Journal of AI and Data Mining
Vol 3, No 2, 2015, 149-156. 10.5829/idosi.JAIDM.2015.03.02.04

Improving the performance of MFCC for Persian robust speech recognition

D. Darabian*, H. Marvi and M. Sharif Noughabi

Department of Electrical Engineering, University of Shahrood, Shahrood, Iran.

Received 15 May 2013; Accepted 27 June 2014
** Corresponding author:danial.darabian1@gmail.com (D. Darabian).*

Abstract
The Mel Frequency cepstral coefficients are the most widely used feature in speech recognition but they are very sensitive to noise. In this paper to achieve a satisfactorily performance in Automatic Speech Recognition (ASR) applications we introduce a noise robust new set of MFCC vector estimated through following steps. First, spectral mean normalization is a pre-processing which applies to the noisy original speech signal. The pre-emphasized original speech segmented into overlapping time frames, then it is windowed by a modified hamming window .Higher order autocorrelation coefficients are extracted. The next step is to eliminate the lower order of the autocorrelation coefficients. The consequence pass from FFT block and then power spectrum of output is calculated. A Gaussian shape filter bank is applied to the results. Logarithm and two compensator blocks form which one is mean subtraction and the other one are root block applied to the results and DCT transformation is the last step. We use MLP neural network to evaluate the performance of proposed MFCC method and to classify the results. Some speech recognition experiments for various tasks indicate that the proposed algorithm is more robust than traditional ones in noisy condition.

Keywords: *MFCC, Autocorrelation, Gaussian Filter Bank, Root, Mean Normalization*.

1. Introduction

Today speech technologies are commercially available for an unlimited range of tasks. The historical background of this technology indicates that the first speech recognition systems were built at Bell's lab in 1950. Improvement in ASR systems capabilities with respect to speech variability factors typically noise was at 1980 - 1990. Nevertheless, it is still a challenge to use ASR systems in real world environment because they are exposed to significant level of noise and it makes mismatch in training and testing conditions in real world applications. Recent research concentrates on developing ASR systems that would be much more robust against factors which make variability in the speech in real world environment.

The mismatch between training and testing condition can be reduced at several levels of ASR system's speech processing chain. Approaches against speech variability factors can be classified in three different groups: 1. Speech enhancement, 2. Speech model adaptation, 3. Robust feature extraction. In this paper, we concentrate on robust feature extraction typically the Mel-frequency cepstral coefficients (MFCCs).

2. Recent methods to improve MFCC

Block diagram of the standard MFCC which includes fundamental steps to derive MFCC from an original input speech shown in figure 1.

Various approaches have been proposed to improve the tolerance of an ASR system with respect to noise and a great deal of work has been done for robust feature extraction typically MFCC.

In some cases, which make significant changes in MFCC the autocorrelation coefficient was mentioned to improve MFCC algorithm in 1999 [1]. The idea was to use one-sided autocorrelation sequences of speech instead of original speech because autocorrelation of the noise in many cases could be considered relatively constant over time so a high pass filtering could lead to suppress the noise furthermore (RAS-MFCC).

The technique mentioned above was used again in 2006 called AMFCC [2]. Since the background noise corrupts the autocorrelation coefficients of the speech signal mostly at lower time lags while the higher-lag autocorrelation coefficients are least affected, this method uses only the higher-lag autocorrelation coefficients. Eliminating the lower order of the noisy speech signal autocorrelations coefficients should lead to removal of the main noise components .The maximum autocorrelation index to be removed is usually found experimentally [3].

Spectral differentiation was applied on the higher-lag autocorrelation coefficients algorithm in 2010 (DRHOASS-MFCC).

Another research was done over log compression in 2001. Results showed that root compression is better than logarithm compression for noise robustness (ROOT-MFCC) [4,5].

In another paper published in 2009, a Gaussian shape filter bank in place of triangular shaped bins was introduced (GMFCC) [6]. The objective was to make a higher amount of correlations between sub-bands outputs. It was shown that the inverted Mel-frequency cepstral coefficients is useful feature set for ASR systems which contain complementary information presented in high frequency region individually as well as in combination with the conventional triangular filter based (IMFCC & IGMFCC)[6].

