Research Journal of Applied Sciences 14 (8): 250-257, 2019
ISSN: 1815-932X

A New Robust Resonance Based Wavelet Decomposition Cepstral Features for Phoneme Recoszgnition

Ihsan Al-Hassani, Oumayma Al-Dakkak and Abdlnaser Assami
Department of Telecommunication, Higher Institute for Applied Science and Technology HIAST, Damascus, Syria

Abstract: Robust Automatic Speech Recognition (ASR) is a challenging task that has been an active research subject for the last 20 years. And still results are very modest in the highly noisy environments. In this study, we propose a new speech parameterization method based on concatenating two wavelet packet decompositions, one decomposition using low Q-factor wavelet and another with high Q-factor wavelet, to extract speech features suitable for ASR task in noisy conditions. Experiments on TIMIT dataset for phonemes recognition show that the proposed wavelet-based features outperform MFCC in all noisy conditions.

Key words: ASR, speech features, wavelet packet decomposition, phoneme classification, Q-factor, TIMIT

INTRODUCTION

Speech is the oldest and most effective means of communication among humans. Hence, the main purpose of speech recognition is to enable humans to communicate with computers more naturally and effectively using some algorithms that convert speech signals into the corresponding text form. Speech recognition techniques are one of the most challenging technologies in the field of computer science.

Early speech recognition systems were based on knowledge-based algorithms. The designers of the systems set rules for interviewing the sound signals of their phonemes as well as high-level cognitive tools (such as the dictionary, etc.).

Speech recognition process consists of three basic steps. The first step is the fragmentation of the acoustic signal into time-stable frames and applying feature extraction to each frame. The second step is building an acoustic model that can describe the different parts of speech signals. The last step is the classification process (decoding) that takes an arbitrary speech and determines the specific sequence of the acoustic models to which it belongs given that a language model is provided to determine that textual sequence that best describes the acoustic signal.

Although, many techniques have been developed for automatic speech recognition, the most widespread technique is based on Gaussian Mixture-Hidden Markov Model GMM-HMM which is the nucleus of the speech acoustic modeling. Each phoneme is represented by its own statistical model. HMM-based ASR is preferred because of better generalization characteristics and low memory requirements (Rabiner and Juang, 1993).

Recently, ASR has found new applications such as speech recognition over the mobile network which needs to run in different environments under different noise conditions and hence the need to develop robust ASR that can perform well in noisy environments. The performance of the ASR depends mainly on the robustness and resistance of the acoustic features to noise. Different robust feature extraction techniques have been proposed in the literature like RASTA-PLP (relative spectra) processing (Hermansky and Morgan, 1994), one-sided autocorrelation LPC (Linear Predictive Coefcients) (Yuo and Wang, 1999) and power difference (Xu and Wei, 2000) and cepstral subtraction method (Rahim *et al*., 1996). All these studies have been carried out using the STFT (Short-Time Fourier transform) based features (Farooq and Datta, 2004). The recognition performance of the plosives is found to be specically poor with STFT based features. This is due to the fact that although we assume that the signal is stationary during the window duration, it is not perfectly true for the case of plosives (Farooq and Datta, 2004).

In this study, we propose a new robust feature extraction algorithm for speech signal based on wavelet transform. The proposed algorithm make use of two different types of wavelet decomposition one with high quality wavelet filters and the other with low quality wavelet filters and make a concatenation of both decompositions.

Literature review: Speech recognition depends on the extraction of characteristic features in the acoustic signal and then represents these features using an appropriate statistical data model.

Corresponding Author: Ihsan Al-Hassani, Department of Telecommunication, Higher Institute for Applied Science and Technology HIAST, Damascus, Syria

The process of feature extracting aims at extracting discriminative characteristics of the linguistic content in the speech signal. In the literature, there are many parameterizing features techniques to characterize speech signal but Mel Frequency Cepstral Coefficients (MFCC) is the most common technique (Davis and Mermelstein, 1980).

