www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

PAN ARCHIVES OF ACOUSTICS
Vol. 40, No. 1, pp. 25–31 (2015)
Copyright © 2015 by PAN – IPPT
DOI: 10.1515/aoa-2015-0004

Phase Autocorrelation Bark Wavelet Transform (PACWT) Features for Robust Speech Recognition

Sayf A. MAJEED, Hafizah HUSAIN, Salina A. SAMAD

Department of Electrical, Electronic and Systems Engineering
Faculty of Engineering and Built Environment, National University of Malaysia
UKM, Bangi, Selangor Malaysia; e-mail: Sayf_alali@yahoo.com

(*received July 17, 2014; accepted November 19, 2014*)

In this paper, a new feature-extraction method is proposed to achieve robustness of speech recognition systems. This method combines the benefits of phase autocorrelation (PAC) with bark wavelet transform. PAC uses the angle to measure correlation instead of the traditional autocorrelation measure, whereas the bark wavelet transform is a special type of wavelet transform that is particularly designed for speech signals. The extracted features from this combined method are called phase autocorrelation bark wavelet transform (PACWT) features. The speech recognition performance of the PACWT features is evaluated and compared to the conventional feature extraction method mel frequency cepstrum coefficients (MFCC) using TI-Digits database under different types of noise and noise levels. This database has been divided into male and female data. The result shows that the word recognition rate using the PACWT features for noisy male data (white noise at 0 dB SNR) is 60%, whereas it is 41.35% for the MFCC features under identical conditions.

Keywords: speech recognition, feature extraction, phase autocorrelation, wavelet transform.

1. Introduction

Speech recognition can be approximately divided into two stages: feature extraction and classification. Feature extraction is a crucial step of the speech recognition process. If vital information is lost during this stage, the performance of the following classification stage is inherently defective (Majeed *et al.*, 2012).

Most conventional features that are designed for speech recognition are based on the power spectrum or the magnitude spectrum of the speech signal, such as the mel frequency cepstrum coefficients (MFCC) algorithm (Rabiner, Juang, 1993). Power or magnitude spectra blindly represent the spectral content of the signal. Hence, with an external noise, the spectral content of the noise is also included, which can make the feature vectors notably sensitive to external noise and cause a bad performance of the speech recognition systems in noisy conditions (Ikbal *et al.*, 2012). Thus, the researchers attempted to find solutions to overcome the weaknesses of feature extraction in noisy speech.

Many methods have been proposed to improve the feature extraction. These methods can be categorized into three groups. The first group is the speech enhancement techniques, working at the signal level. This stage precedes the feature extraction process. However, these approaches improve the speech signal by eliminating or reducing the impact of noise on the speech spectrum immediately before extracting features from it. The best examples of speech enhancement techniques are Spectral Subtraction (Boll, 1979), Nonlinear Spectral Subtraction (Yapanel *et al.*, 2001), Wiener filter (Vaseghi, 2008), Kalman filter (Paliwal, Basu, 1987), and Adaptive noise cancellation (Sambur, 1978; Jie, Zhenli, 2009). The second group is the robust feature extraction which improves the feature extraction algorithm by changing or modifying some inner processes to obtain the feature vectors. Good examples of the robust feature extraction techniques are mel frequency teager cepstral coefficients (MFTCC) (Nehe, Holambe, 2009) and autocorrelation MFCC (AMFCC) (Shannon, Paliwal, 2006). The last group contains feature compensation or feature enhancement techniques at the feature level, which follows the feature extraction process. When these techniques are implemented, a transformation is directly placed on

the feature vectors to compensate the noise effects on the extracted features. Good examples of these techniques are cepstral mean normalization (CMN) (Liu *et al.*, 1993) and principal component analysis (PCA) (Jolliffe, 2005).

