Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2014) 22: 451 – 464

doi:10.3906/elk-1207-76

A new method for the extraction of speech features using spectral-delta characteristics and invariant integration

Hassan FARSI*, Samana KUHIMOGHADAM
Department of Electronics and Communications Engineering, Faculty of Engineering, University of Birjand, Birjand, Iran

Received: 21.07.2012 • **Accepted:** 31.10.2012 • **Published Online:** 17.01.2014 • **Printed:** 14.02.2014

Abstract: We propose a new feature extraction algorithm that is robust against noise. Nonlinear filtering and temporal masking are used for the proposed algorithm. Since the current automatic speech recognition systems use invariant-integration and delta-delta techniques for speech feature extraction, the proposed algorithm improves speech recognition accuracy appropriately using a delta-spectral feature instead of invariant integration. One of the nonenvironmental factors that reduce recognition accuracy is the vocal tract length (VTL), leading to a mismatch between the training and testing data. We can use the invariant-integration feature idea for decreasing the VTL effects. The aim of this paper is to provide robust features that provide improvements in different noise conditions as well as being robust against VTL effect changes. This results in more improvement of the recognition accuracy in comparison with mel-frequency cepstral coefficients and perceptual linear prediction in the presence of different types of noises and scenarios.

Key words: Robust speech recognition, vocal tract length, temporal masking, invariant integration

1. Introduction

Although many speech recognition systems provide satisfactory results, one of the most important problems in speech recognition is recognition accuracy. If the training environment differs from the test environment, the recognition accuracy will be affected. These environmental differences are because of additive noise, diversion channels, and sound differences among various speakers.

The state-of-the-art automatic speech recognition (ASR) systems show excellent performance in a controlled environment. These are established for a certain noise, but, to date, there is no algorithm with acceptable accuracy in different noise environments. Cepstral mean normalization (CMN) [1] and mean and variance normalization [2] are the simplest forms of these techniques [3], in which it is assumed that the mean or the mean and the variance of the cepstral vectors should be equal for all utterances. Histogram equalization [4] is another strong method that assumes that all cepstral vectors have the same probability density function. The proposed algorithms in [5–7] provide acceptable results in quasistatic noise and weak results in very unstable environments, such as background music [8]. Recent studies in nonstatic environments, such as background music and background speaking, have come up with algorithms based on features missing [9] or feature extraction caused by physiological knowledge [10,11].

The ASR, naturally, is a process of pattern matching based on features achieved from nonlinear processing

*Correspondence: hfarsi@birjand.ac.ir

in time domain signals. Nonlinearity indicates that time domain optimization cannot be accurate [12] in the feature domain and, therefore, we use the feature domain for optimizing.

As mentioned, there are many different feature extractions for the robustness of ASR systems. In addition to noise immunity methods, such as relative spectral transform-perceptual linear prediction (PLP) [13], there are many algorithms for nonmatching audio between training and testing data. These methods are generally used for speech enhancement techniques. The CMN, stereo piecewise linear compensation for environments [14], and vector Taylor series [5] are some optimization methods to improve the extraction of speech features. The aim of these techniques is universally to omit noise effects from feature vectors by reducing the mismatch between the training and testing data, such as, for example, parallel model combination [15].

Improved methods for feature extraction in comparison with model adaption have less calculation, and thus they are more useful. One recent feature extraction method based on maximizing the sharpness of the power distribution and flooring power is called power-normalized cepstral coefficients (PNCC) [16].

Another effective factor is the number of speakers that leads to mismatch of the training and testing data. In other words, the vocal tract length (VTL) parameter differs among speakers. To cope with this problem, some approaches such as VTL normalization or maximum likelihood linear regression are used to counteract the distorting effects due to VTL differences that are applied after the feature extraction process. Since noise-robustness construction methods based on the features are required, there are some methods for extracting the invariant features of the VTL, such as invariant-integration features (IIFs) [17]. The invariant-integration method, in fact, was proposed to increase the ASR system's robustness against the VTL effects occurring between individual speakers. In this paper, we propose a new method, which uses asymmetric noise removal. 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, it could be said that speech has a higher modulation frequency spectrum than noise; therefore, many algorithms have been raised by band-pass filtering or high-pass filtering in the modulation spectrum domain [18]. The easiest method is high-pass filtering in each channel, which removes the components containing smooth changes [19,20].

A significant issue in the application of conventional linear high-pass filtering in the power domain is that the output power can become negative, which is mathematically impossible. In addition, it also introduces some problems into speech synthesis unless an appropriate floor value is used for power coefficients [20]. Thus, filtering can be performed after applying log nonlinearity [such as the mel-frequency cepstral coefficient (MFCC) method], but this is not applicable for additive noise as well. Spectral subtraction is another method for decreasing the effects of noise, whose power changes slowly [21]. The noise level is estimated in spectral subtraction techniques from the power speech parts [21] or through using a continuous-update approach [19]. We introduce a method that results in the time-variant estimation of the noise floor using an asymmetric filter, and then it is subtracted from the instantaneous power.

In this paper, we will discuss a method based on the delta-spectral. Although the characteristics of the delta-cepstral increase the ASR accuracy, since dynamic data are discussed in it, they are not robust against noise and reverberation.

Figure 1 shows the structure of the proposed system. As this figure suggests, a nonlinear filter, temporal masking, and delta-spectral feature are used to lead to the improvement of speech recognition. The proposed structure could be promising for option features that are not only robust against noise but also have robustness to the effects of VTL changes.