Because the speech signal is highly unstable, this signal is usually divided into overlapping frames with fixed lengths ranging from 15~30 msec. This partitioning (segmentation) method results in the loss of the phonemes boundaries. The frame can contain two parts of two consecutive phonemes not one.

Farooq and Datta (2001) proposed using a set of wavelet filters to achieve the frequency discrimination of the Mel scale. The two-channel Daubechies (DB) filters were used to derive new sets of features. An improvement was observed in classifying unvoiced phonemes as the recognition performance for the unvoiced fricative phonemes and voiced stops phonemes in TIMIT database is found to be superior using the derived WP (Wavelet Packet) features over MFCC features (Farooq and Datta, 2001, 2003).

Choueiter and Glass (2007) proposed designing new wavelets using filter design methods. Two filter design techniques referred to as filter matching and attenuation minimization are used to design the new wavelets. To improve the exibility in frequency partitioning, they proposed implementing rational filter banks that naturally incorporate the critical-band effect in the human auditory system. Experiments using energy-based measurement show that the designed wavelets outperform off-the-shelf wavelets as well as an MFCC baseline in a phonetic classi cation task (Choueiter and Glass, 2007).

Hung and Fan (2009) proposed a novel scheme that applies feature statistics normalization techniques for robust speech recognition. In the proposed approach, the processed temporal-domain feature sequence is first decomposed into non-uniform subbands using the Discrete Wavelet Transform (DWT) and then each sub band stream is individually processed by well-known normalization methods such as Mean and Variance Normalization (MVN) and Histogram Equalization (HEQ). Experiments show that all the normalized features outperform MFCC, especially in noisy environments.

Pavez and Silva (2012) proposed a new type of phonetic features for speech recognition systems based on wavelet packets called "Wavelet-Packet Cepstral Coefficients (WPCC)". This research is based on the idea of optimal selection of filters for patterns recognition based on the principle of "minimal probability of error recognition". Phonemes recognition experiments were conducted using the proposed features on TIMIT database. The performance of the proposed WPCC features was compared to the MFCC features. The researchers showed that the proposed features outperform the MFCC features. Sahu *et al*. (2014) and Biswas *et al*. (2015), proposed a new set of acoustic features that also depend on wavelet packets called Wavelet packet based ERB (Equivalent Rectangular Bandwidth) Cepstral (WERBC). The main idea was to develop wavelet packet tree decomposition similar to the 24 sub-bands of the ERB filters (Sahu *et al*., 2014). Comparative study with baseline systems is presented to show the robustness of the proposed WERBC features. The multi-resolution property of wavelet allows for a better modeling of phoneme classes, especially for voiceless class. The performance of the new feature is studied for the task of phoneme recognition. WERBC features have shown an overall improvement in recognition performance for English phoneme as compared to Wavelet like MFCC (WMFCC) (Sahu *et al*., 2014) and STFT based features. WERBC is found to be superior compared to the WMFCC, especially in case of noisy condition. The speaker independent results show considerable improvement in recognition of the phoneme classes tested with TIMIT database. Further, the wavelet-based features are found to be robust in the presence of different noises.

Vignolo *et al*. (2016) proposed the use of a multi-purpose genetic algorithm to optimize wavelet-based speech representation where the most relevant parameters are selected from the complete wavelet packet decomposition, so that, the accuracy of classification of phonemes is maximized and the number of used features is minimized (after the vector of features and thus reduced complexity) using the multi-objective genetic algorithm. Experiments have shown that the classification of phonemes using selected wavelet features surpasses the best acoustic features, especially in the noisy environments.

Upadhyaya *et al*. (2018) proposed using a Mel scaled M-band wavelet filter bank structure to extract robust acoustic features for speech recognition application. Results (Biswas *et al*., 2016) shows that the proposed feature extraction from the proposed filter bank shows an improvement in terms of Word Recognition Accuracy (WRA) at all SNR range (20-0 dB) over baseline (MFCC) features.