Cepstral mean normalization and spectral mean normalization technique called SMN-CMN MFCC was another method [7,8].

MFCC standard algorithm was improved in the implementation aspects in 2012[9] because it has a large amount of computation and this is disadvantage in real time applications. An improved MFCC algorithm called MFCC-E was introduced that it reduced computations by 50% and made hardware implementations easy.

In [10] the AGC-MFCC has been used to improve MFCC algorithm.

Improvement in this algorithm is progressing rapidly and the development mentioned above was just only some limited cases. This paper is the complementary efforts, which follow previous work.

According to the recent methods mentioned above MFCC can be classified in three different groups:

1. Modifications in the standard blocks.
2. Modification includes adding some complementary blocks to the standard algorithm.
3. Modification includes reduce in hardware implementation.

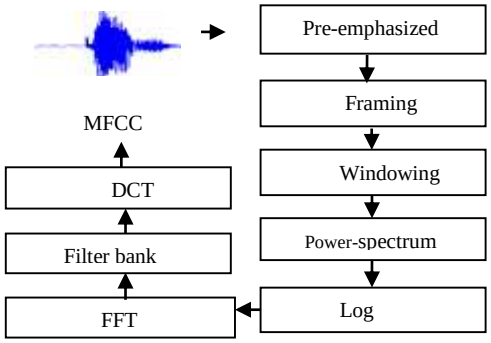


Figure 1. Standard MFCC algorithm.

In this paper, the aim is to improve MFCC algorithm with respect to adding complementary blocks and modification in the standard block.

In the next section, the proposed method is described.

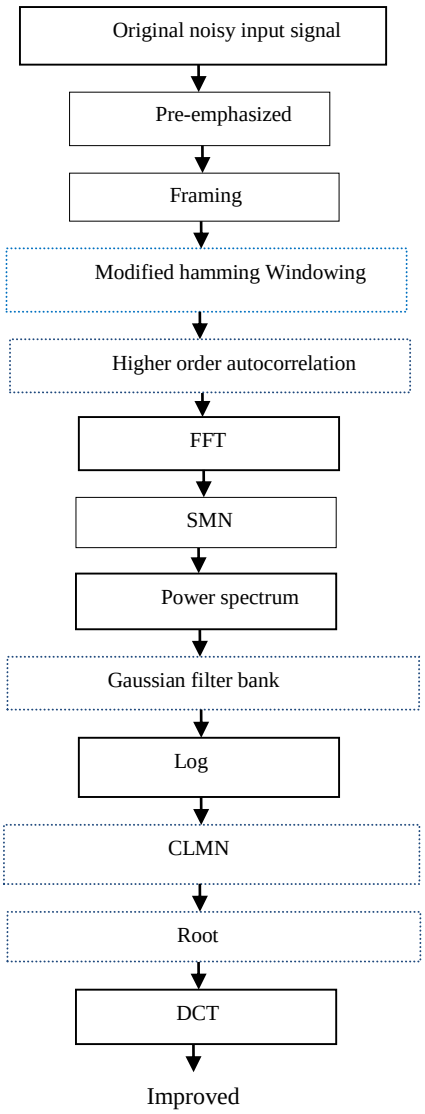


Figure 2. Proposed robust feature extraction algorithm (AGCR-MFCC).

3. Proposed method

This section describes our novel method to obtain new set of MFCC feature vector.

As mentioned in the recent methods section to improve MFCC algorithm, we introduce some methods used previously such as Gaussian filter banks, Modified hamming window, Higher order autocorrelation, Root method, Modified hamming window they are used separately in the standard algorithm without modifying other standard block but no one tried to combine all these advantages together but we try to do and to find out a way to combine last proposed methods: furthermore, we introduce new compensator blocks which they will improve recognition rate.

