International Journal of Soft Computing 8 (1): 56-62, 2013
ISSN: 1816-9503

An Adaptively Enhanced Auditory Transform Based Feature Extraction Algorithm for Robust Speaker Identification

[1]S.D. Umarani, [2]R.S.D. Wahidabanu and [3]P. Raviram
[1]Department of ECE, [2]Government College of Engineering, 636011 Salem, India
[3]Department of CSE, Mahendra Engineering College, 637503 Tiruchengode, India

Abstract: In speech recognition systems, obtaining good performance in noisy environments still remains a very challenging task. The problem is that recognition accuracy degrades significantly if training conditions are not matched to the corresponding test conditions. This study uses auditory transform along with CFCC (Cochlear Filter Cepstral Coefficients). Usually, the performance of acoustic models trained in clean speech drops significantly when tested in noisy speech. The CFCC features have shown strong robustness in this kind of situation. The auditory transform replaces the STFT in CFCC for overcoming the STFT's disadvantage of fixed time-frequency resolution. Thus, a kind of good anti-noisy speech feature coefficient was obtained. In order to enhance the ability to resist the noises of different environments, an adaptive enhancement approach is introduced. The CFCC features with wavelet are applied to a speaker identification task to address the acoustic mismatch problem between training and testing environments. Finally, this experimental results show that noise resilience of the proposed method under small samples circumstance is better than exiting methods at least by 3 dB in worsetcase for lesser word count and at least 1 dB for larger word count. It is observed that CFCC feature with adaptive enhancement can remain better robust to noise. And its performance is more effective under low SNRs. It makes the speech recognition become possible under these conditions.

Key words: Robust speech recognition, auditory transform, adaptive enhancement, cochlear, filter bank, CFCC

INTRODUCTION

In recent decades, speech recognition systems have significantly improved. Nevertheless, obtaining good performance in noisy environments still remains a very challenging task. The problem is that recognition accuracy degrades significantly if training conditions are not matched to the corresponding test conditions. These environmental differences might be due to speaker differences, channel distortion, reverberation, additive noise or other causes. Usually, speech recognition includes three parts: pre-processing, feature extraction and recognition. It consists of filter bank, feature extraction and training (recognition) network. The function of filter bank is dividing speech signal into different frequency band to be good for extraction feature. The good feature can improve the system recognition rate. The training (recognition) network trains (recognizes) the feature vectors according to feature mode and outputs recognition results. The research on noise-robust capability of speech recognition system is a difficult problem that has been limiting the practical application of the speech recognition system (Yao *et al*., 2001).

Because human ear has strong noise-robust capability, it is very important to abstract the features of fitting auditory characters of human ear for improving system noise-robust performance. The bark wavelet overcomes the disadvantage that the common wavelet divides frequency band in octave band and it is more suitable to the auditory characters of human ear. Bark wavelet is a warping wavelet that divides frequency band according to critical band (Fu and Yi, 2000). At the same time, MFCC (Mel Frequency Cepstrum Coefficients) (Lawrence and Rabiner, 1999) and ZCPA (Zero-Crossing with Peak Amplitude) (Kim *et al*., 1999) features themselves have noise-robust performance. But accompany with deterioration of the environmental noisy condition as well as the increase of the speech vocabulary, the MFCC Method's speech recognition capabilities decrease dramatically, showing the unsuitability of the usage in the strong noisy conditions and large vocabulary recognitions. The study introduces a kind of wavelet whose base function obeys the understandability time-frequency optimal uncertainty but the scale function varies according to the critical band. This makes the frequency band in each scale into a frequency group. Researchers name the wavelet as bark wavelet. Associating the Bark wavelet with the MFCC, the

Corresponding Author: S.D. Umarani, Department of ECE, Government College of Engineering, 636011 Salem, India

study presents a new kind of feature coefficients. The experimental results of speech recognition demonstrate that this new feature is more robust than the MFCC feature in noise environment and large vocabulary.

Preliminaries: The most widely used forms of feature extraction are Mel Frequency Cepstral Coefficient (MFCC) and Perceptual Linear Prediction (PLP) (Hermansky, 1990). Figure 1 contains block diagrams of MFCC and PLP. MFCC processing begins with pre-emphasis, typically using a first-order high-pass filter. Short-Time Fourier Transform (STFT) analysis is performed using a hamming window and triangular frequency integration is performed for spectral analysis. The logarithmic nonlinearity stage follows and the final features are obtained through the use of a Discrete Cosine Transform (DCT).