The most popular features in the modern speech recognition systems are most likely the mel frequency cepstral coefficients (MFCCs) (Rabiner, Juang, 1993; Shannon, Paliwal, 2006). The steps to compute the MFCC feature extraction from the speech signal are as follows (Davis, Mermelstein, 1980): (1) Implement short-time Fourier transform (STFT) to the speech signal with a finite-duration window (e.g., a 32 ms Hamming window) and apply the periodogram technique to compute the power spectral estimation of the speech signal; (2) the power spectrum is passed through a mel filter bank to obtain the filter-bank energies; and (3) the discrete cosine transform (DCT) is applied to the log filter-bank energies to obtain the MFCCs. In clean speech, the MFCC features perform well; otherwise, their performance rapidly degrades. However, MFCC extracts only the magnitude of the spectrum. The phase information is usually discarded because we traditionally believe that the human auditory system is phase-deaf (Zhu, Paliwal, 2004). In addition, using the DCT, which is a linear transformation, gives equal weights to all logarithmic energies (Nasersharif, Akbari, 2007). Equal weighting of DCT and discarding the phase spectrum make MFCC features highly sensitive to noise. As it will be shown later, the phase is defined as a nonlinear transformation of the dot product and its use as a measure of correlation results in relative emphasis of the peaks over valleys in the spectral domain. This leads to an improved noise robustness, since the spectral peaks are considered to constitute the noise robust components of the spectrum.

In this paper, a new feature extraction method has been proposed to overcome the limitation of MFCC in noisy speech. This method attempts to achieve robustness based on an alternative measure of autocorrelation, which is known as phase autocorrelation (PAC), and the bark wavelet transform (Zhang *et al.*, 2005; Ikbal *et al.*, 2012), where PAC uses the phase (i.e., angle) difference of the speech signal frame over time to measure the correlation. Conventional autocorrelation computes the correlation coefficients as the dot product of the time-delayed speech vectors. In addition, the bark wavelet transform, which has good time and frequency resolutions, has been used instead of the Fourier transform to alleviate the previously stated issues. We refer to this new feature extraction method as phase autocorrelation bark wavelet transform (PACWT).

The remainder of this paper is organized as follows: Sec. 2 provides a brief overview of the phase autocorrelation. The bark wavelet transform is explained in Sec. 3. In Sec. 4, we introduced our proposed method, the phase autocorrelation bark wavelet transform (PACWT). Section 5 describes the recognition experiments and their results. The paper ends with a conclusion in Sec. 6.

2. Phase autocorrelation

The inspiration to use an angle to measure the correlation depends on the belief that with external noise, the angle changes less than the dot product (Mansour, Juang, 1989). The conventional autocorrelation method is computed as the dot product of the time-delayed speech vectors. Lately, a different measure of autocorrelation, which is known as phase autocorrelation (PAC), has been introduced. This measure depends on the angle between the vectors in the signal vector space (Ikbal *et al.*, 2012). Here, a brief overview of the phase autocorrelation algorithm is provided. For any speech recognition system, the speech signal is divided into a sequence of frames:

$$s = \{s[0], s[1], \ldots, s[N-1], \tag{1}$$

where N is the frame size. Suppose there are two vectors x_0 and x_k as:

$$\begin{aligned} x_0 &= \{s[0], s[1], \ldots, s[N-1]\}, \\ x_k &= \{s[k], \ldots, s[N-1], s[0], \ldots, s[k-1]\}. \end{aligned} \tag{2}$$

Applying the dot product, the autocorrelation coefficients of the speech frame are calculated using:

$$R[k] = x_0^{\mathrm{T}} x_k. \tag{3}$$

$R[k]$ can also be written as:

$$R[k] = |x|^2 \cos(\theta_k), \tag{4}$$

where $|x|^2$ refers to the energy of the frame, and θ_k denotes the angle between vectors x_0 and x_k in the N-dimensional space. The new set of correlation coefficients $P[k]$ is created by using the angle θ_k as the measure of correlation instead of the dot product. These coefficients $P[k]$ are computed as:

$$P[k] = \theta_k = \arccos\left(\frac{R[k]}{|x|^2}\right). \tag{5}$$

Based on the above equations, the PAC coefficients $P[k]$ only depend on θ_k, which is less susceptible to the external noise than $R[k]$ (Ikbal *et al.*, 2012). The inverse cosine transformation can improve the spectral peaks out of spectral valleys and add less weight to some high-frequency information of the spectrum as described in (Ikbal *et al.*, 2012).