As illustrated in Figure 2 at the first step the input original noisy speech signal pass through pre-emphasized block using pre-emphasis filter in (1):

$$P(z) = 1-\alpha z^{-1} \quad (1)$$

Then frame blocking is performed and the modified hamming window is applied to the each frame.

3.1. Modified hamming window

In this paper, we use a family of hamming window, which is introduced in a paper in 2012[11].

If w (n) be a simple hamming window, our using window is in (2):

$$w_{new}(n) = n\, w\,(n) \quad (2)$$

The changes applied to the simple hamming window are in three different aspects:

1. Spectral leakage factor
2. Relative side lobe attenuation
3. Main lobe width

It can be observed that the spectral leakage increases and side lobe attenuation decreases to some extent which they have minor effect in recognition performance but considerably increase in main lobe width and will help to improve recognition performance. The changes in simple hamming window illustrated in figure 3.

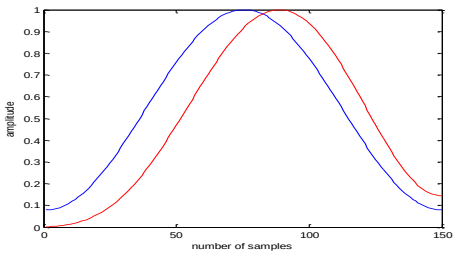


Figure 3. Hamming window (----) and modified hamming window (----).

3.2. Higher order autocorrelation

One-sided autocorrelation sequences of the framed signal passed from modified hamming window, which are obtained, and the lower lags of the autocorrelation sequences are removed [3]. It can further suppress the noise.

If d(m,k) is additive noise and s(m,k) is noise-free speech signal which m is number of frames and k is samples number then :

$$X(m,k) = s(m,k) + d(m,k) \quad (3)$$

If the noise is uncorrelated with the speech it follows that the autocorrelation of the noisy speech is the sum of autocorrelation of clean speech and autocorrelation of the noise:

$$R_{xx(m,k)}=R_{ss(m,k)}+R_{dd(m,k)} \quad (4)$$

If the additive noise is assumed to be stationary the autocorrelation sequences of noise can be considered to be identical for all frames and eliminating the lower order of the noisy speech signal autocorrelation coefficients should lead to removal of the main noise components. The maximum autocorrelation index to be removed is usually found experimentally which is selected in the following experiments section.

$$R_{xx(m,k)}=R_{ss(m,k)}+R_{dd(m)} \quad (5)$$

Then Fourier transform is calculated and power spectrum is found. Next step is SMN block which we use it to suppress the additive noise furthermore. Then we apply a Gaussian shape filter bank.

3.3. Gaussian shape filter bank

Triangular shape filter bank is used in the standard algorithm. A triangular shape filter bank is a symmetric tapered but does not provide any weight outside the sub bands that it covers (Figure 4). As a result, the correlation between a sub band and its nearby spectral component from adjacent sub bands is lost. It is proposed here a Gaussian shape filter bank[6] which provides gradually decaying weights at it's both ends for compensating possible loss of correlation the expression for GF can be written as:

$$\varphi_i = e^{\frac{-(k-kb_i)}{2\sigma_i^2}} \quad (6)$$

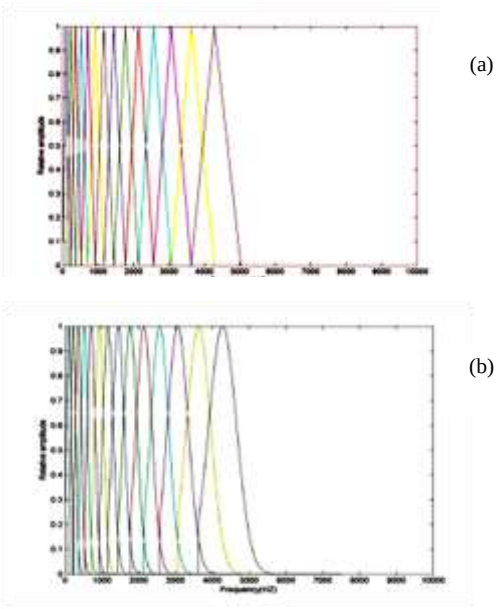
$$kbi=(i+1)\,.\,\Delta_{mel} \quad (7)$$