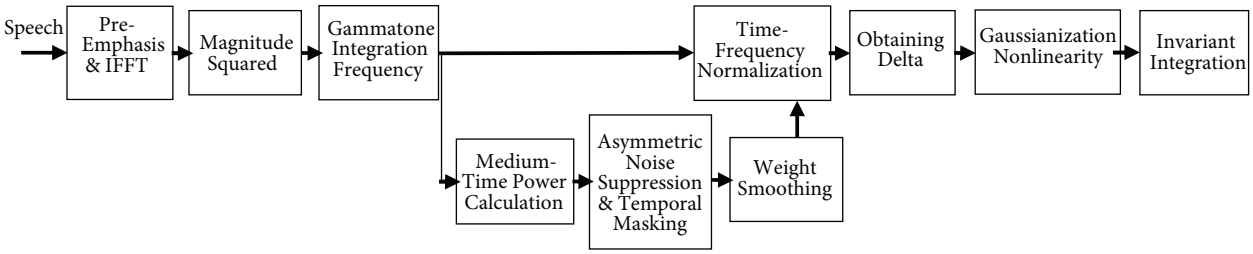

Figure 1. Block diagram of the proposed structure.

The overall structure of this paper is organized as follows: we explain an overall review of the proposed structure in Section 2. In Section 3, we describe the general characteristics of the nonlinear filters used in the proposed structure. Section 4 surveys temporal masking and sets its parameters. The delta-spectral feature and the formulae are described in Section 5. The general characteristics of IIFs are detailed in Section 6. The experimental results are presented in Section 7, and, finally, conclusions are drawn in Section 8.

2. Overall review of the proposed structure

The first stage is based on frequency analysis. A preemphasis filter H(z) = 1 – 0.97 z^{-1} is used and followed by a Hamming window having a time duration of 26.5 ms, with 10 ms between frames. Short-time Fourier transform is then performed and the 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 the frame and channel indices, respectively. It is mathematically represented as follows:

$$P[m,l] = \sum_{k=0}^{N_a-1} \left| X(m, e^{jw_k}) G_l(e^{jw_k}) \right|^2 \tag{1}$$

Here, N_a indicates the size of the fast Fourier transform. We use a 16-KHz sampling rate and N_a= 1024. After weighting the frequency, the power is normalized to the peak power. $G_l[k]$ is the gammatone filter bank for the lth channel and $X(m, e^{jw_k})$ shows the short-time spectrum of the speech signal for the mth frame. The center frequencies are linearly spaced at between 200 Hz and 8000 Hz in an equivalent rectangular bandwidth.

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

$$\hat{Q}[m,l] = \frac{1}{2M+1} \sum_{m'=m-M}^{m+M} P\left[m', l\right] \tag{2}$$

Selection of factor M has a significant effect on the performance (especially in the case of white noise). It is empirically found that if we chose the value of 2 for M, then the recognition accuracy would be optimum. Next, using both an asymmetric nonlinear filter and temporal masking for the compensation of environmental noise, we can improve the features. The effect of smoothing in increasing the recognition accuracy is known.

Next, the delta-spectral feature is used. Since this method alone could not lead to improve recognition, 'Gaussianization' nonlinearity is used.

After Gaussianization nonlinearity, according to [17], the invariant-integration approach is performed. It is a general approach for the construction of invariants for arbitrary transformation groups and its calculation

includes the collection of r functions (probably nonlinear) for all possible converted observations. Finally, the output feature vectors will be robust to both environmental conditions and VTL changes.

3. General characteristics of the asymmetric nonlinear filter

Figure 2 indicates the asymmetric noise suppression (ANS) process and temporal masking. First, we explain the general characteristics of the asymmetric nonlinear filter.

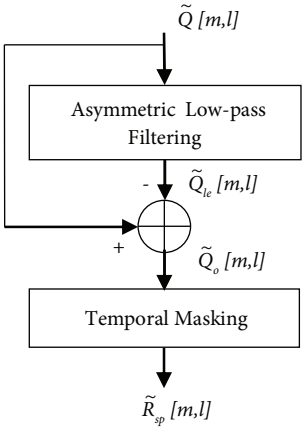

Figure 2. Block diagram to model ANS.

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

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

where m and l are the indices of the frame and channel, respectively, and λ_a and λ_b are constants with values between 0 and 1. If $\lambda_a=\lambda_b$, reviewing Eq. (1) will be easy, and since λ is positive, it will become a low-pass IIR filter, as observed in Figure 3a. If $<\lambda_a<\lambda_b<1$, then the nonlinear filter functions will become upper envelope detectors (Figure 3b), and, finally, as shown in Figure 3c, if $<\lambda_b<\lambda_a<1$, 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, a lower envelope with changes is applied. Therefore, as this envelope reduces in the main input, $\hat{Q}_{in}[m,l]$, slow changes in the nonspeech components are deleted. We use Eq. (4) to represent the nonlinear filter described by Eq. (1).

$$\hat{Q}_{out}[m,l]=AF_{\lambda_a,\lambda_b}\left[\hat{Q}_{in}[m,l]\right] \tag{4}$$