3. Wavelet transform

The wavelet transform uses short windows to determine the high-frequency information in the signal, whereas the low-frequency content of the signal is measured using long windows. This characteristic makes the wavelet transform better than the short-time Fourier transform and Fourier transform (Tufekci, Gowdy, 2000). Consequently, the wavelet transform has been commonly used in speech and image processing.

Whereas the short-time Fourier transform (STFT) provides a fixed resolution at all frequencies, the wavelet transform applies a multi-resolution technique, where different frequencies are analyzed with various resolutions. The continuous wavelet transform (CWT) of a signal $s(t)$ is described as (Addison, 2010):

$$S(a,b) = \frac{1}{\sqrt{a}} \int s(t)\psi^*\left(\frac{t-b}{a}\right) \mathrm{d}t, \tag{6}$$

where $\psi^*_{a,b}(t)$ is the analyzing function, which is the scaled and time-shifted version of the wavelet function $\psi^*(t)$, b is the time-shifting parameter, and a is the scaling parameter. Equation (6) can be easily interpreted in three ways. First, it can be viewed as a scalar product of the signal $s(t)$ and $\psi^*_{a,b}(t)$ analyzing function. Therefore, the signal details can be analyzed with different resolutions (scales) at the time instant $t = b$.

Based on the second interpretation the signal $s(t)$ can be analyzed by a series of linear systems with impulse responses of the form $\frac{1}{\sqrt{a}}\psi(-t/a)$, therefore a wide variety of signal changes in $s(t)$ can be acquired from the slow ($a > 1$) to the rapid ($a < 1$) ones. Equation (6) can be calculated in frequency domain through the use of inverse Fourier transform (Rioul, Duhamel, 1992; Addison, 2010):

$$S(a,b) = \sqrt{a} \int_{-\infty}^{+\infty} S(\omega)\Psi^*(a\omega)e^{jb\omega} \,\mathrm{d}\omega. \tag{7}$$

It leads to a third interpretation because the argument of $\Psi^*(a\omega)$ is in direct proportion to the frequency at a given scale a. Consequently, using the ratio of the bandwidth $\Delta\omega$ and the centre frequency ω_c, the ratio $\Delta\omega/\omega_c$ remains constant, thus (6) simply is a constant relative bandwidth (constant-Q) analysis (Pintér, 1996). For sampled signals the computations can be achieved with inverse DFT at different scales or with direct evaluation of an acceptable approximation of (6):

$$S(n,a) = \frac{1}{\sqrt{a}} \sum_k s(k)\psi^*\left(\frac{k-n}{a}\right). \tag{8}$$

The wavelet transform is merged with MFCC feature extraction algorithm to obtain robust features. Tufekci and Gowdy applied the discrete wavelet transform (DWT) to the mel-scaled log filter-bank energies of a speech frame to achieve good time and frequency localizations (Tufekci, Gowdy, 2000). The bark wavelet transform was used with MFCC by Zhang *et al.* The bark wavelet is particularly designed for speech signal, and it depends on the psychoacoustic bark scale (Zhang *et al.*, 2006).

3.1. Bark wavelet transform

The human auditory system has a nonlinear mapping relation with the actual frequency and a linear relation with the bark frequency. Equation (9) shows the relation between the linear frequency and the bark frequency (Traunmüller, 1990):

$$b = 13\arctan(0.76f) + 3.5\arctan\left(\frac{f}{7.5}\right)^2, \tag{9}$$

where b is the bark frequency, and f is the linear frequency. The fundamental concept of designing a bark wavelet is usually as follows. Because of the equal importance of the time and frequency in the speech analysis, which are lead to the optimality in Gabor sense, the function of (approximately) minimum uncertainty and unity bandwidth is essential for this optimality. As it is well known, this function is the Gaussian (Gabor, 1947; Reid, Passin, 1992) and it can be expressed as:

$$W(b) = e^{-c_1 b^2}. \tag{10}$$

Constant c_1 is selected as 4ln2, when the unit bandwidth is defined as 3 dB. The bandwidths of the mother wavelet are all unit bandwidths on the Bark scale, i.e., 1 Bark. In order to make alterable wavelet windows, $W_k(b)$ can be defined as:

$$\begin{gathered} W_k(b) = W(b - b_1 - k\Delta b), \\ k = 0, 1, \ldots, K-1, \end{gathered} \tag{11}$$

where

$$\Delta b = \frac{(b_2 - b_1)}{K-1},$$

is the translation step-length of $W_k(b)$, k is the scale parameter, $K = 24$ is the total number of sub-bands, b_2 is the highest bark frequency number of the speech signal, and b_1 is the lowest bark frequency number of the speech signal. Then, by substituting (9) and (11) into (10), the bark wavelet function in the linear frequency can be written as:

$$W_k(f) = c_2 2^{-4[13\arctan(0.76f)+3.5\arctan\left(\frac{f}{7.5}\right)^2 - a^*]^2}, \tag{12}$$

where

$$a^* = (b_1 + k\Delta b).$$

The value of c_2 can be found at a given frequency band with the condition below for the perfect reconstruction:

$$c_2 \sum_{k=0}^{K-1} W_k(b) = 1, \qquad 0 < b_1 \leq b \leq b_2. \tag{13}$$

Because of the strict equality in $[b_1,\ b_2]$ in Eq. (13), initially we place one unity bandwidth Gaussian at b_1. Then, starting from a small value of Δb, by its systematic increasing the interval $[b_1,\ b_2]$ is covered, and the resolution of the unity holds in $[b_1,\ b_2]$. In the end, the normalizing constant c_2 is computed as the reciprocal value of the overall sum in Eq. (13).

Then, bark wavelet transform in linear frequency can be expressed:

$$S_k(t) = \int_{-\infty}^{\infty} S(f) W_k(f) e^{j2\pi f t}\, \mathrm{d}f, \tag{14}$$

where $S(f)$ is the frequency spectrum of the speech signal.

4. Phase autocorrelation bark wavelet transform (PACWT) feature extraction method

In this section, the phase autocorrelation bark wavelet transform (PACWT) feature extraction method and its difference from the conventional MFCC feature extraction method are discussed. Figure 1 illustrates the block diagram of the PACWT feature extraction method.

Initially, the speech signal is pre-emphasized using $H(z) = 1 - 0.97z^{-1}$ to increase the signal energy at high frequency given that the low-frequency band is filled by useless/harmful sounds for speech recognition. Frame blocking and frame shift are performed with 200 and 100 samples per frame, respectively. The Hamming window is used on the pre-emphasized signal of a given frame. By applying Eq. (5), a phase autocorrelation sequence $P_n[k]$ is obtained with 25 ms long. Then, the bark wavelet transform is simply applied to the $P_n[k]$ as:

$$S_k(n) = \sum_{l=0}^{N-1} P_n(l) W_k(l) e^{\frac{j2\pi nl}{N}}, \tag{15}$$

where $P_n(l)$ is the frequency spectrum of signal $P_n[k]$, and $W_k(l)$ is the discrete form of $W_k(f)$ in Eq. (10).

Then, the signal $S(n)$ passes through the mel filter bank, which can smooth the frequency spectrum, minimize the harmonic, and emphasize the main formant of the speech signal. Thus, the feature coefficients do not include the tone and the pitch of the speech sound. However, the speech recognition system should not be interfered with by a different pitch of the input speech signal. Finally, 12 PACWT feature coefficients are obtained by applying Eq. (16) as follows:

$$\text{PACWT features} = \sum_{m=0}^{M-1} W_k(m) D(m), \tag{16}$$

where

$$D(m) = \log\left(\sum_{n=0}^{N-1} |S(n)|^2 . H_m(n)\right),$$

$D(m)$ is the log of the mel filter bank output energies $m = 1, 2, \ldots, M$, and $H_m(n)$ represents the response of the mel filter banks, $1 \leq m \leq M$, M is the total number of filters.

