Mel Sub-Band Filtering and Compression for Robust Speech Recognition

Babak Nasersharif1, Ahmad Akbari1, Mohammad Mehdi Homayounpour2

1: Research center for Information Technology, Computer Engineering Department, Iran University of Science and Technology
2: Computer Engineering and IT department, Amirkabir University of Technology, 
{nasser_s, akbari}@iust.ac.ir, Homayoun@aut.ac.ir

Abstract

The Mel-frequency cepstral coefficients (MFCC) are commonly used in speech recognition systems. But, they are high sensitive to presence of external noise. In this paper, we propose a noise compensation method for Mel filter bank energies and so MFCC features. This compensation method is performed in two stages: Mel sub-band filtering 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 in place of logarithm function in conventional MFCC feature extraction in presence of additive noise. Results show that the proposed method significantly improves MFCC features performance in noisy conditions where it decreases average word error rate up to 30% for isolated word recognition on three test sets of Aurora 2 database.

Index terms: Mel sub-bands, Mel sub-band filter, SNR-dependent compression, MFCC

1. Introduction

Feature extraction is a crucial step of speech recognition process which greatly affects the performance of speech recognition systems. Speech features must represent the temporal evolution of the speech spectral envelope. Traditional speech features are extracted from power spectrum or amplitude spectrum of speech signal. When speech spectrum changes due to additive noise, spectral features alter considerably. This effect degrades the performance of speech recognition systems in presence of noise. Several techniques have been proposed to reduce sensitivity of features to external noise. A group of methods work at the spectral level. These methods reduce the effect of additive noise on the speech spectrum before extracting features from it. Spectral subtraction [1] and Wiener filtering [3] are well known examples of such methods.

Another group of robust speech recognition techniques work at feature level. In these methods, a transformation is directly applied to feature vectors to compensate noise effects on them. In some cases, the transformation is applied to cepstral domain such as cepstral mean normalization (CMN) [3] or SNR-dependent cepstral normalization (SDCN) [4]. In some other techniques, the transformation is applied to Mel filter bank energies (MFBE) or log of MFBE. Examples of such methods are: Mel Sub-band spectral subtraction [2,5], root cepstral analysis [2,6,7], and weighting of log for MFBE [8,9,10].

In this paper, we propose a transformation for applying to Mel sub-bands energies in order to remove noise from MFCC. While other methods only weight log energies of Mel sub-bands [8,9,10] or only perform noise filtering [5,7], we apply both of noise subtraction and SNR-dependent compression to Mel sub-band energies. In this manner, we utilize existing information about background noise and at the same time we give more importance to Mel sub-bands that are less affected by noise or distortion. In our proposed method, we first apply noise subtraction to Mel sub-band energies. Then, we introduce an SNR-dependent root function for compressing Mel sub-band energies rather than utilizing log function.

The rest of the paper is organized as follows. In Section 2, we describe our framework for removing noise effects from MFCC. Section 3, explains the used method for Mel sub-band filtering. 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. Compensation Framework

The conventional MFCC have poor performance in noisy conditions. To overcome this problem, we propose a framework to compensate additive noise effects on MFCC. So, we first discuss the general process of MFCC extraction from the speech signal.

Assuming a power spectrum $|X(k)|^2$ for a speech frame, the filter bank energy $E_i$ is calculated as:

$$E_i = \sum_{k} |X(k)|^2 \phi_i(k) \quad 1 \leq i \leq Q$$

(1)

where $\phi_i(k)$ is the $i$-th Mel triangular band-pass filter, $Q$ is number of Mel filters and $k$ is a frequency index. After this, a discrete cosine transform (DCT) is applied to log of filter bank energies. Thus, the Mel frequency cepstral coefficients for frame $x$ can be expressed as:

$$c_{m} = \sum_{i} \log(E_i) \cos\left(\frac{(i - 0.5)m \pi}{Q}\right) \quad 1 \leq m \leq M$$

(2)

where $M$ is desired number of MFCC features.