PLP processing which is similar to MFCC processing in some ways, begins with STFT analysis followed by critical-band integration using trapezoidal frequency-weighting functions. In contrast to MFCC, pre-emphasis is performed based on an equal-loudness curve after frequency integration. The nonlinearity in PLP is based on the power-law nonlinearity proposed by Stevens (Hermansky, 1990). After this stage, Inverse Fast Fourier Transform (IFFT) and Linear Prediction (LP) analysis are performed in sequence. Cepstral recursion is also usually performed to obtain the final features from the LP coefficients (Gold *et al*., 2011).

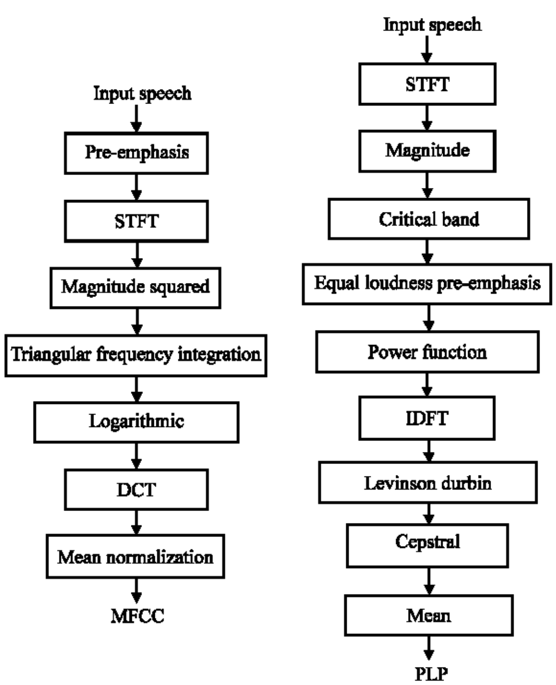

Fig. 1: Block diagrams of MFCC and PLP processing

At a high level, most speech feature extraction methods fall into the following two categories: modeling the human voice production system or modeling the peripheral auditory system. For the first approach, one of the most popular features is a group of cepstral coefficients derived from linear prediction known as the Linear Prediction Cepstral Coefficients (LPCC) (Hermansky and Morgan, 1994; Moore and Glasberg, 1996). The LPCC feature extraction utilizes an all-pole filter to model the human vocal tract with speech formants captured by the poles of the all-pole filter. The narrow band (e.g., up to 4 kHz) LPCC features research well in a clean environment. However, in the earlier experiments the linear predictive spectral envelope shows large spectral distortion in noisy environments (Atal, 1974; Hung *et al*., 2001). This results in significant performance degradation.

For the second approach, there are two groups of features, based on either Fourier Transforms (FT) or auditory-based transforms. Representative for the first group are the MFCCs (Mel Frequency Cepstral Coefficients) where a Fast Fourier Transform (FFT) is applied to generate the spectrum in the linear scale and then a bank of band-pass filters is placed along a Mel frequency scale on top of the FFT output (Pujol *et al*., 2006). Alternatively, the FFT output is warped to a Mel or Bark scale and then a bank of band-pass filters is placed linearly on top of the warped FFT output (Atal, 1974; Hung *et al*., 2001). The proposed algorithm in this study belongs to the second group where the auditory-based transform is defined as an invertible, time-frequency transform. The output from this kind of transform can be in any kind of frequency scale (e.g., linear, Bark, ERB, etc.). Therefore, there is no need to place the band-pass filter in a Mel scale as in the MFCC or warp the frequency distributions as in (Atal, 1974; Hung *et al*., 2001).

The MFCC features (Pujol *et al*., 2006) in the first group are one of the most popular features for speech and speaker recognition. Like the LPCC features, the MFCC features perform well in clean environments but not in adverse environments or mismatched training and testing conditions. Perceptual Linear Predictive (PLP) analysis is another peripheral auditory-based approach.