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

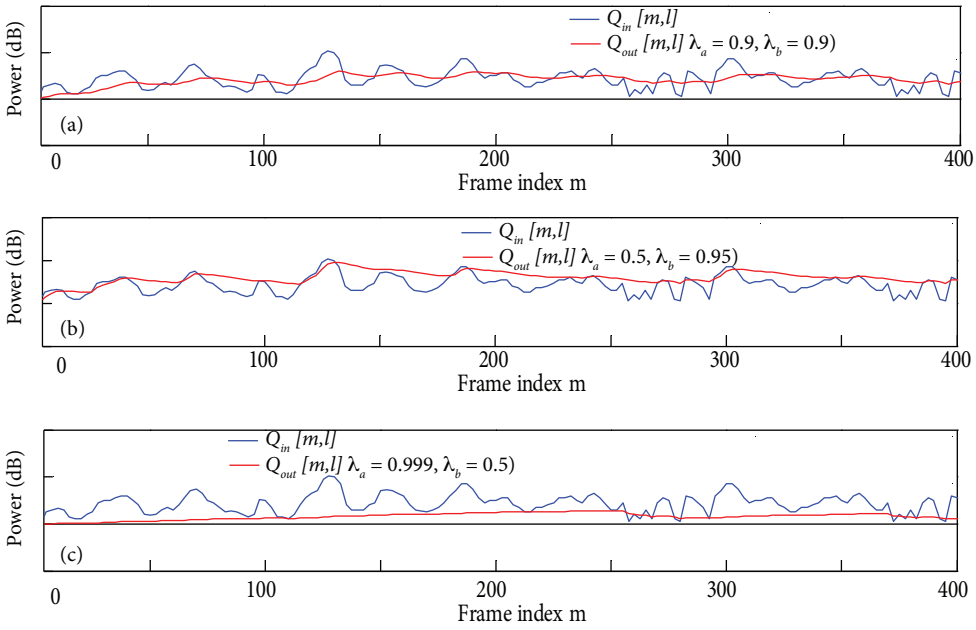

Figure 3. Sample input (solid curves) and output (dashed line curves) of the filter defined in Eq. (1) for different conditions when: a) $\lambda_a = \lambda_b$, b) $\lambda_a < \lambda_b$, and c) $\lambda_a > \lambda_b$.

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

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

Next, $\hat{Q}_{le}\,[m,l]$ is subtracted from $\hat{Q}_{in}\,[m,l]$. We can observe the results of the speech recognition caused by processing with the asymmetric nonlinear filter, after implementing this structure for different values of λ_a and λ_b. We add 3 kinds of noise: white noise, background music, and reverberation (with a delay of about 0.3 s). The experimental results are shown in Figures 4a–4d, where it is observed that the values of λ_b from 0.25 to 0.75 result in good recognition accuracy. According to these figures, the best value for λ_a is 0.9. Therefore, in practice, we consider λ_a= 0.999 and λ_b= 0.5, because the recognition accuracy for speech is maximum in the presence of noise.

4. Temporal masking

Many researchers have found that the human auditory system focuses more on the onset of an incoming power envelope in comparison with the falling edge of the same power envelope [22,23]. In this regard, several algorithms for improving the onset were proposed [20]. In this section, we display a simple method to incorporate this effect in the processing feature vectors extracted. It can be applied using a moving peak for each frequency channel l and omitting instantaneous power if it is under this envelope. This process is shown in a block diagram in Figure 5.

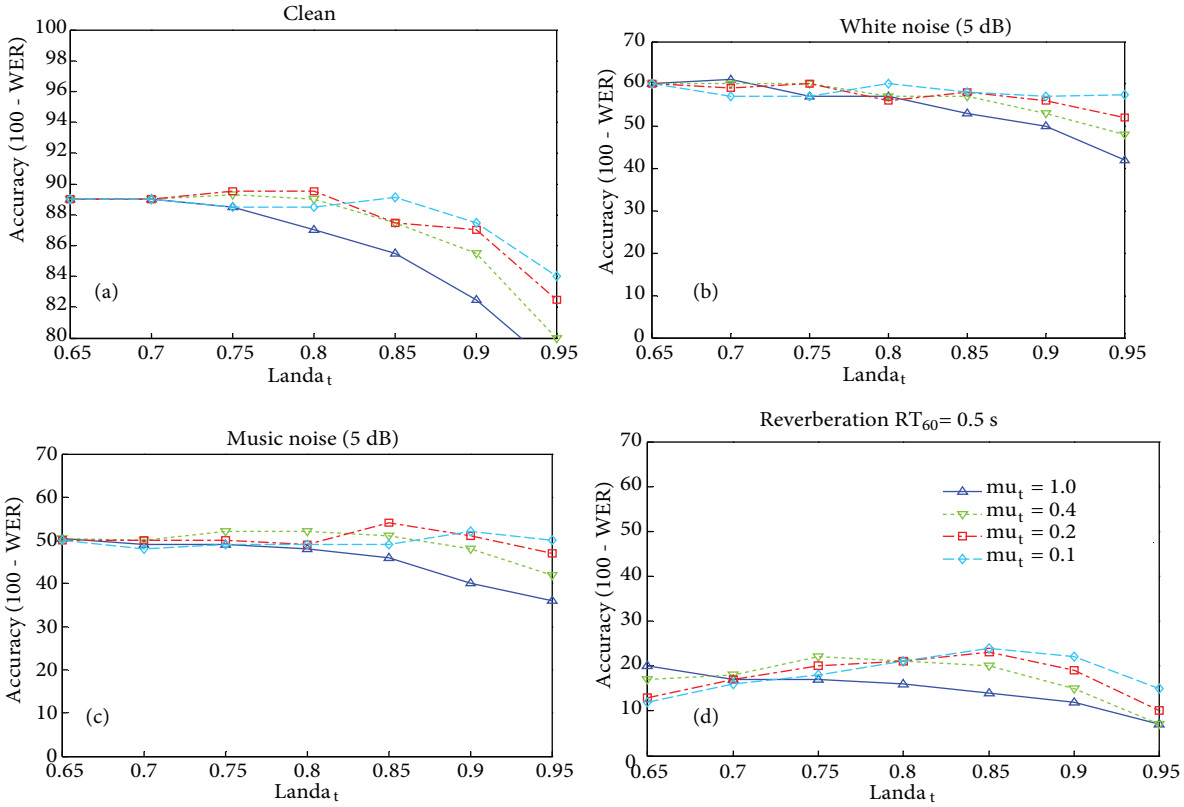

Figure 4. Relation between the forgetting factors (λ_a, λ_b) and recognition accuracy for speech: a) clean, b) 5-dB Gaussian white noise, c) 5-dB music noise, and d) reverberation with RT60 = 0.5.

