A Framework for Robust MFCC Feature Extraction Using SNR-Dependent Compression of Enhanced Mel Filter Bank Energies

Babak Nasersharif, Ahmad Akbari

Computer Engineering Department
Iran University of Science and Technology
{nasser_s, akbari}@iust.ac.ir

Abstract

The Mel-frequency cepstral coefficients (MFCC) are most widely used and successful features for speech recognition. But, their performance degrades in presence of additive noise. In this paper, we propose a noise compensation method for Mel filter bank energies and so MFCC features. This compensation method includes two steps: Mel sub-band spectral subtraction and then compression of Mel-Sub-band energies. In the compression step, we propose a sub-band SNR-dependent compression function. We use this function instead of logarithm function in conventional MFCC feature extraction in presence of additive noise. Experimental results show that the proposed method significantly improves MFCC features performance in noisy conditions where it decreases word error rate about 70% in SNR value of 0 dB for different types of additive noise.

Index Terms: Mel sub-bands, spectral subtraction, SNR-dependent compression, MFCC

1. Introduction

Traditional speech features are typically extracted from power spectrum or amplitude spectrum of speech signal. Then, when speech spectrum is changed due to presence of additive noise, these features show a high sensitivity to the noise. This usually results in performance degradation of speech recognition system in presence of additive noise.

Several techniques have been proposed to reduce sensitivity of features to external noise. In some approaches, a transformation is directly applied to feature vectors to compensate noise effects on them. Sometimes, the transformation is applied to cepstral domain such as cepstral mean normalization (CMN) [11]. In some other kinds of such techniques the transformation is applied to logarithm of spectrum or logarithm of filter bank energies (LFBE) such as vector Taylor series [11] and weighted Mel filter bank analysis [4][9].

Some other groups of methods work at the spectral level. These methods try to reduce the effect of additive noise on the speech spectrum and then extract features. Spectral subtraction [8][11] and different spectral filtering techniques are well known examples of such methods. Spectral subtraction, subtracts an estimation of noise spectrum from speech power spectrum to remove noise effects from it. Phase autocorrelation (PAC) is another example of these techniques that is recently introduced [7]. It tries to make autocorrelation coefficient less sensitive to additive noise [1][7]. Group delay function (GDF), negative derivative of speech phase spectrum, is another technique which can be used for robust feature extraction [6]. In group delay function, features derive from speech phase spectrum instead of speech power or amplitude spectrum [6].

In this paper, we propose a transformation for applying to Mel sub-bands energies in order to remove noise effects from MFCC features. Our proposed method includes two steps: Mel sub-band spectral subtraction and then SNR-dependent sub-band energy compression in place of logarithm function. While other works only use weighted logarithm of Mel filter bank energies [2][3][4][5][9] or only noise subtraction [2][8], we propose to benefit from sub-band spectral subtraction along with SNR-dependent sub-band compression without using logarithm function.

The remainder of this paper is organized as follows. In Section 2, we propose our framework for removing noise effects from MFCC features. Section 3, describes the used method for Mel sub-band spectral subtraction. In section 4, we define our SNR-dependent compression function for Mel-sub-band energies. Section 5 includes our experiments and results. Finally, we give our conclusion in section 6.

2. Proposed Framework for Compensating of Noise Effects on MFCC features

The conventional Mel-frequency cepstral coefficients (MFCC) show a very good performance for clean speech recognition. In spite of their popularity, they have this weakness that they show poor performance in noisy conditions. To overcome this problem, we propose a framework to compensate additive noise effects on MFCC features. So, we first discuss the general process of MFCC feature extraction from the speech signal.