The auditory transform: The auditory transform is the forward transform of a pair of invertible auditory-based transforms as defined and described by Li (2009). It can be implemented as a filter bank. As the foundation of the auditory-based feature extraction algorithm, the forward auditory transform is used to replace the Fast Fourier transform. The auditory transform models the traveling wave in the cochlea where the sound waveform is decomposed into a set of sub band signals.

Let f(t) be a speech signal. A transform of f(t) with respect to a cochlear filter ψ(t), representing the Basilar Membrane (BM) impulse response in the cochlea is defined as:

$$T(a,b) = f(t) \times \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right)dt \qquad (1)$$

where, * denotes the convolution operation, a and b are real, both f(t) and ψ(t) belong to $L^2(R)$ and T(a, b) representing the traveling waves in the BM is the decomposed signal and filter output. The above equation can also be written as:

$$T(a,b) = f(t) \times \psi_{a,b}(t)dt \qquad (2)$$

Where:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t\text{-}b}{a}\right) \qquad (3)$$

Like in the wavelet transform, the factor a is a scale or dilation variable. By changing a, researchers can shift the central frequency of ψ to receive a band of decomposed signals. Factor b is a time shift or translation variable. For a given value of a, factor b shifts the function $\psi_{a,0}(t)$ by an amount b along the time axis.

Note that $1/\sqrt{|a|}$ is an energy normalizing factor. It ensures that the energy stays the same for all a and b therefore, researchers have:

$$\int_{-\infty}^{\infty}\left|\psi_{a,b}(t)\right|^2 dt = \int_{-\infty}^{\infty}\left|\psi(t)\right|^2 dt \qquad (4)$$

The cochlear filter as the most important part of the transform is defined as:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t\text{-}b}{a}\right) \qquad (5)$$

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|\alpha|}}\psi\left(\frac{t\text{-}b}{\alpha}\right)^{a}\exp\left[-2\pi f_L\beta\left(\frac{t\text{-}b}{\alpha}\right)\right] \cos\left[2\pi f_L\left(\frac{t-b}{\alpha}\right)+\theta\right]u(t\text{-}b) \qquad (6)$$

where, α>0 and β>0, u(t) is the unit step function, i.e., u(t) = 1 for t≥0 and 0 otherwise. Parameters α and β determine the shape and width of the cochlear filter in the frequency domain. The value of θ should be selected such that the following condition is satisfied:

$$\int_{-\infty}^{\infty}\psi(t)\,dt = 0 \qquad (7)$$

This is required by the transform theory to ensure no information is lost during the transform (2009). The value of a can be determined by the current filter central frequency, f_c and the lowest central frequency, f_L in the cochlear filter bank:

$$a = \frac{f_L}{f_c} \qquad (8)$$

Since, researchers contract $\psi_{a,b}(t)$ with the lowest frequency along the time axis, the value of a is in the range 0<a≤1. If researchers stretch ψ, the value of a would be constrained to a>1. The frequency distribution of the cochlear filter can be in the form of linear or non-linear scales such as ERB (equivalent rectangular bandwidth) (Moore and Glasberg, 1983), Bark (Zwicker and Terhardt, 1980), Mel scale (Davis and Mermelstein, 1980) or log. For a particular band number i, the corresponding value of a is represented as a_i which needs to be pre-calculated for the required central frequency of the cochlear filters at band number i. The value of β controls the filter band width, i.e., the Q-factor. This makes the Auditory Transform (AT) different than the Gammatone function (Johannesma, 1972) in which the Q-factor is fixed.

Researchers note that the inverse of the above transform exists. It has also been proven mathematically and validated experimentally (Li, 2009). This property ensures that the forward transform implemented by the cochlear filter bank can avoid any information loss and thus qualifies as a platform for feature extraction.

Following is the comparative analysis of the Auditory Transform (AT) and the well-known Fourier Transform (FT) with the comparison to the features derived from the AT such as the CFCCs and from the FT such as the MFCCs. It is observed that there are no pitch harmonics and there is less computational noise in the spectrums generated from the auditory transform. In addition, all formant information has been kept. This is due to the variable length of cochlear filters and the selection of parameter β in Eq. 5. The harmonics in FFT spectrogram are due to the fixed window length for all frequency bands. For robust speaker identification, researchers do need a more robust time-frequency transform as the foundation for feature extraction. The transform should generate less distortion from background noise and less computation noise from selected algorithms such as pitch harmonics while also retaining the useful information. Here, the auditory transform provides a robust solution to replace the Fourier transform.