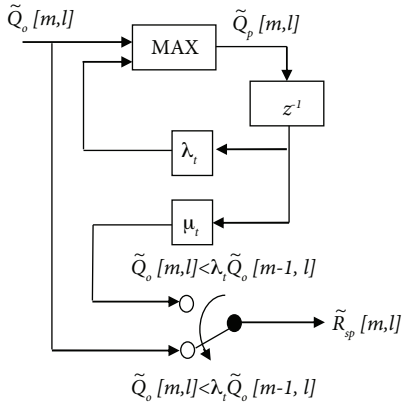

Figure 5. Block diagram of the model temporal masking.

In the first stage, the power of the online peak, $\hat{Q}_p[m,l]$, is calculated for each channel by:

$$\hat{Q}_p[m,l] = max\left(\lambda_t \hat{Q}_p[m-1,l]\ ,\ \hat{Q}_0[m,l]\right). \tag{6}$$

Here, λ_t is a forgetting factor for the calculation of the online peak, and m and l are the frame index and channel index, respectively.

Temporal masking for different parts of speech is obtained through the following equation.

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

Figures 6a–6d indicate the relationship between the recognition accuracy and the forgetting factor (λ_t), and also the elimination coefficient (μ_t). We represent the results of the recognition system using the complete structure in Figure 1 and just change the coefficients of the forgetting and elimination factors (λ_t, μ_t).

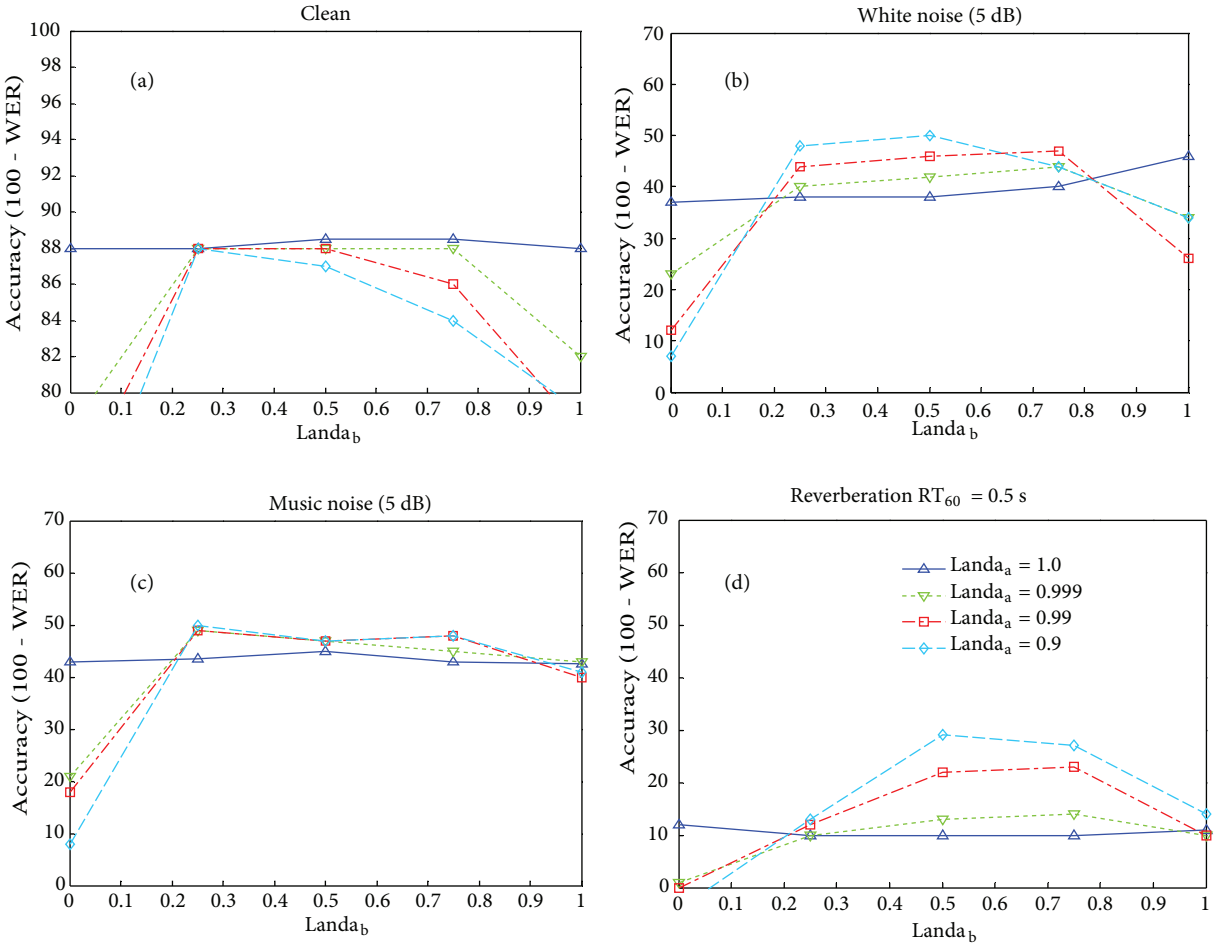

Figure 6. Relation between the speech recognition accuracy and the forgetting factor (λ_t) and elimination factor (μ_t): a) clean, b) 5-dB Gaussian white noise, c) 5-dB musical noise, and d) reverberation with RT60 = 0.5.

In a clean environment, as observed in Figure 6a, if $\lambda_t \leq 0.85$ and $\mu_t \leq 0.2$, the recognition accuracy will almost remain constant. However, if $\lambda_t > 0.85$, the performance will be degraded. In an additive noise environment such as weight or music noise, as shown in Figures 6b and 6c, the performance is the same. However, for the reverberation, as shown in Figure 6d, the application of the temporal masking scheme provides considerable improvement.