Assume that x(n) represents the frame of a speech signal that is pre-emphasized and hamming windowed. The frame x(n), where $1 \le n \le N$, is transformed from time domain to frequency domain by applying an N-Point fast Fourier transform (FFT) and the resulting amplitude spectrum is shown by |X(k)|, where $1 \le k \le N$, and k is frequency index. Then, the filter bank energy E_i^X passing through i-th Mel scaled critical band band-pass filter $\psi_i(k)$ is calculated as follows:

$$E_i^X = \sum_{k=1}^{N} |X(k)|^2 \cdot \psi_i(k) \qquad (1)$$

where $1 \le i \le M$, and M is number of Mel-scaled triangular band-pass filter. After this, a discrete cosine transform (DCT) is applied to logarithm of filter bank energies. Thus, the Mel frequency cesptral coefficients for frame x can be expressed as:

$$c_l^x = \sum_{i=1}^{M} \log(E_i^X) \cos\left[l.\frac{(2i-1)\pi}{2M}\right] \qquad (2)$$

where 1≤l ≤L, and L is desired number of cepstral features.
Assuming that x(n) is noisy speech frame, we define our proposed noise compensation framework based on filter bank energies calculated in relation (1). The general form of proposed method can be shown by:

$$\hat{E}_i^X = F(E_i^X, w_i, b_i) = (E_i^X - b_i)^{w_i} \qquad (3)$$

where $\hat{E}_i^X$ is compensated Mel filter bank output and w_i and b_i are compensation parameters. The parameter w_i is the compression factor and the bias b_i depends on noise spectral characteristics.
Relation (3) includes two steps: the subtraction and then filter bank energy compression. In the subtraction step, we reduce the filter bank energy increased due to presence of additive noise. After that, in the compression step, we emphasize those filter bank energies less affected by noise and distortion generated by the subtraction.
In the subtraction step, we must estimate parameter b_i for each Mel sub-band and then do the subtraction. For this purpose, we use the noise estimation in Mel sub-bands and then we perform Mel sub-band spectral subtraction. We will discuss this method in section 3. In the compression step, we use from sub-band SNR-dependent compression factor in order to put emphasis on Mel sub-bands less affected by noise and distortion created after the subtraction.
After these two steps, we can calculate the compensated MFCC using following equation:

$$\hat{c}_l^x = \sum_{i=1}^{M} \hat{E}_i^X \cos\left[l.\frac{(2i-1)\pi}{2M}\right] = \sum_{i=1}^{M} (E_i^X - b_i)^{w_i} \cos\left[l.\frac{(2i-1)\pi}{2M}\right] \qquad (4)$$

where $\hat{c}_l^x$ show compensated MFCC.
It can be seen from equation (4) that we replace logarithm function by the proposed compression function. This function discriminates filter bank energies better than logarithm function in presence of additive noise. The usefulness of a compression function in comparison to logarithm function has shown in root cepstral analysis [5][12]. In addition, from the viewpoint of psychoacoustic, the compression process is also performed in human's ear, where the sound intensity is converted to the perceived loudness [5].
Fig. 1 shows our general proposed method for removing noise effects from MFCC features. In the following sections, we discuss each of two mentioned compensation steps in detail.

3. Mel Sub-Band Spectral Subtraction

Conventional power spectral subtraction is defined as follows [8][11]:

$$|\hat{S}(k)|^2 = \begin{cases} |X(k)|^2 - \alpha |N(k)|^2 & if\ |X(k)|^2 > \dfrac{\alpha}{1-\beta}|N(k|^2 \\ \beta |S(k)|^2 & otherwise \end{cases} \qquad (5)$$

where $|\hat{S}(k)|^2$, $|X(k)|^2$ and $|N(k)|^2$ are the power spectra of enhanced speech, noisy speech and estimated noise, respectively. α is over-estimation factor and β is to define spectral flooring.