Palo and Mohanty (2017) proposed a combination of reduced features for emotional speech recognition. The baseline features like wavelet, LPCC (Linear Prediction Cepstral Coefficient) and MFCC coefficients are extracted. These features are also extracted from wavelet coefficients instead of the direct signal and are named as WLPCC and WMFCC. Next to it, VQ (Vector Quantization) method of reduction is applied to the baseline MFCC, LPCC and resultant WLPCC and WMFCC. Different combination of these reduced feature sets are attempted and compared for enhancement in

accuracy (Palo and Mohanty, 2017). Experiments show that the combined features exhibit superior performance in terms of Mean Square Error (MSE) for classification task using Radial Basis Function Network (RBFNN) classifier (Palo and Mohanty, 2017).

Wang *et al*. (2018) proposed a feature compression algorithm entitled Suppression by Selecting Wavelets (SSW), to achieve the two main goals of Distributed Speech Recognition (DSR): minimizing memory and device requirements. And maintaining or even improving the recognition performance (Wang *et al*., 2018). The first step of SSW applies DWT to decompose the full-band input feature stream into Low-modulation Frequency Component (LFC) and High-modulation Frequency Component (HFC). The second step of SSW discards the HFC information and only preserves the LFC information prior to transmitting across a network to a remote server. As soon as the LFC feature sequence is received on the server side, the third step of SSW normalizes the LFC sequence to alleviate the environmental mismatch between training and testing phases. Next, a feature vector with all-zero elements is prepared as the HFC which works together with the normalized LFC to reconstruct the new feature stream via. inverse DWT (IDWT) (Wang *et al*., 2017). The reconstructed feature stream is further compensated via a high-pass filter which aims to alleviate possible over-smoothing effects. The resulting features are then used for speech recognition. Experiments in this study reveals that SSW can accomplish the main goals of DSR: improving performance on the back-end server while also providing up to a 50% compression rate (by discarding the HFC information) during the transmission stage (Wang *et al*., 2017).

In all previous works that used wavelet-based feature in ASR the quality factor of the wavelet filter was no taken into account. The Q-factor of an oscillatory pulse is the ratio of its center frequency to its bandwidth. The Q-factor of a wavelet transform should be chosen in part according to the oscillatory behavior of the signal to which it is applied (Selesnick, 2011a, b). For example, when using wavelets for the analysis and processing of oscillatory signals (speech, EEG, etc.), the wavelet transform should have a relatively high Q-factor. On the other hand when processing signals with little or no oscillatory behavior (such as a scan-line from a photographic image), the wavelet transform should have a low Q-factor (Selesnick, 2011a, b). Selesnick (2011a, b) proposed a wavelet transform with tunable Q-factor; the transform can be tuned according to the oscillatory behavior of the signal to which it is applied. The tunable-Q wavelet transform is based on the multirate filter banks illustrated in Fig. 1. It is composed of two filters: a Low Pass Scaling filter (LPS) with scaling

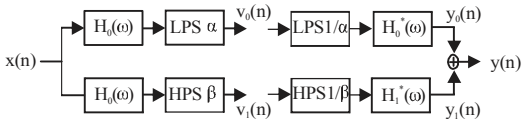

Fig. 1: Analysis and synthesis filter banks for the tunable-Q wavelet transform (Selesnick, 2011)

parameter α that preserves the low-frequency content of the x(n) signal and a High Pass Scaling filter (HPS) with scaling parameter β. The low-pass sub-band signal $v_0(n)$ and high-pass sub-band signal $v_1(n)$ have sampling rates of αf_s and βf_s, respectively where the sampling rate of the input signal is f_s.

We propose using the tunable-Q wavelet transform to derive new speech features that we entitle Resonance Wavelet Decomposition Cepstral Coefficients (RWDCC). The algorithm to derive the new feature is explained in the following section.