where, in (6) and (7) sigma is variance of any sub bands and kb is boundary points in triangular filter bank derived from equations below (i, is the number of Gaussian):

$$\Delta_{mel}= \frac{f_{max(mel)}}{i+1} \quad (8)$$

In (8) f_{max} is maximum sampling frequency rate and it is calculated in Mel-Scale through (9):

$$f_{mel}=2595 \log (1+\frac{f}{700}) \quad (9)$$



(a)

(b)

Figure 4. a: Triangular filter bank , b: Gaussian filter.

3.3. CLMN and root blocks

The proposed algorithm uses spectral mean normalization to suppress the additive noise and uses cepstral log mean normalization after logarithm to remove the effect of convolution noise. Combination of CLMN and SMN can inhibit additive and convolution noise at the same time. In this paper SMN block applies after FFT block and CLMN applies after logarithm function to compensate vulnerability of logarithm to convolution noise we name that CLMN (cepstral logarithm mean normalization).

The calculations of SMN and CLMN are based on this fact that expectation of noisy part is constant so it can be removed in CLMN and SMN process which is shown in equations below:

$$X(m,k) = S(m,k) + d(m,k) \quad (10)$$

$$\hat{\boldsymbol{x}}(m, k) = x(m, k)-E[x(m, k)]$$

$$= \{s(m, k)+d(k)\}-\{E[s(m, k)+d(k)]\} \quad (11)$$

$$= s(m, k)-E[s(m, k)] = \hat{\mathbf{s}}(m, k)$$

Logarithm function in the MFCC generation is very sensitive to noise and is one reason for poor noise performance of MFCC. After logarithm function CLMN is used. The root compression block is the next block in our proposed algorithm due to generating values close to zero after CLMN [4,5]. The log function gives large negative values for input close to zero and this leads to spreading of the energy. CLMN doesn't change these values and its task is just to suppress convolution noise and they are still close to zero (furthermore CLMN makes data more close to zero). So root compression is used and followed by DCT leads to better compaction of the energy.

The large negative excursion of CLMN outputs for values close to zero leads to a splattering of energy whereas root compression, which express as $(.)^{\alpha}$ with $0 <\alpha< 1$ leads to better compaction of energy. Algorithm uses root block after CLMN to achieve this aim. The application of CLMN is defined in the following equations:

$$X(m,k) = s(m,k) * d(k) \quad (12)$$

$$X(m,k) = s(m.k) . d(k) \quad (13)$$

$$LogX(m,k)=\log(S(m,k).H(k))=\log S(m,k)+\log H(k) \quad (14)$$

$$\text{Log } X(m,k)-E(\text{Log } X(m,k)) =$$

$$\log S(m,k)+\log H(k)–E(\log S(m,k))-\log H(k)= \quad (15)$$

$$\log S(m,k) –E(\log S(m,k))$$

In (13) the original signal is under convolution noise then the FFT applied and (14) is resulted then logarithm performance make the conversion of multiplying to the adding and expectation function suppress the noise according to (15).

We call our proposed method as AGCR-MFCC which A stands for "Autocorrelation" G stands for "Gaussian shape filter bank" and C stands for "CLMN" and R stands for "Root".

4. Experimental setup

In order to evaluate the performance of proposed algorithm and to classify, we use MLP neural network with one input layer, two hidden layer and one output layer. We experiment some other hidden layer values but the results show that it has the best results. Number of neurons in the two hidden layer can be chosen by a user in the MATLAB code. We spot them both 50 because at this value network has the best response.

60 words which are chosen through 10 different speakers with 15 repetition in each word have been chosen so we have 60 classes (and so 60 output neurons), and 900 words.

70% of the entire data (630 words) is used for training and 30% (270 words) is used for testing.