Since, the MFCC features are popular features in both speaker and speech recognition, the proposed CFCCs is compared with the MFCCs as follows: it is understood that the MFCC features use the FFT to convert the time domain speech signal to the frequency domain spectrum. The power spectrum is calculated and then triangle filters are applied to produce filter bank energy estimates. The triangle filters are distributed in the Mel scale. In contrast, the proposed CFCC features use a bank of cochlear filters to decompose the speech signal into multiple bands. The frequency response of a cochlear filter has a bell-like shape rather than a triangle shape. The shape and width (the Q-factor) of the filter in the frequency domain can be adjusted by parameters α and β from Eq. 5. In each of the bands, the decomposed signal is still in the time domain, represented by real numbers. The central frequencies of the cochlear filters can be arranged in any distribution including Mel, ERB, Bark or log.

When using the FFT to compute a spectrogram, the window size must be fixed to all frequency bands due to the fixed point FFT. When researchers compute a spectrogram from the decomposed signals generated by the cochlear filters, the window size can be different for different frequency bands. For example, a longer window is used for a lower frequency band to average out the background noise and a shorter window is for a higher frequency band to protect high-frequency information. Furthermore, the MFCCs use a logarithm as the nonlinearity while the CFCCs use a cubic root.

In this research, the Speech Separation Challenge (SSC) database is used (Cooke and Lee, 2006) to report results because the database has several mismatched conditions. Also, this allows us to compare the obtained results with other reported results on the same database. Note that the training set consists of only clean speech while both the development set and the testing set consist of clean speech and noisy speech at five different SNR levels.

Adaptive wavelet threshold: To solve the problem of poor understandability of the speech signals processed by the fixed wavelet threshold, a new speech enhancement method of adaptive wavelet thresholds is presented. The types of additive noise are to be ascertained firstly according to the differences in the spectrum amplitude between white noise (including color noise with flatting spectrum amplitude) and color noise with varying spectrum amplitude:

$$\text{thr}(d_i, \lambda) = \begin{cases} \text{sgn}(d_i)(|d_i| - \lambda) & |d_i| \geq \lambda \\ 0 & |d_i| < \lambda \end{cases} \qquad (9)$$

Where:

d_i = The wavelet coefficient before de-noising on scale i

thr (d_i, λ) = The wavelet coefficient after thresholding

λ = The soft threshold function

Generally, the soft threshold function can be defined as:

$$\lambda = \delta\sqrt{2\log_{10}^{N}} \qquad (10)$$

where, N is the sequence length of the input signal and the main noises are white noise and colored noise. So, the threshold algorithms presented by Yu-Zheng *et al.* (2009) are separately used in the background noise of white noise and the colored noise.

White noise: When the background noise is white noise (including color noise with flatting spectrum amplitude), the spectrum amplitude is flat. As the Lipschitz exponent less than zero and the wavelet coefficient is inversely proportional to the scale, so the threshold function can be modified as follows:

$$\lambda = \delta\sqrt{2\log_{10}^{N}}\; g(i) \qquad (11)$$

where, g(i) is the inverse proportion function of variable i. the soft threshold function of Eq. 11 can vary according to the variety of the scale and the variety of noise variance, so it can be called as adaptive threshold algorithm related to the scale, via choosing proper g(i) function. With lots of tests in the laboratory in the experimentation it is found (Yu-Zheng *et al*., 2009) that the inverse proportion function is as follows:

$$g(i) = \frac{1}{\left[2^{\frac{i-1}{2}} \ln(i+1)\right]} \qquad (12)$$

Colored noise: When the background noise is colored noise, the spectrum amplitude is not flat. So, the threshold function is modified with the flatness of the spectrum amplitude as:

$$\lambda' = \delta\sqrt{2\log_{10}^{N}}\; \varphi(\gamma) \qquad (13)$$

Where:

γ = Denote the flatness

$\varphi(\gamma)$ = The correction function

As different colored noise has different flatness of the spectrum amplitude, via repetitious tests, $\varphi(\gamma)$ has been introduced (Yu-Zheng *et al*., 2009) as follows:

$$\varphi(\gamma)=\frac{2}{\left[\log_{10}^{0.05\gamma}\right]} \tag{14}$$

The calculation process is as follows:

- Exerting pre-emphasis, framing and window adding processing to the original speech signal x(n)
- Making the auditory transform to each frame
- Judge the style of the background noise according to the flatness of the background noise and then choose proper threshold function according to different background noise to achieve the speech enhancement processing
- Spectral combination is done with:

$$s_k(n)=\sum_{l=0}^{N-1} S(l)\,W(l)\,e^{j\frac{2\pi nl}{N}} \tag{15}$$

Where:

N = Length of the spectrum
S (l) = Spectrum of the input speech signal
W (l) = Fourier transform of Auditory function

- The frequency synthesis is performed by using equation:

$$s(n)=\sum_{k=0}^{K-1} s_k(n) \tag{16}$$

to obtain s(n) is the spectrum of frequency synthesis
- Making the DCT to the cubic energy, the final speech features are obtained

MATERIALS AND METHODS

In the method, the speech signal is undergone usual preprocessing stage where the signal is divided into various subbands using STFT. The new auditory transform designed with actual cochlear filter structure is applied on each frame of the sub banded signal. In order to minimize the effect of added noise in the final speech recognition. Thresholding is done with adaptiveness based on the flatness of the noise.

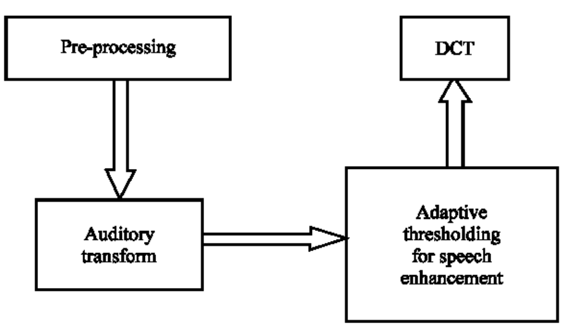

Fig. 2: A Block diagram for the proposed method

The frames after thresholding will be combined back by using inverse FFT and will be passed on Mel Filter bank for making the features not interfered by the different pitch of the input signal. Doing DCT, finally gives the noise robust features for speech recognition. Block diagram for the proposed method is shown in Fig. 2. The resultant features are called as Enhanced MFCC (EMFCC).

HMM is the statistical model based on the time sequence structure of speech signal and it can simulate reasonably speech time changing process and describe the whole non-stationary and local stationary of speech well. But a shortcoming of HMM is that distinguish ability is not strong enough. HMM is influenced by the number of training samples along with the increase in number of samples, the recognition rate will be improved greatly. Wavelet Neural Network (WNN) on other hand has great advantages in recognition. Not only the structure is simple, algorithm is easy to be implemented and the recognition rate is high but also linear least square method is used to train weight value, the convergence speed is fast and training time is only a few minutes (Zhang and Meng, 2007).

RESULTS AND DISCUSSION

The proposed algorithm is implemented with and without adaptive enhancement and tested for four different noise conditions. Performance of the algorithm for three different word counts and for five different noise conditions including clean speech are tabulated in Table 1. It is found that the performance in terms of SNR is greatly improved when the algorithm is used with adaptive enhancement both in MFCC and in CFCC. The use of Auditory Transform gives at least 3-5 dB improvement when the enhancement is not introduced for maximum noise disturbance of 30 dB. At the same time

Table 1: The performance (%) of algorithms for various noise conditions and different word counts

		SNR (dB)				
No. of words	Methods	15	20	25	30	Clean
10	MFCC	86.6	91.9	92.8	93.3	95.2
	EMFCC	92.7	93.2	95.1	96.0	97.9
	CFCC	94.4	95.1	96.4	97.2	98.0
	ECFCC	95.0	95.9	96.6	97.8	98.5
20	MFCC	83.8	88.5	90.4	91.4	93.5
	EMFCC	91.7	92.1	93.2	94.3	95.0
	CFCC	93.5	94.3	95.2	96.3	97.5
	ECFCC	94.3	95.0	96.2	97.1	98.0
30	MFCC	83.3	87.7	90.4	91.3	92.7
	EMFCC	90.3	91.5	92.1	93.4	94.8
	CFCC	91.2	92.7	93.0	94.0	95.3
	ECFCC	92.5	93.4	94.6	95.4	96.6