We define our noise compensation framework based on MFBE calculated in equation (1). Assuming that $|X(k)|^2$ is power spectrum of noisy speech, the proposed framework can be shown by following function:

$$E_i^x = F(E_i^x, w_i, b_i) = (E_i^x - b_i)^{w_i}$$

(3)

where $E_i^x$ is compensated Mel filter bank output and $w_i$ and $b_i$ are compensation parameters. The parameter $w_i$ is the compression factor with a value between 0 and 1. The bias $b_i$ depends on noise spectral characteristics. Equation (3) includes two steps: subtraction and energy compression. Using subtraction step, we reduce the filter bank energy increased due to presence of additive noise. In this step, a power estimation of noise in corresponding Mel sub-bands ($b_i$) is
needed. In Section 3, we give more details for estimation of \( b_i \). In the compression step, we use a sub-band SNR-dependent factor \( w_i \), less than 1 in order to put emphasis on Mel sub-bands that are less affected by noise and distortion. After these two steps, we calculate the compensated MFCC using:

\[
\hat{c}_i = \sum_{m} \left( E_i^m - b_i \right) \cos \left[ \frac{(i - 0.5) \pi m}{Q} \right]
\]

where \( \hat{c}_i \) shows compensated MFCC.

It can be seen from equation (4) that we replace log function with the proposed compression function. This function discriminates filter bank energies better than log function in presence of additive noise. The usefulness of a compression function in comparison to log function has been shown in root cepstral analysis [6, 7]. The compression process in human speech perception is well-known from the psychoacoustics point of view, where the sound intensity is converted to the perceived loudness [11].

2. Mel Sub-Band Filtering

The parameter \( b_i \) depends on energy of estimated noise and type of filter in \( i \)-th Mel sub-band. In order to estimate noise power spectrum at Mel-sub-bands, we estimate the noise power spectrum at the duration of 300 ms where only the noise is present. We use following smoothing equation for the noise power spectrum estimation [12]:

\[
| N(k) |^2 = P_i(k) = \lambda P_{i-1}(k) + (1 - \lambda) | B_i(k) |^2
\]

where \( P_i(k) \) and \( B_i(k) \) are estimated noise power spectra in previous \( i-1 \) frames and current frame, respectively. \( \lambda \) is a forgetting factor and \( k \) is the frequency index. The estimation of noise in \( i \)-th Mel sub-band is shown by \( \hat{E}_i \). \( \hat{E}_i \) is defined as the output of \( i \)-th triangular Mel filter, assuming that estimated noise \( |N(k)|^2 \) is passed through Mel filter bank. It can be computed as follows:

\[
\hat{E}_i = \sum_{m} N(k) \phi(k)
\]

We use sub-band filters that are reported to be more useful than full-band filters [5]. These filters are defined as in (7) for using in Mel sub-bands:

\[
H_{\varphi} = \sqrt{\frac{\operatorname{MAX} \left( (E_i^o)^{\varphi} - a_{\varphi} \beta_{\varphi} \right)}{(E_i^o)^{\varphi}}}
\]

where \( \varphi \) can be equal to 1 or 2. \( a_{\varphi} \) and \( \beta_{\varphi} \) are over-estimation factor and spectral flooring parameter in \( i \)-th Mel sub-band, respectively. \( E_i^o \) and \( E_i^\varphi \) are defined in equations (1) and (6).

Using equation (7), we can compute enhanced energy in \( i \)-th Mel sub-band as:

\[
E_i' = E_i^\varphi - b_i = E_i^o H_{\varphi}
\]

where \( E_i' \) is filtered sub-band energy by \( H_{\varphi} \), where \( \rho \) can be 1 or 2.

As a consequence, the parameter \( b_i \) in equations (3) and (4) can be computed as follows:

\[
b_i = E_i^o \cdot (1 - H_{\varphi})
\]

3. Mel Sub-Band Compression

MFCC computation is based on using logarithmic compression of Mel filter bank energies. An important effect of logarithmic compression of Mel filter bank energies is reduction of their dynamic range. This effect has two drawbacks in presence of additive noise. First, it can not highlight sub-bands energies that are less affected by noise. Second, some distortions that are insignificant in power spectrum domain become important after the logarithmic compression of Mel filter bank energies. DCT that is utilized in MFCC computation is a linear transform that gives equal weights to all compressed sub-band energies. Equal weighting of DCT and drawbacks of logarithmic compression make MFCC highly sensitive to additive noise. One solution to this problem is weighting of filter bank energies [8, 9, 10]. Another solution is root cepstral analysis [6,7] which substitutes log function with a constant root function and yields root Mel-frequency cepstral coefficients. The root Mel-frequency cepstral coefficients are computed as follows:

\[
\hat{c}_i = \sum_{m} \left( E_i^o \right)^{\gamma} \cos \left[ \frac{(i - 0.5) \pi m}{Q} \right]
\]

where \( \hat{c}_i \) denotes the root MFCC (RMFCC) features, \( \gamma \) is a constant between 0 and 1, and \( i \) is Mel filter index.