5. Delta-cepstral features

In this section, we improve the speech recognition accuracy in practical environments by placing delta features in the spectral domain instead of in the cepstral domain. Delta-cepstral features were suggested in a different way in [24], in the dynamics information for collecting static features. They also improve recognition accuracy by adding one characteristic of temporary attachment to hidden Markov model (HMM) frames that are supposed to be statically independent from each other.

Delta-cepstral features are defined by the following equation for a short-time cepstral sequence, C[n].

$$D[m] = C[m+M] - C[m-M] \tag{8}$$

Here, m is the index of the analysis frame and M is empirically gained at about 2 or 3. Similarly, delta-delta features are defined in terms of a subsequent delta operation on the delta-cepstral features. From Eq. (8), it can be easily proven that E{D[m].C[m]} = 0. Since E{.} is the expectation operator, the delta features are uncorrelated with the static features and help the frame to be independent from the HMM assumption in ASR.

The total delta-cepstral coefficients improve the ASR accuracy and also provide good robustness against noise [25].

The human ear can detect speech sounds in the presence of background noise. In other words, due to differences in the stationary characteristics of speech and noisy signals, the human ear can largely ignore the noise and concentrate on the speech signal. Since the noise spectral values are relatively flat, although the speech spectral values change quickly, taking a different approach across frames strongly attenuates the noise components.

5.1. Formal analysis of cepstral coefficients

In this subsection, formal analysis for improving the signal-to-noise ratio (SNR) is discussed using the spectral features in white noise. Suppose that the noise is a white Gaussian sequence sample distribution w_i of the form $N(0,\sigma^2)$, that power P for an independent set of N samples is $E[P]=\frac{1}{N}E\left[\sum_{i=1}^{N} w_i^2\right]$, and that power P indicates a chi-square distribution with N degrees of freedom that approximately has a Gaussian distribution for large N.

Under the assumed Gaussian for P, E[P] is given by the following.

$$E[P]=\frac{1}{N}E\left[\sum_{i=1}^{N} w_i^2\right]=\sigma^2 \tag{9}$$

$$var[P]= E\left[P^2\right]-E[P]^2=\frac{E\left[\sum_{i,j} w_i^2 w_j^2\right]}{N^2}-\sigma^2 \tag{10}$$

$$=\frac{1}{N^2}\left(\sum_i E\left[w_i^4\right]+\sum_{i,j,\, i\neq j} E\left[w_i^2 w_j^2\right]\right)-\sigma^4=\frac{2\sigma^4}{N^2} \tag{11}$$

Therefore, assuming Gaussian power, we could say that P is approximately distributed as $N(\sigma^2 \frac{2\sigma^4}{N^2})$. The total P can be considered as the sum of the DC and AC powers. The DC power is the square of the mean, σ^4, while

the AC power is the variance $\frac{2\sigma^4}{N^2}$. Processing in the spectral domain leads to removing DC power; therefore, a higher improvement is obtained. As is known, the largest part of noise is its DC power. In other words, we have the following.

$$Noise\,cancelling = -10\,log_{10}\left(\frac{P_{DC}}{P_{AC}+P_{DC}}\right) = 10log_{10}\left(1+\frac{N}{2}\right) \tag{12}$$

As previously mentioned, a window with a 25.6-ms length and sampling frequency of 16 KHz is used, so the number of sample durations, N, is equal to 410. This means that for the cancellation of white noise with delta-spectral alone, the maximum possible increase in accuracy is a 26.05-dB SNR.

5.2. Delta-spectral power coefficients

Now the delta-spectral power coefficients for the ASR are discussed. Since the short-time power of speech changes is faster than the short-time power of noise, we consider delta-spectral power coefficients for the proposed method. In fact, these large differences in speech power changes or noise make it possible to understand speech distinctions for humans in the noise.

The target is to use delta features for increasing recognition accuracy. With the delta function given by Eq. (8) in the spectral domain, rapid changes in the speech components are improved and slow changes in the noise components are inactivated. Figures 7a–7f show the effect of using delta features for increasing the recognition accuracy. Figure 7e indicates the result of a delta operation in the spectral domain in Figure 7b. It is obvious that this kind of delta in Figure 7e (caused by applying a delta operation in the spectral domain) is more useful than that in Figure 7c (caused by applying a delta operation in the cepstral domain).