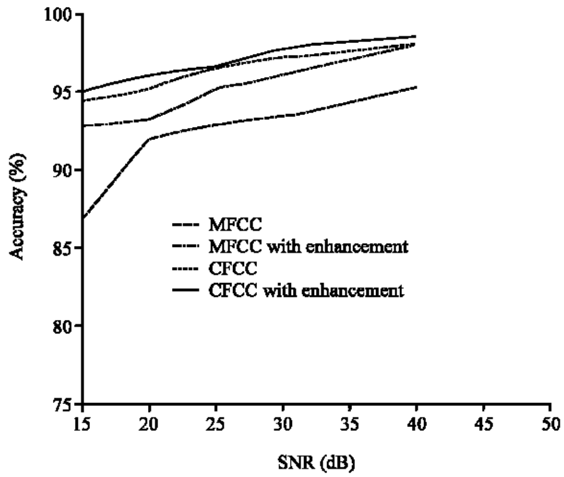

Fig. 3: Comparison of acuracy for 10 words

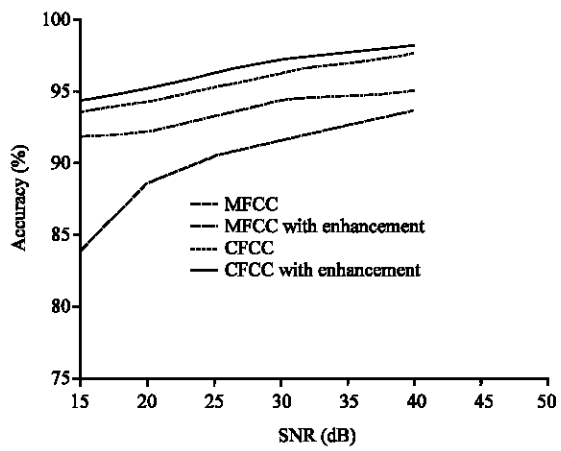

Fig. 4: Comparison of acuracy for 20 words

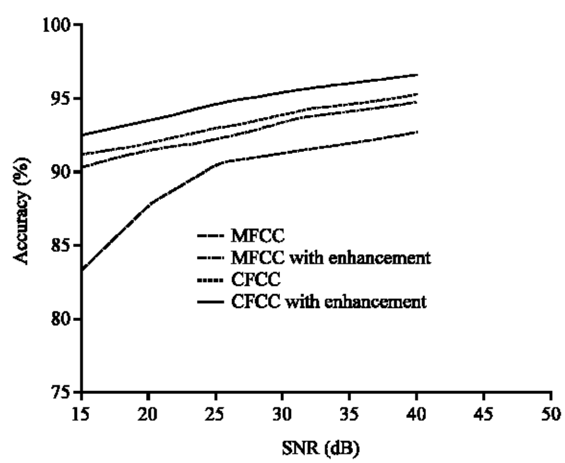

Fig. 5: Comparison of acuracy for 30 words

under low noise disturbance, i.e. at 15 dB, the improvement is 7-10 dB. Even with adaptive enhancement, the improvement from existing MFCC to CFCC is from 1-3 dB when the noise disturbance is more and at least 3 dB when the noise disturbance is less. From the Fig. 3-5, it is observed that the performance of the proposed method of feature extraction with CFCC outperforms the existing MFCC. With adaptive enhancement, the new algorithm gives still better performance compared to MFCC as evident from Fig. 3-5. The performance of the proposed algorithm with adaptive enhancement declines as the word count increases from 10-30. It is due to cumulative error. This is shown in Fig. 6.