Although a constant root function performs better than log function in presence of noise, it doesn't take SNR of different sub-bands into account. We propose a compression function that is computed based on SNR in Mel sub-bands. This function replaces \( w_i \) in equations (3) and (4) and is formulated as:

\[
w_i = \gamma \left[ 1 - \exp \left( - \frac{SNR}{\xi_i} \right) \right] = \gamma \cdot G(SNR, \xi)
\]

where \( \gamma \) is a constant root and \( \xi_i \) is a parameter that controls the steepness of the compression function. \( G \) is an SNR-dependent function with values between 0 and 1. \( SNR_i \) is signal to noise ratio in \( i \)-th Mel frequency sub-band that can be estimated as in:

\[
SNR_i = \left( 1 + \frac{E_i'}{E_i^o} \right)^{\xi_i}
\]

where square root has been used for reducing the dynamic range of energy ratio. \( \rho \) is equal to 1 or 2 based on selected filter in equation (7). \( E_i' \) and \( E_i^o \) have been defined in equations (6) and (8).

Using equation (11), we need high compression at sub-bands with low \( SNR_i \) values and less compression or equalization at sub-bands with high \( SNR_i \) values. Therefore, the parameter \( \xi_i \) in equation (11) should be near to zero for high \( SNR_i \) values and near to one for low \( SNR_i \) values. So, we formulated \( \xi_i \) based on \( SNR_i \) as follows:

\[
\xi_i = \frac{1}{1 + \exp \left( - \frac{\xi \cdot SNR_i - \mu_{\xi \cdot SNR_i}}{\sigma_{\xi \cdot SNR_i}} \right)}
\]

where \( \mu_{\xi \cdot SNR_i} \) and \( \sigma_{\xi \cdot SNR_i} \) are mean and standard deviation of \( SNR_i \), computed from all Mel sub-bands of a speech frame. Function \( f \) was chosen as a sigmoid function, because a sigmoid function \( f \) satisfies asymptotic behavior of being \( f=0 \) at low \( SNR_i \) and \( f=1 \) at high \( SNR_i \). Moreover this function is monotonic and smooth enough.

Fig.1 shows values of \( \xi_i \) and \( G(SNR, \xi_i) \) versus \( SNR_i \) values in Mel sub-bands of a noisy speech frame where \( \rho \) is equal to 1. In this frame, \( SNR_i \) value varies between 1 and 3. It can be seen from the figure that when \( SNR_i \) is high, \( \xi_i \) has a small value. In such cases, as shown in the figure, the value of \( G \) in equation (11) is near to one and so \( w_i \) is very close to constant root \( \gamma \). In
this figure, when SNR is low, wi has a value near to 0.7. As shown in Fig. 1, this makes the value of G lower than one in equation (11). Consequently, in such cases, wi becomes a fraction of γ.

\[
G(SNR_i, w_i) = 0.7 \quad \text{when } SNR_i \text{ is low}
\]

Therefore, according to equation (11) and as shown in Fig. 1, the compression root wi changes in accordance with the sub-band SNR. When sub-band SNR is small, the compression root wi decreases and becomes less than γ. On the other hand, for large SNR values, compression root wi tends to the constant root γ as in root cepstral analysis.

5. Experiments and Results

We report our results on Aurora 2 database [13] for isolated word recognition. Only clean data are used for HMM training. Our recognizer is CDHMM with 16 states and 3 Gaussian mixtures per state. There are 8 types of noises in the 3 test sets: sets A,B and C [13]. Our feature vector in all cases (conventional or compensated form) contains 12 MFCC and 12 delta-MFCC and so its length is 24.

For evaluating our proposed compensation method, we have tested Mel sub-band filtering in company with the conventional log function. These features can be expressed by:

\[
se_c = \sum_{j} \log(E_j) \cos\left\{\frac{(i-0.5)m\pi}{Q}\right\} 
\]

(14)

where se denotes the obtained MFCC feature and \(E_j\) is as in equation (8). We denote these features by \(\text{LMSBF}\) which stands for Logarithm and Mel Sub-Band Filtering. Based on sub-band filtering type in equation (7), we use LMSBF1 for filtering by H1 and LMSBF2 for filtering by H2. LMSBF1 is the same feature set studied at [5] and our pervious work [2].