Additionally, the log energy of the windowed signal is calculated and added to the 12 PACWT feature coefficients to obtain 13 base features. The first and second derivatives (delta and delta-delta) of the time sequence of each base feature are also calculated. These derivatives are concatenated to the base feature set to have the final PACWT feature coefficients set (with 39 features). The MFCCs are the most widely used features for speech recognition, and the block diagram

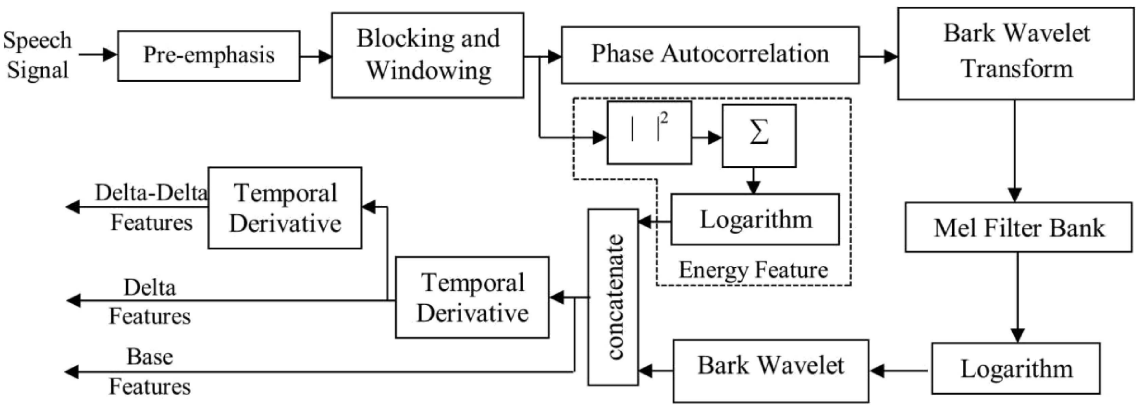

Fig. 1. Block diagram of the PACWT feature extraction algorithm.

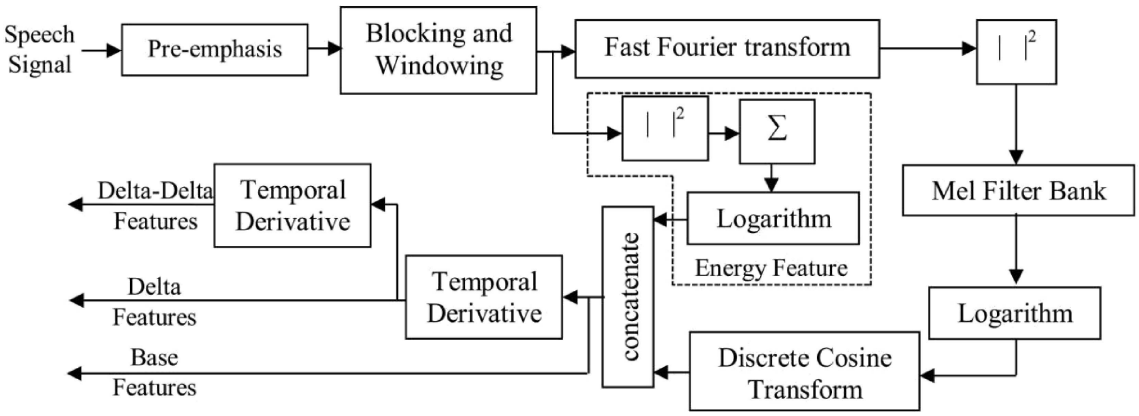

Fig. 2. Block diagram of the MFCC feature extraction algorithm.

of the MFCC feature extraction method is shown in Fig. 2.