MATERIALS AND METHODS

Resonance Wavelet Decomposition Cepstral Coefficients (RWDCC): Speech signal is composed of phonemes with different acoustic characteristics, for example some phonemes have high resonance (oscillatory) behavior like vowels while other like stops contains two parts, one with low resonance and another with high resonance. The high-resonance components differ from the low-resonance ones by the duration to which their oscillations are sustained. By a high-resonance component, we mean a signal consisting of multiple simultaneous sustained oscillations. In contrast by a low-resonance component, we mean a signal consisting of non-oscillatory transients of unspecified shape and duration (Selesnick, 2011a, b). As Selesnick showed in (Selesnick, 2011a, b) a low-resonance pulse may be either a high frequency signal (pulse 1-Fig. 2) or a low frequency signal (pulse 3-Fig. 2). Low-resonance pulses are not restricted to any single band of frequencies. Therefore, the low-resonance component of a signal cannot be extracted from the signal by frequency-based filtering. Likewise, a high-resonance pulse may be either a high frequency signal (pulse 2) or a low frequency signal (pulse 4). Figure 3 shows an example of the decomposing the phoneme (p) into high-resonance and low-resonance components using (*q factor* function in MATLAB).

Selesnick (2011a, b) suggested decomposing signals according to the degree of resonance (resonance-based decomposition) which is a new nonlinear signal analysis method based not on frequency or scale as provided by the Fourier and wavelet transforms but on resonance. This method decomposes a complex non-stationary signal

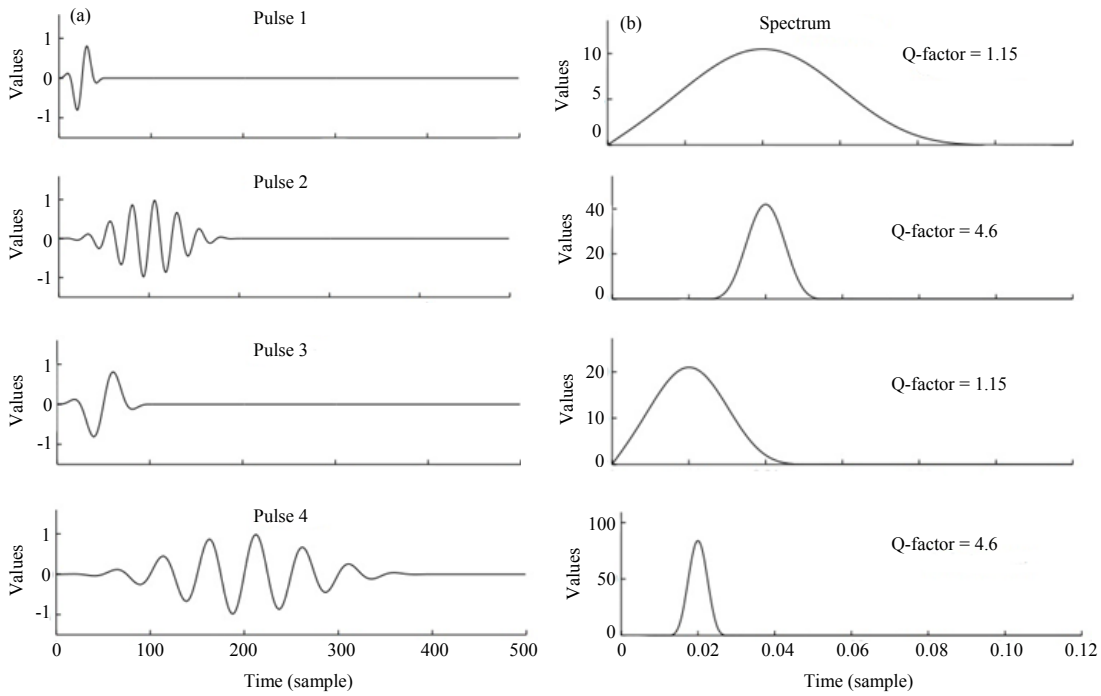

Fig. 2(a, b): A low Q-factor wavelet transform is suitable for the efficient representation of pulses 1 and 3. The efficient representation of pulses 2 and 4 calls for a wavelet transform with higher Q-factor. Selesnick (2011a, b)