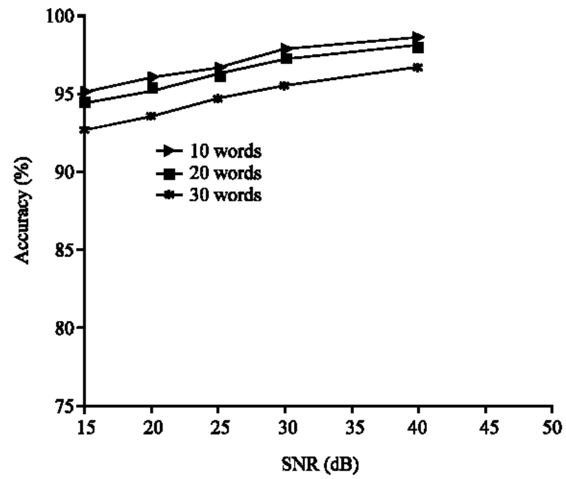

Fig. 6: Comparison of accuracy of CFCC with enhancement for different word lengths

CONCLUSION

Usually, the performance of acoustic models trained in clean speech drops significantly when tested in noisy speech. Researchers have shown that the CFCC feature with adaptive enhancement can remain high robust under high SNRs and large vocabulary conditions. And its performance is more effective under low SNRs too. The improvement is as high as 10 dB and goes to low as 1 dB in a worst case. This algorithm can further be improved by replacing DCT by DWT.

REFERENCES

Atal, B.S., 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. Acoust. Soc. Am., 55: 1304-1312.

Cooke, M. and T.W. Lee, 2006. Speech separation challenge. http://staffwww.dcs.shef.ac.uk/people/M.Cooke/SpeechSeparationChallenge.htm.

Davis, S. and P. Mermelstein, 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process., 28: 357-366.

Fu, Q. and K.C. Yi, 2000. Bark wavelet transform of speech and its application in speech recognition. Electronics, 28: 102-105.

Gold, B., N. Morgan and D. Ellis, 2011. Speech and Audio Signal Processing: Processing and Perception of Speech and Music. John Wiley and Sons Inc., New York, USA.

Hermansky, H. and N. Morgan, 1994. RASTA processing of speech. IEEE Trans. Speech Audio Proc., 2: 578-589.

Hermansky, H., 1990. Perceptual linear prediction analysis of speech. Acoust. Soc. Am., 87: 1738-1752.

Hung, X., A. Acero and H.W. Hon, 2001. Spoken Language Processing, a Guide to Theory, Algorithm and System Development. 1st Edn., Prentice Hall Inc., USA., ISBN-10: 0130226165, pp: 980.

Johannesma, P.I.M., 1972. The Pre-Response Stimulus Ensemble of Neurons in the Cochlear Nucleus. In: IPO Symposium on Hearing Theory, Cardozo, B.L., E. de Boer and R. Plomp (Eds.). IPO, Eindhoven, The Netherlands, pp: 58-69.

Kim, D.S., S.Y. Lee and R.M. Kil, 1999. Auditory processing of speech signal for robust speech recognition in real-world noisy environments. Speech Audio Process., 7: 55-68.

Lawrence and Rabiner, 1999. Fundamentals of Speech Recognition. Tsinghua University Press, ISBN-10: 0130151572, Beijing, China.

Li, Q., 2009. An auditory-based transform for audio signal processing. Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 18-21, 2009, New Paltz, NY, pp: 181-184.

Moore, B.C. and B.R. Glasberg, 1983. Suggested formula for calculating auditory-filter bandwidth and excitation patterns. Acoust. Soc. Am., 74: 750-753.

Moore, B.C.J. and B.R. Glasberg, 1996. A revision of Zwicker's loudness model. Acustica Acta Acustica, 82: 335-345.

Pujol, P., D. Macho and C. Nadeu, 2006. On real-time mean-and-variance normalization of speech recognition features. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 1, May 14-19, 2006, Toulouse, France, pp: 773-776.

Yao, W., T. Yao and T. Han, 2001. Development and prospect of robust speech recognition. Signal Process., 17: 484-497.

Yu-Zheng, Z., L. Go-Liang, Z. Jie and L. Xiao-Ying, 2009. The application of adaptively enhanced bark wavelet MFCC and the introduction of a novel noise-robust speech recognition system. Proceedings of the International Workshop on Information Security and Application, November 21-22, 2009, Qingdao, China, pp: 168-172.

Zhang, X. and W. Meng, 2007. The Research of Noise-Robust Speech Recognition Based on Frequency Warping Wavelet. In: Robust Speech Recognition and Understanding, Grimm, M. and K. Kroschel (Eds.). I-Tech, Vienna, Austria.

Zwicker, E. and E. Terhardt, 1980. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. J. Acoust. Soc. Am., 68: 1523-1525.