The main differences in these two methods are the process to estimate the power spectrum of the speech signal and the decorrelation of the spectral vectors. In the MFCC scenario, the power spectrum is estimated by applying a STFT to the Hamming windowed speech signal; the Discrete Cosine Transform (DCT) is used to decorrelate the mel-spectral vectors. In the scenario of the PACWT feature coefficients, the power spectrum is estimated by calculating the bark wavelet transform of the phase autocorrelation, which is derived from the Hamming windowed speech signal, whereas the bark wavelet transform is used instead of the DCT. These differences can be observed in the block diagrams in Fig. 1 and Fig. 2.

5. Experiments and results

In this section, the recognition performance of the PACWT features is evaluated at different SNRs using the TI-Digits database. To obtain the noisy speech, the speech signal is corrupted using three different types of additive noises: pink, white, and babble noises, and are added to both training and testing sets.

5.1. Database

The TI-Digits database is used as a benchmark dataset for isolated word recognition. It was collected at Texas Instruments in the early 1980's to develop and evaluate algorithms for speaker-independent recognition of the connected digit sequences (Leonard, 1984). The version in the experiments is down sampled to 8000 samples per second. Each speaker pronounces each digit twice. The dataset was divided into a training set and a test set and into male and female speakers. The "o" digits will not be used; to represent a zero, the "zero" digits are used, which creates 10 digit classes from 0 to 9.

5.2. Experimental setup

In the experiment, the speech sampling frequency is 8000 Hz, and the frame length and frame shift are 200 and 100 samples per frame, respectively. Hamming window is used as the window type. The bark wavelet transform, which has the property of multi-resolution, is used to process the speech data.

An MFCC feature is computed using the *melcepst* function in the Voicebox toolbox of Matlab. A similar window length and a similar function are used for the spectrogram experiments. 12 coefficients are extracted; furthermore, the log energy, delta, and delta-delta coefficients are computed.

For the classification, the support vector machine (SVM) is used. For a set of training samples, each of which is labeled as belonging to one of the ten classes, an SVM training algorithm builds a model that predicts whether a new sample belongs to one class or the other. In this work, the LIBSVM (Chang, Lin, 2011) library is used. This library supports multiclass classification.

5.3. Results

The recognition performance of the PACWT feature extraction method is compared with that of the MFCC method. We evaluate the recognition performance of the PACWT method using the speech that is corrupted with three types of noises, two of which are stationary: white noise, which is artificial, and pink noise, which is real. The last type of noise is non-stationary noise, which is real babble noise. Furthermore, we divided the speech data into male and female data, and the word recognition rate results for male and female data are listed in Tables 1 and 2, respectively.

From these tables, it is obvious that white noise is the worst type of noises because it includes all audible frequencies. To make the comparison between the MFCC and PACWT features more convenient, we plot

Table 1. Word recognition rate of male data in MFCC and PACWT.

PACWT features			
SNR, in dB	Word recognition rate, in %		
	White	Babble	Pink
Clean	91.43	91.43	91.43
20	92.86	91.00	91.43
15	91.43	90.71	91.14
10	87.86	90.00	91.09
5	77.86	82.14	90.71
0	60.00	68.57	82.86

MFCC features			
SNR, in dB	Word recognition rate, in %		
	White	Babble	Pink
Clean	98.57	98.57	98.57
20	93.18	95.80	95.20
15	86.52	93.06	92.46
10	77.99	84.63	84.03
5	58.00	69.32	67.43
0	41.35	45.95	44.92

Table 2. Word recognition rate of female data in MFCC and PACWT features.

PACWT features			
SNR, in dB	Word recognition rate, in %		
	White	Babble	Pink
Clean	82.14	82.14	82.14
20	78.57	81.43	82.14
15	73.57	76.43	77.14
10	68.57	72.86	66.43
5	61.43	65.00	63.57
0	51.43	56.43	60.43

MFCC features			
SNR, in dB	Word recognition rate, in %		
	White	Babble	Pink
Clean	97.86	97.86	97.86
20	92.32	96.41	96.22
15	86.45	93.88	92.01
10	74.39	86.40	83.41
5	60.23	71.49	69.57
0	47.14	51.71	54.07

the recognition accuracies for the male and female data that are corrupted by white noise as a function of SNR in Figs. 3 and 4, respectively.