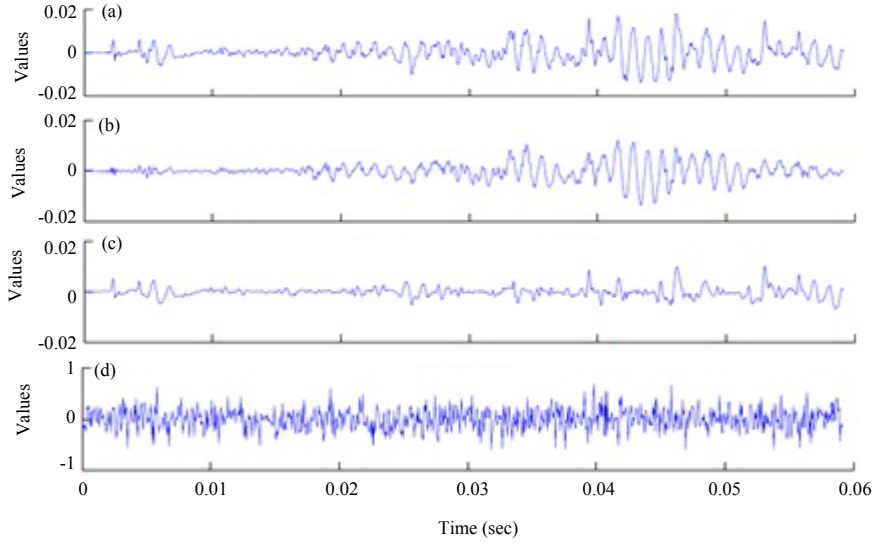

Fig. 3 (a-d): Phoneme (p) resonance component, (a) Speech signal, (b) High-resonance component, (c) Low-resonance component and (d) Residual

(such as speech signal) into a high-resonance component (multiple simultaneous sustained oscillations) and low-resonance component (non-oscillatory transients of unspecified shape and duration) using a combination of low and high Q-factor filters.

Therefore, it is not appropriate to use one type of wavelet filters when extracting wavelet-based features. This is the fundamental point in the proposed acoustic features extraction algorithm. We propose using two wavelet decomposition: one with high Q-factor wavelet to characterize the highly oscillatory parts in speech (high resonance) and another with low Q-factor wavelet to characterize the low oscillatory parts.

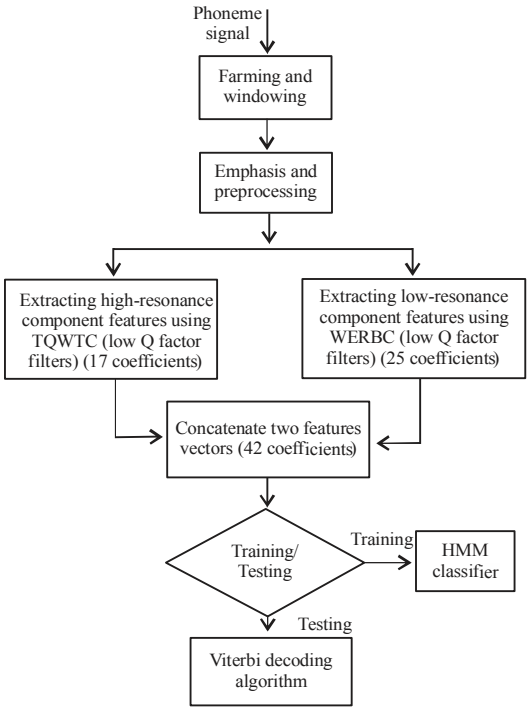

Fig. 4: RWDCC (Resonance Wavelet Decomposition Cepstral coefficient) Proposed algorithm flowchart

Wavelet transform-based feature extraction is first performed using the low Q-factor filters and high Q-factor filters. Then we concatenate the two features vectors we obtained into one single vector that captures the distinctive features of both high-resonance and low-resonance components. Figure 4 shows the flowchart of the proposed algorithm.

The proposed features extraction algorithm starts with preprocessing stage. A high pre-emphasis filter is applied on the whole speech signal, then a frame size 24 msec with 14 msec overlap is used, windowed by Hamming widows, to derive cepstral features. To extract the low resonance component features, the whole frequency band is decomposed using ERB like WP decomposition proposed by Sahu *et al*. (2014) as shown in Fig. 5.