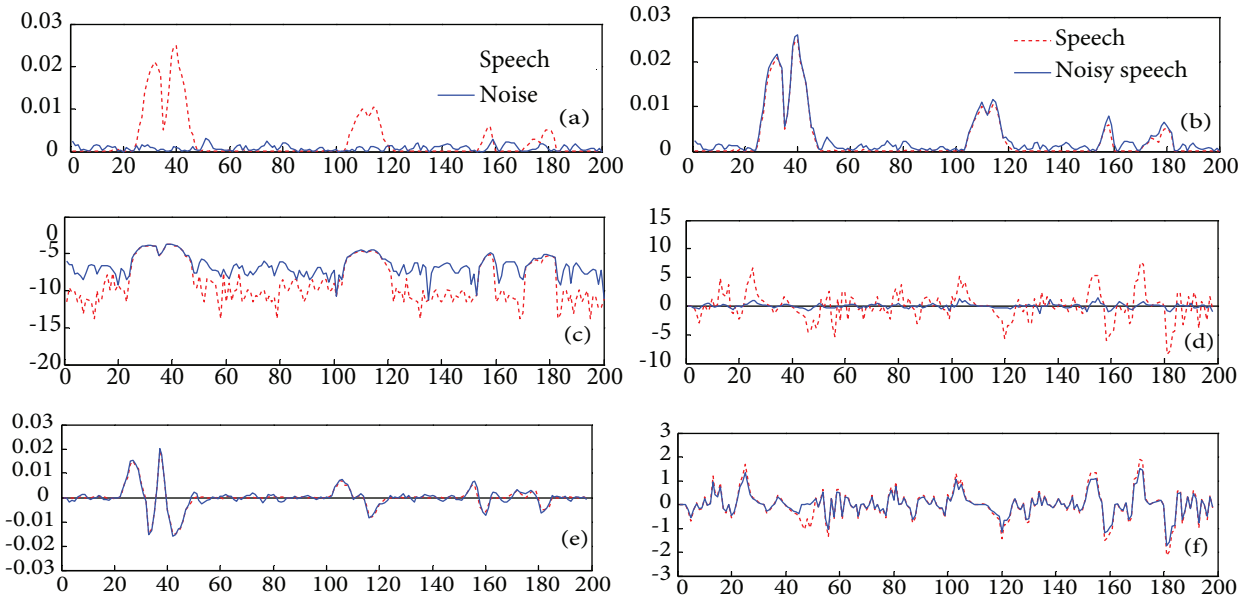

Figure 7. a) Short-time power for clean speech and a part of the white noise, 0 dB; b) short-time power for clean speech and noisy speech with the noise of Figure 7a; c) logarithmic power using clean speech and noisy speech in Figure 7b; d) temporal difference operation on the signals of Figure 7c; e) temporal difference operation on the signals of Figure 7b; and f) Gaussianization operation on the signals of Figure 7e.

Due to the compressive nature logarithmic of the nonlinearity, spectral peaks are almost identical for the clean and noisy speech. However, for the rest of the frames, the amount of mismatch is high, as shown in Figure 7c.

Overall, therefore, a delta-spectral outcome equivalent to the MFCC features in the cepstral domain or 'mel-filter spectrum' is obtained by applying a temporal difference operation to the spectral values of mel, which could be considered as a filtering operation with the mel-spectral feature sequences. This filter is shown as:

$$H_d(z) = z^d - z^{-d}. \tag{13}$$

It is experimentally found that parameter d has the best performance between 2 and 4. Since the use of these features alone would not be appropriate for speech recognition applications and as delta-spectral cepstral features are non-Gaussian, histogram normalization for these features is used to match the delta-spectral features with the speech recognition and then they are converted to Gaussianization nonlinearity. Hence, we apply it instead of the logarithmic nonlinearity used in the MFCC.

Though logarithmic nonlinearity is closer to the human auditory model, it is more vulnerable to noise. This Gaussianization nonlinearity is applied on an utterance-by-utterance basis. Figure 7f shows the Gaussianized delta-spectral features.

6. Invariant integration

In practice, with a given time-frequency representation, $y_l(m)$, the monomials are defined as in [17], where m and l are the frame and channel indices, respectively.

$$\hat{r}\left(m; w, \vec{l}, \vec{h}, \vec{r}\right) = \left[\prod_{i=1}^{M} y_{l_i+w}^{h_i}(m+r_i)\right]^{1/\sum_{i=1}^{M} h_i} \tag{14}$$

It is shown that vectors $\vec{l} \in N^M$, $\vec{h} \in N_0^M$, and $\vec{r} \in N^M$ describe the used subbands, integer components, and temporal offsets, respectively, and w introduces the subband-index offset. In addition, a monomial is evaluated on the several converted versions of each frame and then on a window with a size of 2w + 1, and the results are averaged:

$$A_{\hat{r}}(n) = \frac{1}{2w+1} \sum_{w=-w}^{w} \hat{r}\left(m; w, \vec{l}, \vec{h}, \vec{r}\right) \tag{15}$$

The final feature vector $\vec{A} \in R^M$ is a combination of these averages.

$$\vec{A}(n) = \left(A_{\hat{r}_1}(n), A_{\hat{r}_2}(n), \ldots, A_{\hat{r}_N}(n)\right) \tag{16}$$

The last stage of the calculations of invariant integration is a component-wise mean subtraction. The IIF parameters are computed by the iterative method of the feature selection based on the linear classifier in [26]. In [17], it was shown that an IIF set results in a significant rise in the accuracy of the MFCC, whether the training and testing data are not matching in the VTL or are a matching average. In the analysis of the IIF, the discrete cosine transform is replaced with an invariant integration to make a shift in the time-frequency display resulting from the VTL differences. For quantitative measurements, experiments of recognition are performed under different noise conditions and training-testing scenarios toward the average VTL.

In order to select the monomial parameters properly, they are calculated under clean speech conditions with no matching between the training and testing data. Less improvement is obtained in the noise (especially with less SNR); however, the result is better than with the PNCC and MFCC.

7. Experimental results

In this section, the simulation results are compared with the common feature extraction methods, MFCC, and PLP, and the new PNCC method suggested recently provides good improvements. To allow for an assessment of the performance of the feature types under mismatching training-test conditions with respect to the average VTL, 2 different scenarios are defined: the first is the VTL match between the training and testing data and the second is the VTL mismatch between those data. The mismatching VTL scenario uses only the male utterances from the training set for training and only the female utterances from the test set for testing. In the training, a clean database with matching is applied.

For the experiments, the Persian Large Vocabulary Speech Recognition System is used. The modeling of the phonetic units is performed by the HMM. HMM models using a continuous density are combined with the Gaussian. The mentioned phonetic unit is a word and 1 HMM is trained for each word.

We choose clean environment training and a test database from FarsDat [27] that includes 140 h of speech uttered by 300 speakers with 10 different dialects. Each person expresses 4000 words. All of the existent signals of this database are labeled in the word.