The PACWT feature extraction method is generally more noise-robust than the MFCC, particularly in high-noise (low-SNR) environments. However, MFCC has a higher recognition rate than the PACWT features in clean speech (high SNR) because of the nonlinear transformation used to compute the angle θ_k in PACWT, which deemphasizes the noise sensitive components that otherwise would have been useful during the clean speech recognition.

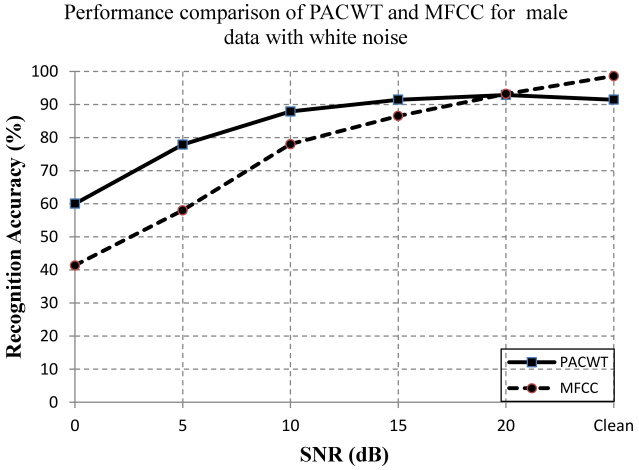

Fig. 3. Performance comparison of PACWT and MFCC features for white noise of male data.

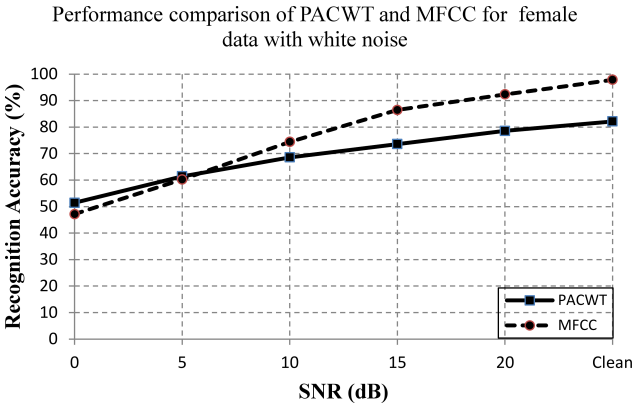

Fig. 4. Performance comparison of PACWT and MFCC features for white noise of female data.

In contrast, the female data have lower PACWT performance than the male data possibly because of the differences among males and females in voice quality. Female speakers have higher formant frequencies and breathier quality than male speakers, and female speakers speak faster on average than male speakers. The PACWT is more evident at high-noise conditions.

6. Conclusion

In this paper, a new method of feature extraction has been introduced for robust speech recognition. This method applies the phase autocorrelation technique and the bark wavelet transform. The speech

recognition performance of the PACWT features has been evaluated using the TI-Digits database. Comparing with MFCC, the PACWT features perform much better in low-SNR conditions, and the recognition performance was significantly better for male data than for female ones. To further improve the PACWT method in female data, we will analyze in-depth the characteristics of speech and the factors that affect the speech recognition performance in females. Furthermore, we will attempt to find a good method to enhance speech recognition in environments with low noise (high SNR).

