (typically additive noise), NOISEX-92 database is used [19]. In order to evaluate the performance of the proposed system against noise, we use three different additive noises: white noise, background music and reverberation. The background music has been taken from DARPA Hub 4 [19].

This part shows the recognition accuracy of the proposed system in presence of various SNRs from noise under various training styles. As mentioned in Sec. 6.2., the optimum w depends on training style, SNR and type of noise. Table 1 shows different values of w for white noise.

It is observed that optimum selection of w manually for each SNR, noise type, training condition, etc is impossible. We found if instead of blindly selecting optimal values of w, we spot that w is fixed, recognition accuracy still improves. This amount is good overall for each training style but it may not be the best choice for any particular noise type, training condition, or SNR. Related results to w=0.7 are shown in Fig. 9.

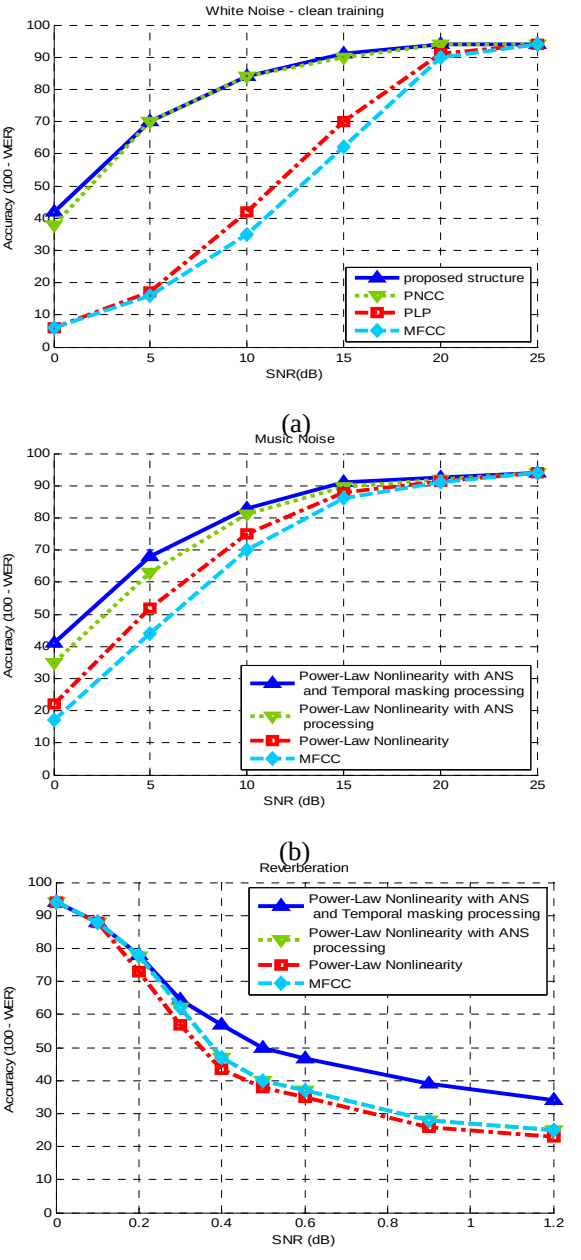


Fig. 9. Speech recognition accuracy that damaged with white noise and w=0.7 under different training styles: (a) training with clean data, (b) multi-style training, (c) matched training

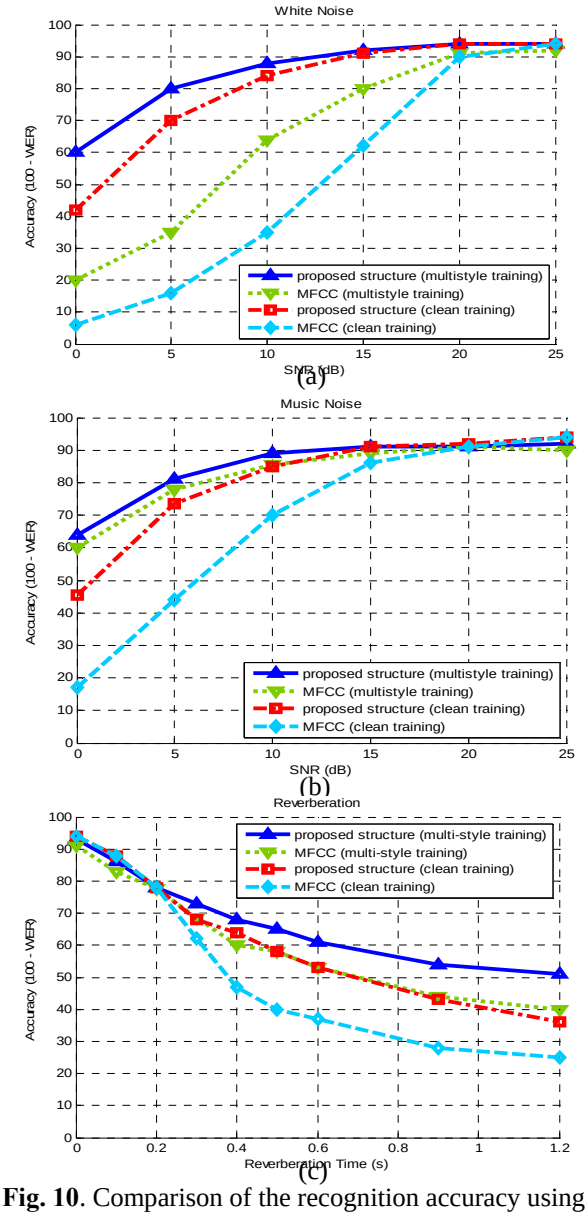


Fig. 10. Comparison of the recognition accuracy using MFCC features as clean training and multi-style training for different noises (a) white noise, (b) music noise, (c) reverberation.