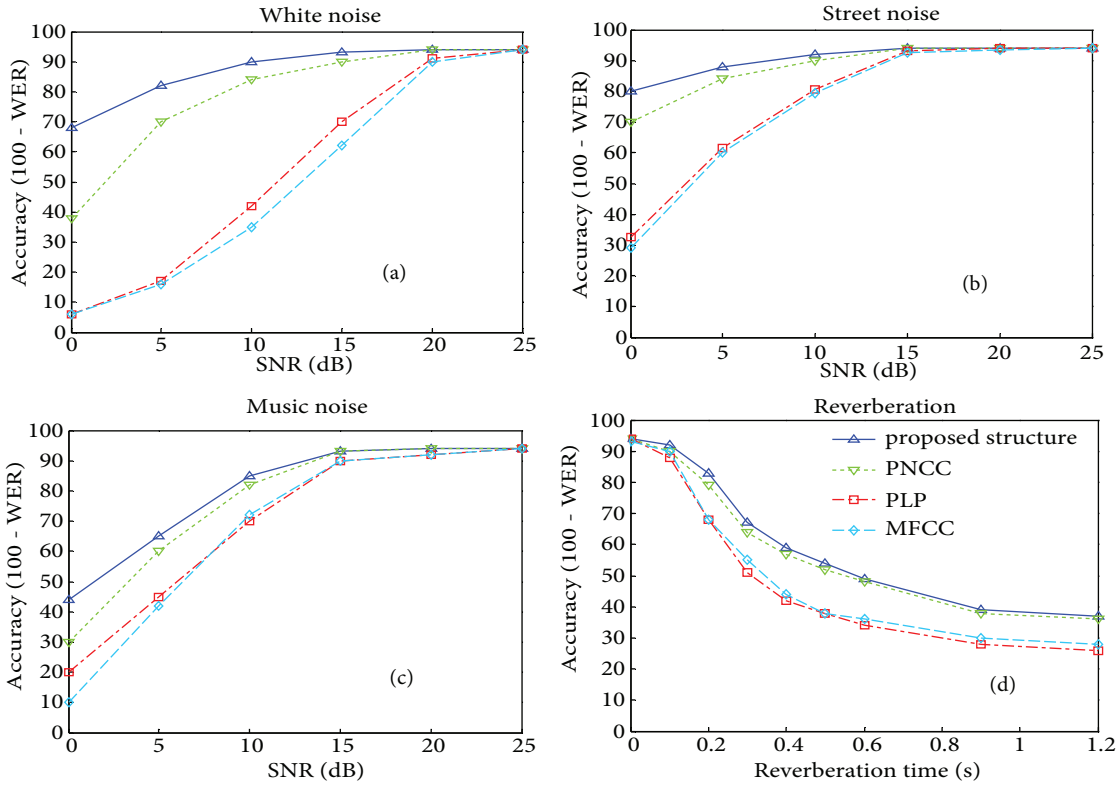

Figure 8. Speech recognition accuracy obtained from speaking under noise with different kinds of noises and the VTL matching scenarios between the training and test data: a) white noise, b) street noise, c) background music, and d) reverberation.

The recognition accuracy is compared between the proposed method and different kinds of feature extraction methods including MFCC, PLP, and PNCC under noise condition, match, and mismatch scenarios.

Figures 8a–8d show the recognition results of the VTL matching scenarios between the training and testing in the presence of white noise, street noise, background music, and reverberation. As observed, the proposed structure has the best performance in the presence of noise. These improvements are more than those of the PNCCs. More specifically, it provides more improvement under white noise, and compared to the MFCC, the recognition accuracy is increased by a value of about 16 dB. For the street noise and background music, it has also improved the accuracy by about 10 and 5 dB, respectively. Under noise of reverberation, there is no difference when compared with the PNCC; however, it provides more improvement than the MFCC.

Figures 9a–9d show the recognition results for the VTL mismatch scenarios between the training and test data in the presence of white noise, street noise, background music, and reverberation. The obtained results show that the proposed structure provides more improvements in recognition accuracy under mismatch conditions.