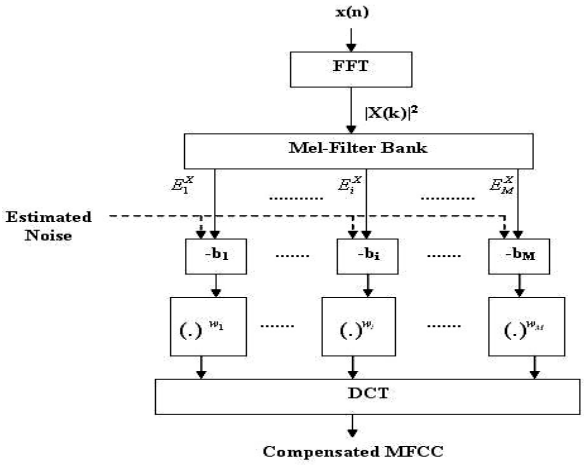

Fig.1. Block diagram of proposed method for compensating of noise effects on MFCC features

In this paper, we use Mel sub-band spectral subtract ion that is more useful than full-band spectral subtraction as mentioned in [8]. Mel sub-band spectral subtraction relation is defined as:

$$E_i^{ss} = E_i^X - b_i = \begin{cases} E_i^X - \alpha_i E_i^N & if\ E_i^X > \dfrac{\alpha_i}{1-\beta_i} E_i^N \\ \beta_i E_i^X & otherwise \end{cases} \qquad (6)$$

where E_i^{ss} is enhanced filter bank energy after Mel-sub band spectral subtraction and α_i and β_i are over-estimation factor and spectral flooring parameter in i-th Mel sub-band, respectively. E_i^N is the output of i-th triangular Mel scaled band-pass filter when estimated noise $|N(k)|^2$ is passed through Mel filter bank. It can be defined as follows:

$$E_i^N = \sum_{k=1}^{N} |N(k)|^2 \cdot \psi_i(k) \qquad (7)$$

Using equation (6), we can compute parameter b_i in equation (3) as:

$$b_i = \begin{cases} \alpha_i E_i^N & if\ E_i^X > \dfrac{\alpha_i}{1-\beta_i} E_i^N \\ (1-\beta_i) E_i^X & otherwise \end{cases} \qquad (8)$$

According to equation (8), the parameter b_i depends on energy of estimated noise in i-th Mel sub-band and its corresponding over-estimation and spectral flooring factors. For estimating noise at Mel-sub-bands, we firstly estimate the noise power spectrum at the duration of 300 ms where only the noise is present. We use from following smoothing equation for the noise power spectrum estimation:

$$|N(k)|^2 = P_t(k) = \lambda P_{t-1}(k) + (1-\lambda)|B_t(k)|^2 \qquad (9)$$

where $P_{t-1}(k)$ *and* $|B_t(k)|^2$ are estimated noise power spectra in previous t-1 frames and current frame, respectively. λ is forgetting factor and k is frequency index. We have selected λ =0.98 in this work as in [10]. The estimation of noise in i-th Mel sub-band is obtained using equation (7).

4. Compression of Mel Sub-Band Energies

One property of logarithmic compression of Mel filter bank energies is reduction of their dynamic range. This property has

two drawbacks in presence of additive noise. First, it can not emphasize on sub-bands energies that less affected by noise. Second, some distortions that are insignificants in power spectrum domain may become important after the logarithmic compression of Mel filter bank energies. In other hand, DCT is a linear transform that gives equal weights to all compressed sub-band energies. These disadvantages of DCT and logarithmic compression make MFCC features highly sensitive to additive noise. One solution to this problem is weighting of logarithm of filter bank energies as done in [3][4][9]. Another existing solution is root cepstral analysis [12] which substitutes the logarithm function with a root function. The root function uses a constant root for filter bank compression and is more immune to noise in comparison to logarithm function. The root Mel-frequency cepstral coefficients are computed as follows:

$$rc_l^x = \sum_{i=1}^{M} (E_i^X)^{\gamma} \cos\left[l.\frac{(2i-1)\pi}{2M}\right] \qquad (10)$$

where rc_l^x denotes the root MFCC (RMFCC) features, γ is constant root with a value between 0 and 1,and i is Mel filter index.