The proposed approach was implemented on Farsdat speech data base. Frame length is appropriate to speech length but the number of frames is constant 60 and length of window is 50ms and sampling frequency is 22000.

To obtain the noisy speeches the clean speech corrupted by artificial white Gaussian noise (WGN) in four different signals to noise ratio (SNR) levels. Silence speech parts are removed

using a general silence detection technique. Figure 5 illustrates a general form of MLP neural network, which is used to classify in this paper.

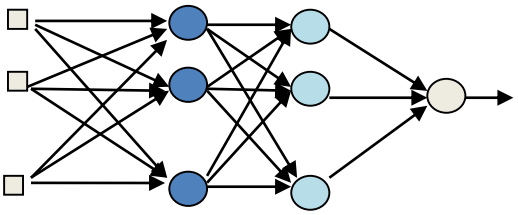


Figure 5. Neural network with two hidden layer [12].

As mentioned the Data base is divided into training set and testing set. Features vector sets of size 14 are extracted using different family of MFCC: standard MFCC, RAS-MFCC, AMFCC, ROOT-MFCC, GMFCC, AGMFCC and AGCR-MFCC (proposed method) and their performances are compared. As describe above, adding the artificial Gaussian noise at four SNR levels generate the polluted testing utterances. Using a random number generation program generates the white noise.

4.1. Experimental results

As described in the section 3.2, the maximum autocorrelation index to be removed is usually found experimentally, Table 1 shows experiment results which lead to selecting the best index to be removed .

Table 1. Various index was experimented to select the best and appropriate index (T: threshold) to be removed at autocorrelation segment in AGCR-MFCC. Experiment results show that the highest noisy average recognition ratio belong to T=100.

Index	20dB	10dB	5dB	0dB	Noisy average recognition ratio
T=10	77	72	65.6	55	67.4
T=30	75.3	73.2	67.5	64.6	70.15
T=50	73.5	73	72.5	40	64.75
T=70	77.3	77	61.6	55.5	67.85
T=100	83.5	80.8	77.7	76.91	78.91

In the Table 1 variable T (threshold) is the index whose experiments are performed on it and the results show that when T=100 is selected the best speech recognition occurred.
Figure 6 and Figure 7 shows a comparative results to select the best index to be removed as it is shown in the T=100 the best speech recognition rate is achieved. The process of experiments is explained at experimental setup and the other details are explained in the following section.

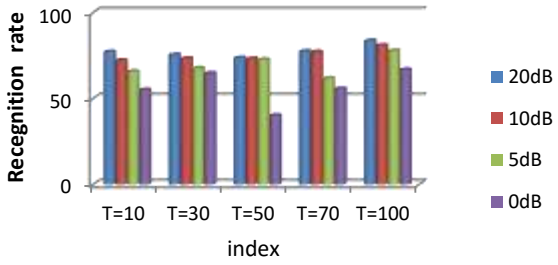


Figure 6. A comparative results to select the best index to be removed as it is shown T=100 has the best recognition rate.

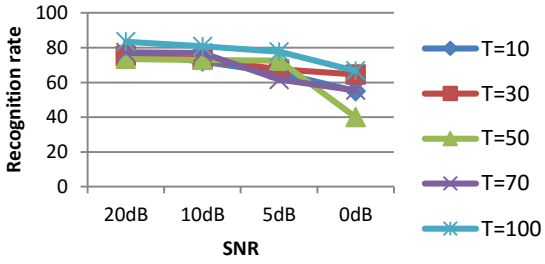


Figure 7. Results which depict that T=100 has the best speech recognition rate.

In order to use the root compression block in the modified algorithm it should be determined variable α.

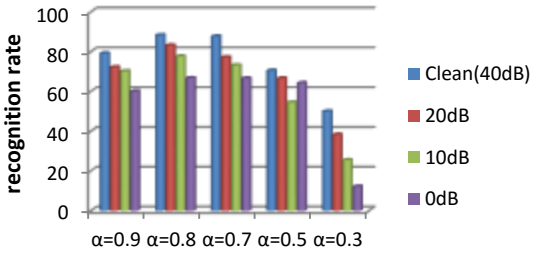


Figure 8. Various α values was experimented and α=0.8 was selected because of better recognition rate in some certain SNR.