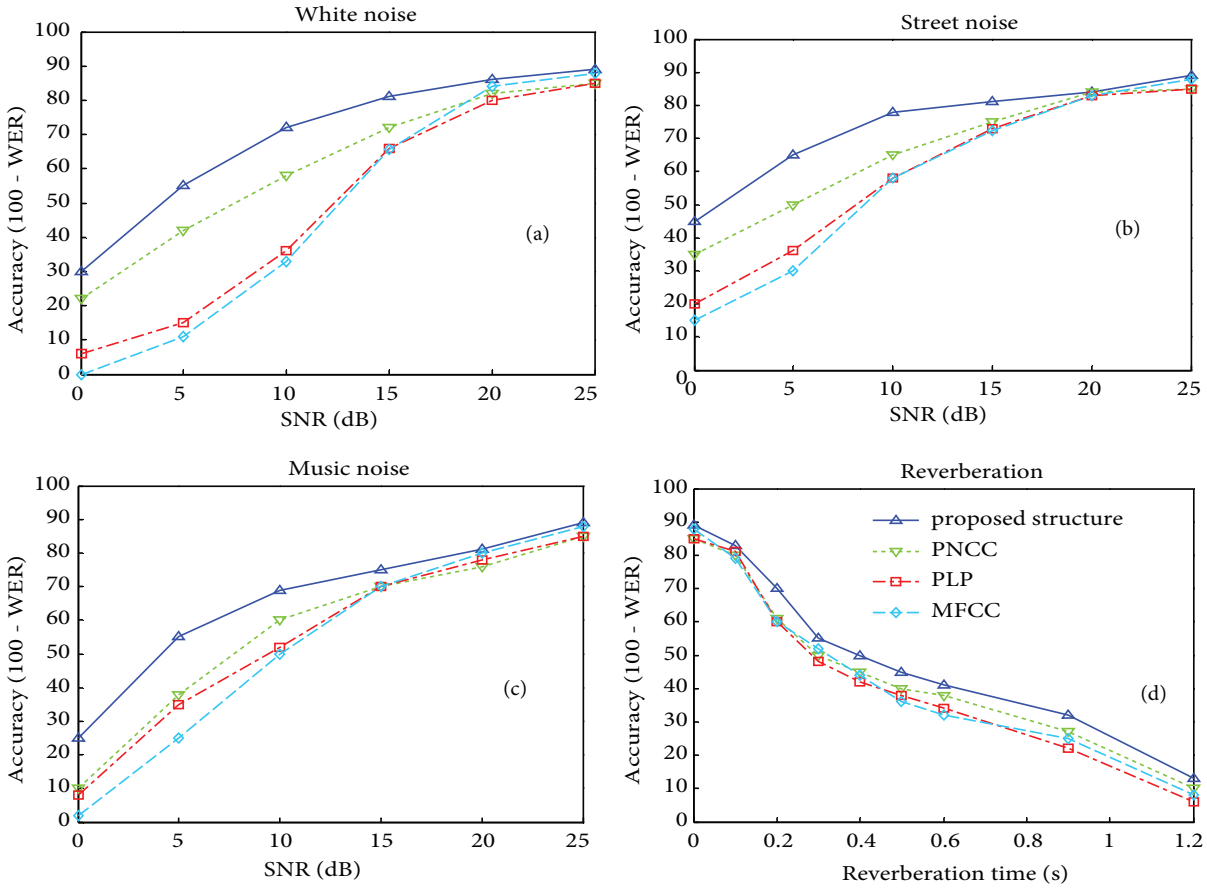

Figure 9. Speech recognition accuracy obtained from speaking under noise with different kinds of noises and the VTL mismatching scenarios between the training and test data: a) white noise, b) street noise, c) background music, and d) reverberation.

8. Conclusions

In this paper, we proposed a new structure for the extraction of speech features using spectral-delta characteristics and the invariant-integration method. This structure makes speech robust against a noise environment by

integrating the delta-spectral method and normalized distribution of the power function. Therefore, we were able to increase the robustness of speech to changes over the VTL that depend on speakers using IIFs. In other words, this structure can not only cause the speech feature to be more robust against noise, but can also provide more robustness under VTL mismatch conditions. The experimental results showed that the proposed structure provides better performance under different noise conditions and in both match and mismatch VTL scenarios between the training and test data.

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, H. Hermansky, "Improved mean and variance normalization for robust speech recognition", Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001.

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

[4] Y. Obuchi, N. Hataoka, R.M. Stern, "Normalization of time-derivative parameters for robust speech recognition in small devices", IEICE Transactions on Information and Systems, Vol. 87, pp. 1004–1011, 2004.

[5] P.J. Moreno, B. Raj, R.M. Stern, "A vector Taylor series approach for environment-independent speech recognition", Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 733–736, 1996.

[6] R.M. Stern, B. Raj, P.J. Moreno, "Compensation for environmental degradation in automatic speech recognition", Proceedings of the ESCA Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels, pp. 33–42,1997.

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

[8] B. Raj, V.N. Parikh, R.M. Stern, "The effects of background music on speech recognition accuracy", Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 851–854, 1997.

[9] B. Raj, R.M. Stern, "Missing-feature methods for robust automatic speech recognition", IEEE Signal Processing Magazine, Vol. 22, pp. 101–116, 2005.

[10] H. Hermansky, "Perceptual linear prediction analysis of speech", Journal of the Acoustical Society of America, Vol. 87, pp. 1738–1752, 1990.

[11] C. Kim, H. Chiu, R.M. Stern, "Physiologically-motivated synchrony-based processing for robust automatic speech recognition", InterSpeech, pp. 1975–1978, 2006.

[12] K. Kumar, "A spectro-temporal framework for compensation of reverberation for speech recognition", PhD, Carnegie Mellon University, Pittsburg, PA, USA, 2011.

[13] H. Hermansky, N. Morgan, "RASTA processing of speech", IEEE Transactions on Audio Speech and Language Processing, Vol. 2, pp. 578–58, 1994.

[14] L. Deng, A. Acero, M. Plumpe, X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments", Proceedings of the International Conference on Spoken Language Processing, pp. 806–809, 2000.

[15] M.J.F. Gales, "Model-based techniques for noise robust speech recognition", PhD, Cambridge University, Cambridge, UK, 1995.

[16] C. Kim, R.M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring", Proceedings of the International Conference on Audio, Speech, and Signal Processing, pp. 4574–4577, 2010.

[17] F. Muller, A. Mertins, "Contextual invariant-integration features for improved speaker-independent speech recognition", Speech Communication, Vol. 53, pp. 830–841, 2011.

[18] B.E.D. Kingsbury, N. Morgan, S. Greenberg, "Robust speech recognition using the modulation spectrogram", Speech Communication, Vol. 25, pp. 117–132, 1998.

[19] H.G. Hirsch, C. Ehrlicher, "Noise estimation techniques or robust speech recognition", IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 153–156, 1995.

[20] C. Kim, R.M. Stern, "Nonlinear enhancement of onset for robust speech recognition", InterSpeech, pp. 2058–2061, 2010.

[21] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 27, pp. 113–120, 1979.

[22] C. Lemyre, M. Jelinek, R. Lefebvre, "New approach to voiced onset detection in speech signal and its application for frame error concealment", IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 4757–4760, 2008.

[23] S.R.M. Prasanna, P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 17, pp. 556–565, 2009.

[24] F.Müller, A. Mertins, "Noise robust speaker-independent speech recognition with invariant-integration features using power-bias subtraction", Speech Communication, Vol. 53, pp. 830–841, 2011.

[25] S. Furui, "Speaker-independent isolated word recognition based on emphasized spectral dynamics", Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 1986.

[26] T. Gramss, "Word recognition with the feature finding neural network (FFNN)", Proceedings of the IEEE Workshop on Neural Networks for Signal Processing, pp. 289–298, 1991.

[27] M. Bijankhan, J. Sheikhzadegan, "FARSDAT – The speech database of Farsi spoken language", Proceedings of the 5th Australian International Conference on Speech Science and Technology, Vol. 2, pp. 826–831, 1994.