Although, using a constant root γ is better than logarithm function in presence of noise, but it is also a sub-optimal approach. Because, it doesn't notice to way that noise affects on Mel frequency sub-bands. In this paper, we propose a compression function that uses from SNR in Mel sub-bands. We define our proposed compression function for determining w_i in equations (3) and (4) as:

$$w_i = \gamma.\left[1-\exp(-\frac{SNR_i}{\xi_i})\right] \qquad (11)$$

where γ is a constant root and ξ_i is the gain to control the steepness of the compression function. SNR_i is signal to noise ratio in i-th Mel frequency sub-band computed as:

$$SNR_i = \left(1+\frac{E_i^{ss}}{E_i^N}\right)^{0.5} \qquad (12)$$

where square root has been used for reducing the dynamic range of energy ratio and E_i^{ss} and E_i^N have been defined in equations (6) and (7).

The parameter ξ_i in the compression function is calculated based on SNR_i as follows:

$$\xi_i = \frac{1}{1+\exp\left(\frac{SNR_i-\mu_{SNR}}{\sigma_{SNR}}\right)} \qquad (13)$$

where μ_{SNR} and σ_{SNR} are mean and standard deviation of SNR_i computed from all of Mel frequency sub-bands of the speech frame. Fig.2 shows ξ_i values in different Mel sub-bands of a speech frame and corresponding SNR_i values. It can be seen from the figure that when SNR_i is high, ξ_i has a small value. In such cases, w_i in equation (11) is very close to constant root γ. In this figure, when SNR_i is low, ξ_i has a value near to 0.6. This cause that w_i in equation (11) becomes a fraction of γ and so becomes less than γ.

Therefore, according to equation (11), the compression root w_i increases with a slope in accordance with the sub-band SNR. When sub-band SNR is smaller, this slope is steeper. So, when SNR_i is low, the compression root w_i decreases. On the other hand, for large SNR_i values, compression root w_i is simplified to the constant root γ presented in root cepstral analysis.

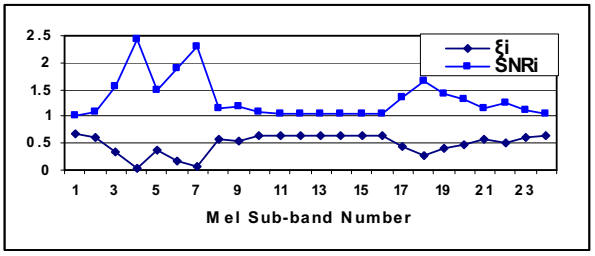

Fig. 2. SNR_i and ξ_i values in different Mel sub-bands for a noisy speech frame in presence of white noise with SNR= 0 dB

5. Experiments and Results

We report our results on TIMIT database for isolated word recognition. Two sentences from speakers in two dialect regions were selected and were segmented into words. In this way, we have 21 words spoken by 151 speakers including 49 females and 102 males. These speakers were divided into train and test speakers according to TIMIT speakers division. Our training set contains 2349 utterances spoken by 114 speakers. The testing set includes 777 utterances spoken by 37 speakers. Our recognizer is CDHMM with 6 states and 8 Gaussian mixtures per state which is trained on clean speech. Three types of additive noises were used: white, pink and factory noises selected from NOISEX92 database. We added these noises to both training and testing sets. Our feature vector in all cases (conventional or compensated form) contains 12 MFCC features and 12 delta-MFCC features and so its length is 24.

For evaluating our proposed compensation method, we have also used Mel sub-band spectral subtraction in company with the conventional logarithm function. This can be expressed by following equation:

$$sc_l^x = \sum_{i=1}^{M} \log(E_i^{SS}) \cos\left[l.\frac{(2i-1)\pi}{2M}\right] \qquad (14)$$

where *sc* denotes the obtained MFCC feature. E_i^{SS} was also defined in equation (6). We show this feature by ***LMSBS*** that is an abbreviation for Logarithm and Mel Sub-Band Spectral subtraction.