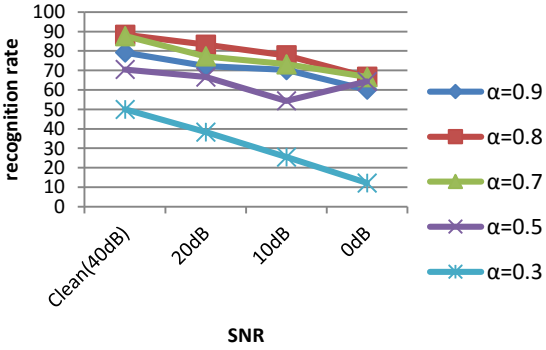


Figure 9. Results which depict that α=0.8 has the best speech recognition rate.

We study some various α rates and choose the best one in noisy condition. We tried various α rates .Some experiments were done to select the best value of α for the best speech recognition application. The results of corresponding experiments α sets 0.8. Table 2 includes the experiment results to select the best root value to use after CLMN block.

Table 2. The experiments to select the best root for using after CLMN block was performed. Results show that α=0.8 yields better speech recognition accuracy in noisy condition. The highest noisy average recognition ratio occurred in α=0.8.

α values	Clean (40dB)	20dB	10dB	0dB	Noisy average recognition ratio
0.9	79.2	72.2	70.2	60.1	67.5
0.8	88.3	83.2	77.7	76.91	78.91
0.7	87.7	77.2	73.2	66.6	72.3
0.5	70.5	66.6	54.4	64.3	61.7
0.3	50	38.3	25.5	12.2	25.3

Results show that α=0.8 yields better speech recognition accuracy in noisy condition. The highest noisy average recognition ratio occurred in α=0.8. Figure 8 and 9 indicate that α=0.8 is an appropriate value in our Farsi speech recognition experiments. Then the general experiments performed to evaluate the performance of our novel method to obtain a new set of MFCC feature vectors with these determined values.

We compare the performance of MFCC, AMFCC, GMFCC, ROOT MFCC, CMN-SMN MFCC, AGMFCC, and AGCR-MFCC (proposed method) when training data and testing data are in clean (40dB) environment and after adding artificial noise at 4 SNR levels. The noises are added to the clean speech signal at 20,10,5 and 0dB SNRs table 3 indicates the results obtained using MFCC, AMFCC, GMFCC, ROOT MFCC, CMN-SMN MFCC, AGMFCC, AGCR-MFCC (proposed method) front-ends. For the case of speech sounds corrupted by white noise shown in Figure 10 and table 3 the performance of MFCC degrade most significantly among all features in presence of the noise and it was found to be worse among other robust features. Evidence depicts that the performance of MFCC degrades significantly compared with other feature vectors when added noise increases. It is due to standard MFCC is sensitive to noise and it was not an unexpected result whereas in the clean environment standard MFCC has still the best application than other suggested methods. Figure 10 shows a remarkable improvement especially in noisy condition (5dB, 0dB) for our proposed method. The best performance comes from AGCR-MFCC with improvement in recognition score of %3.3 at 20dB %7.2 at 10dB 17.6% at 5dB and 27.41 % at 0dB in comparison with standard MFCC due to variations applied to the standard algorithm which makes it robust to noise such as including SMN block, Gaussian shape filter instead of triangular shape filter, autocorrelation and removing the lower orders. CLMN and ROOT compression block to compensate logarithm function but in clean condition the standard algorithm has still the best results and this is obvious because we know that standard MFCC feature has no problem in clean condition and its application degrade in the noisy condition and our proposed method has been organized to overcome this problem therefore we don't expect our proposed algorithm be better in clean condition. Our proposed algorithm running duration is more than standard algorithm but it is ignored. In the standard algorithm, the average of processing time is less than 1 minute but in our proposed one is less than 1.30 minutes to extract features.