In fact, Fig. 9 indicates the performance of the proposed system in comparison with MFCC, PLP and PNCC methods. As observed, the proposed system provides better recognition accuracy compared to the popular methods. The obtained results also show that taken an accurate mechanism to select w creates significant improvements with various SNRs of white noise in multi-style and matched training. As shown in Fig. 9, selection of fixed w (even if it is not optimized) leads to improvement of the recognition accuracy, specifically in multi-style and matched training.

After selection of w for each type of noise, we examine the recognition accuracy for three types of noise (white noise, background music and reverberation) and for both clean and multi-style training compared to MFCC method. The obtained results have been shown in Fig. 10. As observed, the proposed system provides better performance for three different noises.

9. CONCLUSIONS

In this paper, we proposed a new system containing a set of features which results in higher recognition accuracy in comparison with the popular methods such as MFCC, PLP and PNCC in noisy environment. The obtained results show that use of an accurate mechanism for selection of weighting parameter, w, improves the performance of the proposed system with different SNRs. Also, setting a constant value for weighting parameter, w, even if it is not the optimum value, results in more improvement on the recognition accuracy especially in multi-style training.

REFERENCES

[1] B. Atal, "**Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification**," *Journal of the Acoustical Society of America*, vol. 55, pp. 1304–1312, 1974.

[2] P. Jain and H. Hermansky, "I**mproved mean and variance normalization for robust speech recognition,**" in *IEEE Int. Conf. Acoust., Speech and Signal Processing*, 2001.

[3] X. Huang, A. Acero, and H-W Won, **"Spoken Language Processing: A Guide to Theory, Algorithm, and System Development**,*"Upper Saddle River, NJ: Prentice Hall*, 2001.

[4] Y. Obuchi, N. Hataoka, and R. M. Stern, "**Normalization of time-derivative parameters for robust speech recognition in small devices,**" *IEICE Transactions on Information and Systems*, vol. 87-D, no. 4, pp. 1004-1011, 2004.

[5] C.Kim and R.M Stern,"**Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring,"***in IEEE* Int. Conf. On

[6] R. Balchandran and R.J. Mammone, "**Non-parametric estimation and correction of non-linear distortion in speech systems**," in *IEEE Int. Conf. on Acoust., Speech and Signal Processing*, May 1998.

[7] S. Dharanipragada and M. Padmanabhan, "**A nonlinear unsupervised adaptation technique for speech recognition,**" in *Int. Conf. on Spoken Language Processing*, 2000, vol. 4, pp. 556–559.

[8] A. de la Torre et al., "**Non-linear transformations of the feature space for robust speech recognition,**" in *IEEE Int. Conf. on Acoust., Speech and Signal Processing*, 2002, pp. 401–404.

[9] F. Hilger**, "Quantile based histogram equalization for noise robust speech recognition**," *Ph.D. thesis*, Computer Science Department, RWTH Aachen University, Aachen, Germany, 2004.

[10] H. Hermansky and N. Morgan, "**RASTA processing of speech,**" *IEEE. Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–58, 1994.

[11] B. E. D. Kingsbury, N. Morgan, and, S. Greenberg, "**Robust speech recognition using the modulation spectrogram,**" *Speech Communication*, vol. 25, no. 1–3, pp. 117–132, 1998.

[12] H. G. Hirsch and C. Ehrlicher, "**Noise estimation techniques or robust speech recognition,"** in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, May 1995, pp. 153–156.

[13] C. Kim and R. M. Stern, "**Nonlinear enhancement of onset for robust speech recognition**," in *INTERSPEECH-2010*, Sept. 2010, pp. 2058–2061.

[14] S. F. Boll, "**Suppression of acoustic noise in speech using spectral subtraction,**" *IEEE Trans. Acoust., Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[15] C. Kim and R. M. Stern, "**Power function-based power distribution normalization algorithm for robust speech recognition**," in IEEE Automatic Speech Recognition and Understanding Workshop, pp.188-193, Dec. 2009.

[16] R.C. Gonzalez and R.E.Woods," **Digital Image Processing**, *Pearson Prentice Hall, Upper Saddle Ridge, NewJersey*, third edition, 2008.

[17] C. Kim, K. Kumar, and R.M. Stern, "**Robust speech recognition using a small power boosting algorithm,**" in *IEEE Automatic Speech Recognition and Understanding Workshop*, December 2009.

[18] M. Bijankhan and J. Sheikhzadegan, "**FARSDAT – The Speech Database of Farsi Spoken Language**," *Proc. 5th Australian Int. Conf. On Speech Science & Tech*., vol. 2, pp. 826-831, 1994.

[19] SPIB, SPIB noise data. Available from: <http://spib.rice.edu/spib/select_noise.html>