We also use from word ***CMSBS*** for our proposed method which stands for Compression and Mel Sub-Band Spectral subtraction. We have chosen α_i =1 and β_i =0.1 for all Mel sub-bands in equations (6) and (8) based on empirical results. Additionally, we have given the value of 0.5 to constant root γ in order to determine w_i in equation (11). Moreover, we compare CMSBS method with constant root cesptral analysis where we choose the constant root equal to 0.5. This means that γ is equal to 0.5 in equation (10). This can be written as:

$$rc_l^x = \sum_{i=1}^{M} (E_i^X)^{0.5} \cos\left[l.\frac{(2i-1)\pi}{2M}\right] \qquad (15)$$

We show the MFCC features obtained form equation (15) by ***RMFCC*** (Root MFCC).

Furthermore, we have performed Mel spectral subtraction together with constant root γ =0.5 that can be shown by:

$$src_l^x = \sum_{i=1}^{M} (E_i^{SS})^{0.5} \cos\left[l.\frac{(2i-1)\pi}{2M}\right] \qquad (16)$$

where *src* denotes the obtained MFCC feature. E_i^{SS} was also introduced in equation (6). We use the abbreviation ***RSMFCC*** for MFCC features obtained from equation (16).

Fig. 3 shows word error rate in presence of 3 different noise types (factory, pink and white) for different SNR values. The results are reported for all 3216 utterances of testing and training noisy database. We have shown the baseline MFCC results in top of figures in order to demonstrate proposed method results more clearly. As can be seen in the figure, the proposed method CMSBS has the lowest word error rate among other methods in presence of all three noise types. It can be seen that results of all methods are very significant and noticeable in comparison to conventional MFCC, especially in low SNR values. In SNR value of 0 dB, CMSBS word error rates are 10.15%, 7.75% and 6.84% for white, pink and factory noises,

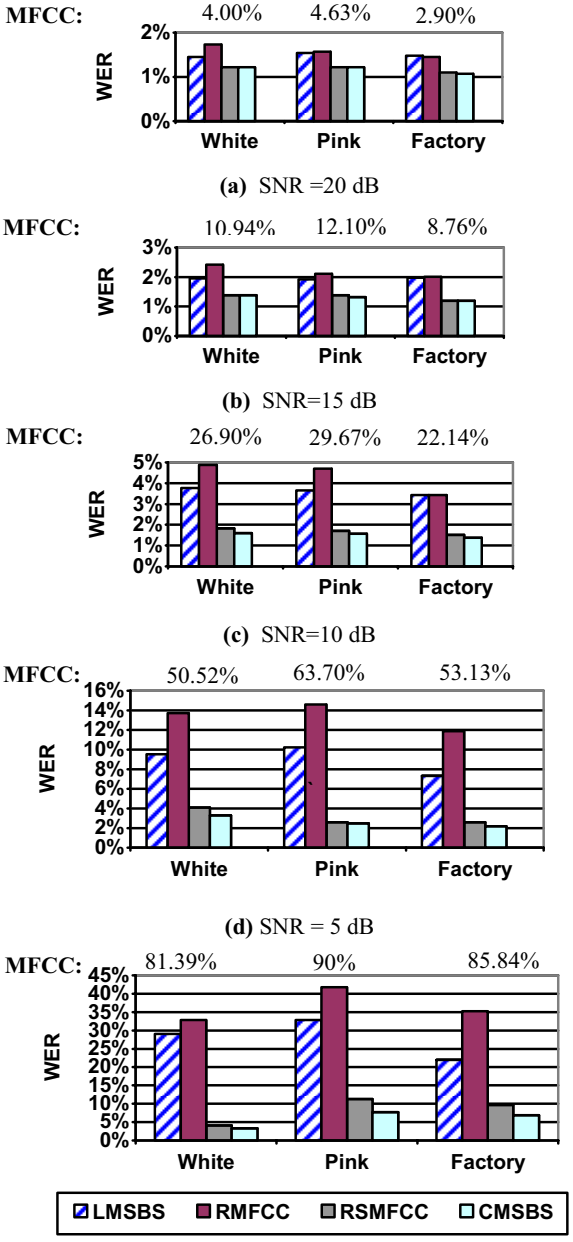

Fig. 3. Word error rate in presence of white, pink and factory noises for different SNR values