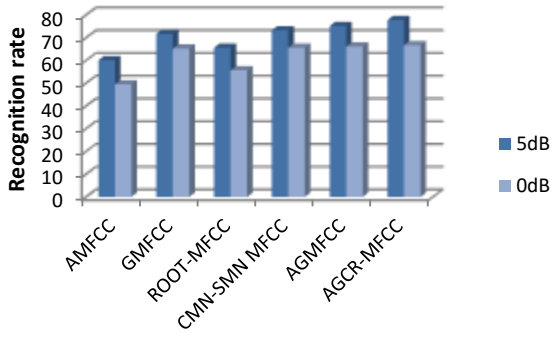


Figure 10. This figure shows that the results of experiments in the two noisy condition and as it is shown AGCR-MFCC has better recognition rate at noisy condition in comparative with other extracting MFCCs methods.

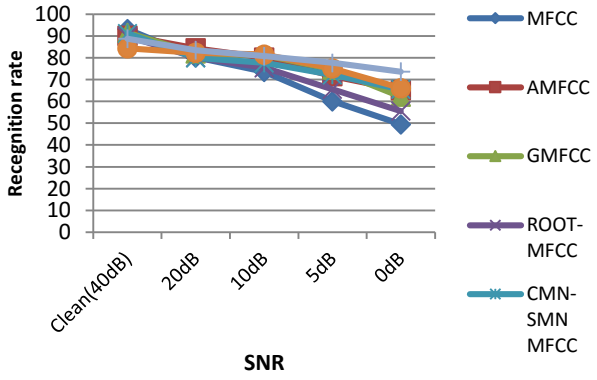


Figure 11. AGCR-MFCC has the best speech recognition especially in noisy condition such as 5dB and 0Db which the results are clearer in these SNRs.

Table 3. The entire training and testing data was used for experiments. Results for AGCR-MFCC show improvement in comparison with other traditional method in recognition rate when artificial white Gaussian additive noise increase. Noisy average recognition ratio proves that proposed method is more robust to noise than traditional methods.

SNR/feature	MFCC	AMFCC	GMFCC	ROOT-MFCC	CMN-SMN MFCC	AGMFCC	AGCR-MFCC (proposed method)
Clean(40dB)	93.2	90.2	91.5	89.9	91.2	84.4	88.8
20dB	80.2	84.4	82	80.1	80	82.3	83.5
10dB	73.6	80.1	81.3	75.5	77.7	81.5	80.8
5dB	60.1	71.6	74.6	65.55	72.2	75.09	77.7
0dB	49.5	65.2	62.2	55.6	65.6	66.08	76.91
Noisy average recognition ratio	65.85	75.32	75.02	69.18	73.87	76.24	78.91

5. Conclusion and future work

This paper modified one of the most common features for robust speech recognition application to improve ASR accuracy under noisy condition.

To evaluate the experiments, we use the MLP neural network for classification. In proposed method triangular, filter bank has been replaced by Gaussian shape filter bank then to compensate the undesirable effect of the noise and we use the CLMN and root compression blocks. Spectral mean normalization (SMN), Autocorrelation and eliminating the lower Order was other works which all made improve the noise-robustness of MFCC standard blocks.

Although these variations make computational costs because we will have more multiplying and adding computations typically in the autocorrelation, eliminating the lower order and Root block, Consequently more hardware logical gate is needed in hardware implementation but we pay these costs and certainly it is reasonable because the powerful application of MFCC algorithm is undeniable and as we know the standard algorithm degrade in presence of Noise drastically. If we Pay the computational and implementations costs, we can impart this feature even in presence of noise and we can keep it as a powerful feature in the future works [13,14].

Our research improvement model contains complementary blocks and modifications in the standard blocks were performed but there are still some blocks which were not examined and the question which has been still remained is that: is there any better replacement blocks for them?

Future works would involve these examinations. Further studding about hardware implementation which is an important necessity should be conducted.