References

1. Addison P.S. (2010), *The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance*, CRC Press.
2. Boll S. (1979), *Suppression of acoustic noise in speech using spectral subtraction*, IEEE Transactions on Acoustics, Speech and Signal Processing, **27**, 2, 113–120.
3. Chang C.C., Lin C.J. (2011), *LIBSVM: a library for support vector machines*, ACM Transactions on Intelligent Systems and Technology (TIST), **2**, 3, 27.
4. Davis S., Mermelstein P. (1980), *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*, IEEE Transactions on Acoustics, Speech and Signal Processing, **28**, 4, 357–366.
5. Gabor D. (1947), *Acoustical quanta and the theory of hearing*, Nature **159**, 4044, 591–594.
6. Ikbal S., Misra H., Hermansky H., Magimai-Doss M. (2012), *Phase AutoCorrelation (PAC) features for noise robust speech recognition*, Speech Communication, **54**, 7, 867–880.
7. Jie Y., Zhenli W. (2009), *On the application of variable-step adaptive noise cancelling for improving the robustness of speech recognition*, Computing, Communication, Control, and Management, CCCM 2009, ISECS International Colloquium on, IEEE.
8. Jolliffe I. (2005), *Principal component analysis*, Wiley Online Library.
9. Leonard R. (1984), *A database for speaker-independent digit recognition*, IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'84, IEEE.
10. Liu F.H., Stern R.M., Huang X., Acero A. (1993), *Efficient cepstral normalization for robust speech recognition*, Proceedings of the workshop on Human Language Technology, Association for Computational Linguistics.
11. Majeed S., Husain H., Samad S., Hussain A. (2012), *Hierarchical K-Means Algorithm Applied On Isolated Malay Digit Speech Recognition*, International Proceedings of Computer Science & Information Technology, **34**, 33–37.
12. Mansour D., Juang B.H. (1989), *A family of distortion measures based upon projection operation for robust speech recognition*, IEEE Transactions on Acoustics, Speech and Signal Processing, **37**, 11, 1659–1671.
13. Nasersharif B., Akbari A. (2007), *SNR-dependent compression of enhanced mel sub-band energies for compensation of noise effects on MFCC features*, Pattern recognition letters, **28**, 11, 1320–1326.
14. Nehe N.S., Holambe R.S. (2009), *Isolated Word Recognition Using Normalized Teager Energy Cepstral Features*, International Conference on Advances in Computing, Control, & Telecommunication Technologies, ACT '09.
15. Paliwal K., Basu A. (1987), *A speech enhancement method based on Kalman filtering*, IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE ICASSP'87.
16. Rabiner L., Juang B.H. (1993), *Fundamentals of speech recognition*, PTR Prentice-Hall, Inc, Englewood Cliffs, New Jersey, USA.
17. Reid C.E., Passin T.B. (1992), *Signal processing in C*, John Wiley & Sons, Inc.
18. Rioul O., Duhamel P. (1992), *Fast algorithms for discrete and continuous wavelet transforms*, IEEE Transactions on Information Theory, **38**, 2, 569–586.
19. Sambur M. (1978), *Adaptive noise canceling for speech signals*, IEEE Transactions on Acoustics, Speech and Signal Processing, **26**, 5, 419–423.
20. Shannon B.J., Paliwal K.K. (2006), *Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition*, Speech Communication, **48**, 11, 1458–1485.
21. Traunmüller H. (1990), *Analytical expressions for the tonotopic sensory scale*, The Journal of the Acoustical Society of America, **88,** 1, 97–100.
22. Tufekci Z., Gowdy J. (2000), *Feature extraction using discrete wavelet transform for speech recognition*, Proceedings of the IEEE, Southeastcon 2000.
23. Vaseghi S.V. (2008), *Advanced digital signal processing and noise reduction*, Wiley.
24. Yapanel U., Hansen J.H., Sarikaya R., Pellom B. (2001), *Robust digit recognition in noise: an evaluation using the AURORA Corpus*, Proc. Eurospeech.
25. Zhang X., Jiao Z., Zhao Z. (2005), *The speech recognition based on the bark wavelet front-end processing*, Fuzzy Systems and Knowledge Discovery, Springer, 302–305.
26. Zhang X., Bai J., Liang W. (2006), *The speech recognition system based on bark wavelet MFCC*, 8th International Conference on Signal Processing IEEE.
27. Zhu D., Paliwal K.K. (2004), *Product of power spectrum and group delay function for speech recognition*, Proceedingson IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04).