respectively, while the conventional MFCC has a word error rate higher than 80% for all three types of noise. After CMSBS, RSMFCC (Mel sub-band spectral subtraction followed by constant root) has the highest recognition result for all SNR values. RSMFCC is a special case of CMSBS that uses from constant root instead of SNR-dependent compression root. Furthermore, LMSBS method (Mel sub-band spectral subtraction followed by logarithm function) has lower word error rate in comparison to RMFCC method that only uses constant root without any noise subtraction.

6. Conclusion

We proposed a general framework for compensating of noise effects on MFCC features. This method included two steps. First, we applied a sub-band spectral subtraction to energies of Mel filter bank. After that, we used a sub-band SNR-dependent compression function instead of logarithm function in conventional MFCC features for more robustness to noise. Results show that the proposed method significantly decreases word error rates in presence of different additive noises with different SNR values. In SNR value of 0 dB, it decreases word error rates about 70% in comparing to conventional MFCC features. As future work, we plan to optimize our proposed compression function and use voice activity detectors to obtain a better estimation of noise in Mel sub-bands.

7. References

[1] Nasersharif, B., Akbari, A., "Sub-band weighted projection measure for robust sub-band speech recognition", *Proceeding of EUROSPEECH,* pp. 945-948, 2005.

[2] Choi, E.H.C, "A generalized framework for compensation of Mel-filter bank outputs in feature extraction for robust ASR", *Proceeding of EUROSPEECH,* pp. 933-936, 2005.

[3] Zhu, D., Nakamura, S., Paliwal, K.K., Wang, R, "Maximum likelihood sub-band adaptation for robust speech recognition", *Speech Communication,* Vol. 47, Iss. 3, pp. 243-264, November 2005.

[4] Cho, H.Y., Oh, Y.H., "On the use of channel-attentive MFCC for robust recognition of partially corrupted speech", *IEEE signal processing letters,* Vol.11, No. 6, pp. 581-584, June 2004.

[5] Chu, K.K., Leung, S.H., "SNR-dependent non-uniform spectral compression for noisy speech recognition", *IEEE International Conference on Acoustics, Speech, and Signal processing,* Vol. 1, pp. 973-976, 2004.

[6] Zhu, D., Paliwal, K., "Product of power spectrum and group delay function for speech recognition", *IEEE International Conference on Acoustics, Speech, and Signal processing,* Vol. 1, pp. 125-128, 2004.

[7] Ikbal, S., Misra, H., Bourlard, H., "Phase autocorrelation derived robust speech features", *IEEE International Conference on Acoustics, Speech, and Signal processing,* Vol. 2, pp. 133-136, 2003.

[8] Chen, J., Paliwal, K.K., Nakamura, S.," Sub-band based additive noise removal for robust speech recognition", *Proceeding of EUROSPEECH,* pp. 571-574, 2001.

[9] Hung W.W, Wang, H.C., "On the use of weighted filter bank analysis for the derivation of Robust MFCCs", *IEEE signal processing letters,* Vol.8, No. 3, pp. 70-73, March 2001.

[10] Martin, R., "Noise power spectral density estimation based on optimal smoothing and minimum statistics", *IEEE Transactions on Speech and Audio Processing,* Vol. 9, No. 5, pp. 504-512, July 2001.

[11] Huang, X., Acero, A., Hon, H., *Spoken Language processing*, Prentice Hall, 2001.

[12] Alexandre, P., Lockwood, P., " Root cepstral analysis: A unified view. Application to speech processing in car noise environments", *Speech Communication,* Vol. 12, Iss. 3, pp. 277-288, July 1993.