Once, the WP decomposition is performed, energy in each frequency band is calculated and the log of weighted energy is applied resulting in 24 cepstral coefficients. Discrete Cosine Transform (DCT) is applied to decelerate the 24 coefficients of filter bank energies. Variance Feature (VF) of the 24 coefficients has been calculated. Finally, a total of 25 features that describe the low resonance component are obtained per each frame. We tried different types of wavelet filters with different degrees such as Daubechies and Coifflet filters.

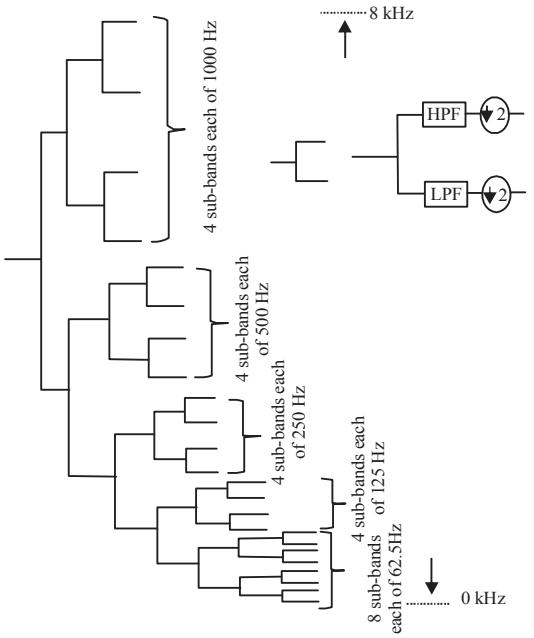

Fig. 5: The 24 sub-band wavelet packet tree based on ERB scale (Sahu *et al*., 2014)

Experiments showed that the coif5 filter gives the best performance in terms of classification accuracy; therefore, it is adopted in all the reported results below.

The quality factor of the coifflet5 was Q = 1.4, indicating that the features extracted using these filters would capture the characteristics of the low-resonance components.

As for the high resonance component feature extraction stage, we first apply the Tunable Q factor Wavelet (TQWT) filters to the preprocessed speech frames. These filters are implemented by iteratively applying a two-channel filter bank (Selesnick, 2011). A high quality factor Q = 5 was selected to capture the characteristics of the high-resonance components. The over sampling coefficient (redundancy coefficient) according to Selesnick (2011a, b) must verify $r \geq 3$, the chosen value is r = 3. Thus, the scaling parameters β and α (based on the previous values for Q and r) are:

$$\beta = \frac{2}{Q+1} = \frac{2}{6} = 0.3333 \qquad \alpha = 1 - \frac{\beta}{r} = 0.89$$

The maximum number of the decomposition levels J of the wavelet transform is given by Selesnick (2011a, b):

$$J_{max} = \left\lfloor \frac{\log(\beta N/8)}{\log(1/\alpha)} \right\rfloor$$

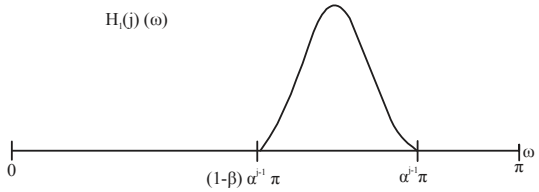

Fig. 6: Filter $H1^{(j)}$ with the tunable quality factor of phase j (Selesnick, 2011a, b)

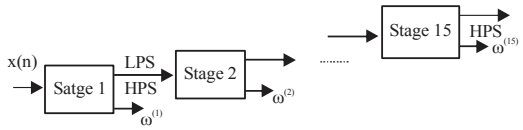

Fig. 7: Distribution of wavelet filter bank (Selesnick, 2011)

where, N is the number of samples per frame. In our experiments N = 384 (as the sampling frequency is 16 KHz and the frame size is 24 msec). Therefore, $J_{max} = 23$. We have tested different values {12, 15, 18, 20, 22} for J and found that J = 15 gives the best performance in terms of classification accuracy. This value is chosen in our experiments for the reported results.