References

[1] You, K. H. & wang, H. C. (1999). Robust features for noisy speech recognition based on temporal trajectory filtering of short time autocorrelation sequences. Speech communication, vol. 28, pp. 13-24.

[2] Shannon, B. J. & Paliwal, K. K. (2006). Feature extraction from higher lag autocorrelation coefficient for robust speech recognition, vol. 48, no. 11, pp. 1458-1481.

[3] Devand, A. & bansal, P. (2010). Robust feature extraction for noisy speech recognition from magnitude spectrum of higher order autocorrelation coefficients. International journal of computer application(0975-8887), vol. 10, no. 8.

[4] Lim, J. S. (1979). Spectral root holomorphic deconvolution system, IEEE Trans. ASSP, vol. 27, no. 3, pp. 223–233.

[5] Alexandre, P. & Lockwood, P. (1993). Root cepstral analysis: A unified view. Speech Communication, vol. 12, no. 3, pp 277–288.

[6] Chakroborty, S. & Saha, G. (2009). Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter. International Journal of Signal Processing, vol. 5, no. 1, pp. 11-19.

[7] Sarikaya, R. & John, H. L. H. (2001). Analysis of the root-cepstrum for acoustic modeling and fast decoding in speech recognition. In: Euro speech, Aalborg, Denmark.

[8] Liu, F. H., Acero, A. & Stern, R. (1992). Efficient joint compensation of speech for the effects of additive noise and linear filtering. Proc. Of IEEE ICASP, vol. 1, pp. 257–260.

[9] Xie, C., Cao, X. & Lingling, H. (2012). Algorithm of Abnormal Audio Recognition Based on Improved MFCC. International workshop on information and electronic engineering (IWIEE), pp 731-737.

[10] Marvi, H., Darabian, D. & Sharif Noughabi, M. (2013). Robust speech recognition using modified MFCC and neural network. 11th ICIS conference Tehran.

[11] Sahidullah, M. D. & Saha, G. (2012). A novel windowing technique for efficient computation of MFCC for speaker recognition. Signal Processing Letters, IEEE, vol. 20, no. 2, pp. 149 – 152.

[12] Xiong, X. PhD thesis. (2009). Robust Speech Features and acoustic Models for Speech Recognition. Computer engineering Department, Nanina's Technological University.

[13] Zunjing, W. & Zhigang, C. (2005). Improved MFCC-Based Feature for Robust Speaker Identification. Identification Tsinghua Science and Technology ISSN, vol. 10, no. 2, pp.158-161.

نشریه هوش مصنوعی و داده کاوی

دانشگاه شاهرود

بهبود عملکرد ضرایب مل-کپستروم در تشخیص گفتار فارسی

دانیال دارابیان*، حسین مروی و مجتبی شریف نوقابی

دانشکده مهندسی برق، دانشگاه شاهرود، شاهرود، ایران.

ارسال ۲۰۱۳/۰۵/۱۵؛ پذیرش ۲۰۱۴/۰۶/۲۷

چکیده:

ضرایب مل-کپستروم یکی از رایج‌ترین ویژگی‌ها در سیستم‌های تشخیص گفتار می‌باشند. این ضرایب در عین قدرت بالا در به‌کارگیری در سیستم‌های تشخیص گفتار، بسیار به نویز حساس هستند. در این مقاله ما یک روش مقاوم به نویز برای استخراج این ضرایب پیشنهاد نموده‌ایم که شامل استفاده از چند بلوک جبران‌گر، همچنین تغییر در چند بلوک در الگوریتم پایه می‌باشد. در بخش آزمایش روش پیشنهادی از شبکه عصبی استفاده شده است. آزمایش‌های تشخیص گفتار صورت گرفته نشان‌دهنده بهبود نرخ تشخیص گفتار در محیط نویزی، نسبت به سایر روش‌ها در استخراج این الگوریتم هستند.

کلمات کلیدی: ضرایب مل-کپستروم، خودهمبستگی، فیلتربانک گوسی، ریشه‌یابی، تفریق از میانگین.