\(\text{CMSBF}\) which stands for Compression and Mel Sub-Band Filtering shows our proposed features that are described by equations (6) to (13). We use CMSBF1 and CMSBF2 for filtering by H1 and H2, respectively. We chose \(a_{i,1} = 1\) and \(\beta_{i,1} = 0.1\) in H1 for all Mel sub-bands. Furthermore, we selected \(a_{i,2} = 1\) and \(\beta_{i,2} = 0.01\) in H2. Additionally, we compared CMSBF with root Mel-frequency cepstral coefficients (RMFCC) where we choose the constant root equal to 0.5 in equation (10). Furthermore, we have performed Mel sub-band filtering together with constant root γ = 0.5 that can be shown by:

\[
se_c = \sum_{j} (E_j)^{\gamma} \cos\left\{\frac{(i-0.5)m\pi}{Q}\right\} 
\]

(15)

where src denotes the obtained MFCC feature and \(E_j^\gamma\) is as in equation (8). We use \(\text{RMFCC}\) for MFCC obtained from equation (15) that is Mel sub-band filtering followed with constant root. It can be seen that RMFCC is a special case of CMSBF where we use a constant compression root in place of SNR-dependent root. We use RMFCC1 and RMFCC2 for filtering using H1 and H2, respectively.

Fig. 2 shows average word error rate (AWER) for 3 test sets separately. The results are averaged over all SNR values (-5, 0, 5, 10, 15 and 20 dB) and noise types. We have shown the baseline MFCC and RMFCC results in top of figure in order to demonstrate our results more clearly. In this way, we can demonstrate our results more clearly. As can be seen in Fig. 2, AWER of RMFCC is higher than MFCC. On the other hand, AWER variation range for compensation methods in the figure is between 14% and 36% that is noticeable in comparison to MFCC and RMFCC results. By comparing between methods using H2 filter and methods using H1 filter, it can be seen that H2 (\(\rho=2\)) outperforms H1 (\(\rho=1\)). The CMSBF2 method has the lowest AWER among other methods. After CMSBF2, RMFCC2 (Mel sub-band filtering by H2 followed with constant root) has the highest recognition rate. As mentioned above, RMFCC is a special case of CMSBF. Furthermore, LMSBF2 method (Mel sub-band filtering by H2 followed with log function) has lower AWER in comparison to CMSBF1 and RMFCC1 methods. This shows the superiority of H2 filter to H1 filter. On the contrary, between those methods that use H1 filter, CMSBF1 and RMFCC1 have better results than LMSBF1.

\[
\begin{align*}
\text{MFCC} & \quad 45.58\% & \quad 40.53\% & \quad 50.35\% \\
\text{RMFCC} & \quad 56.86\% & \quad 57.42\% & \quad 57.49\% 
\end{align*}
\]

Fig. 2. Average word error rate over all SNR values and noise types separated for three test sets (A, B, C)

Fig. 3 shows average word errors rate over all SNR values which are separated for different types of noise and three test sets. It can be seen form the figure that for all types of noise, CMSBF2 has lowest AWER. Furthermore, after CMSBF2, RMFCC2 and LMSBF2 have the best recognition results. This shows that H2 performs better than H1. Additionally, for all types of noises, CMSBF1 has highest recognition rate among those methods using H1 filter.

Fig. 4 shows AWER over all noise type and test sets which are separated for SNR values. It can be seen from the figure that...
CMSBF2 has the lowest AWER among other methods for all SNR values.

6. Conclusion

We proposed a general framework for compensation of noise effects on MFCC feature vectors in speech recognition systems. In our proposed method, we had two steps. First, we applied a sub-band filter to energies of Mel sub-bands in order to reduce noise effects on them. Then, we compressed Mel sub-band energies using an SNR-dependent root function to overcome noise effects on MFCC feature vectors in speech recognition systems. We proposed a general framework for compensation of noise effects on MFCC feature vectors in speech recognition systems.

6. Conclusion

We proposed a general framework for compensation of noise effects on MFCC feature vectors in speech recognition systems. In our proposed method, we had two steps. First, we applied a sub-band filter to energies of Mel sub-bands in order to reduce noise effects on them. Then, we compressed Mel sub-band energies using an SNR-dependent root function to overcome drawbacks of DCT and log function in conventional MFCC. Results show that the proposed techniques significantly improve MFCC performance on Aurora 2 database. The proposed methods decreased average word error rates for isolated word recognition on 3 test sets of Aurora 2 database up to 30% in comparison to conventional MFCC. As our 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. Acknowledgement

The authors would like to thank the Iran Telecommunication Research Center (ITRC) for their support.

8. References