Figure 6 shows the frequency response of filter $H1^{(j)}$ with the tunable Q factor of phase j. Figure 7 shows how the levels of the wavelet filter bank are distributed.

As in the case of a low-resonance, the energy in each 15 sub-band signals ($\omega^{(1)}$, ..., $\omega^{(15)}$) which is the output of the 15 tunable Q factor wavelet filters is calculated and the log of weighted energy is applied resulting in 16 cepstral coefficients. Discrete Cosine Transform (DCT) is applied to decelerate the 16 coefficients of filter bank energies. Variance Feature (VF) of the 16 coefficients has been calculated. Finally, a total of 17 features that describe the high resonance components are obtained per each frame. At the last, the two feature vectors for both low and high resonance components were concatenated to obtain single feature vector with 42 coefficients.

RESULTS AND DISCUSSION

Experiments

TIMIT Dataset: For all the experiments, the TIMIT corpus was adopted presented in this research. TIMIT is widely used as a standard corpus to evaluate the performance of new acoustic features in ASR because it is a phonetically balanced database and has good coverage of speakers and dialects. All of these make TIMIT a sufficiently challenging corpus to evaluate new ASR methods which justifies its wide adoption by the community. The TIMIT corpus consists of 6300 utterances for 8 major dialects of the United States. There are 630 different speakers, each one speaking 10 sentences. For this experiment, a subset of 39 English language phonemes with 1000 utterances from complete training set were used for training have been carried out. Also, all phonemes with 100 utterances from complete test set were used for testing. The speech signal was pre-emphasized to ensure that all formants of acoustic signals have similar amplitudes, so that, they get equal importance in subsequent processing stages.

Table 1: Phonemes recognition rate of studied features in clean environment

Feature	WERBC	TQWTC	RWDCC	MFCC
Recognition rate (%)	**75.02**	**71.26**	**76.63**	74.28

Experimental results: Experiments were performed on a subset derived from the TIMIT database which includes the 39 English language phonemes. The performance of the proposed RWDCC features was compared after implementing the proposed algorithm in terms of phoneme recognition rate with the performance of three other acoustic features: MFCC features, the WERBC features proposed by Sahu *et al*. (2014) and the TQWTC-based features proposed by Selesnick (2011). Experiments are conducted in three different environments:

- Clean environment without noise
- In the presence of white noise

An environment with unstable noise. Where six types of non-stationary noise were tried from a dataset available at:

- The noise of chatter
- F16 aircraft noise
- Tank noise
- Factory noise
- Noise of HF radio channel
- Volvo engine noise

Table 1 shows a comparison between the performance of feature types without noise in terms of recognition rate while Table 2 shows a comparison between performance of feature types in the seven studied types of noise.

From Table 1, we note that in the absence of noise, the proposed RWDCC features outperform, in case of recognition rate, wavelet-packet based WERBC and TQWTC by a large proportion. The RWDCC outperform WERBC by (1.6%), TQWTC by (5.4%) and MFCC by (2.4%). We also note that the WERBC features outperform TQWTC by 3.7%.

Table 2 shows that if white noise is present, the proposed RWDCC algorithm outperforms the studied features extraction algorithms for all noise levels.

Table 2: Phonemes recognition rate in the case of white noise

Noise type	dB	WERBC	TQWTC	RWDCC	MFCC
White noise	-5	52.75	52.83	**55.79**	53.23
	0	58.52	59.48	**59.64**	57.88
	5	64.41	62.36	**65.61**	61.96
	10	66.85	65.57	**67.73**	65.89
	15	70.62	66.69	**71.90**	65.97

Table 3: Phoneme recognition accuracy rate for the different features in the presence of 6 different kinds of non-stationary noise

Noise type	dB	WERBC	TQWTC	RWDCC	MFCC
Volvo	-5	71.90	69.49	**72.38**	66.13
1	0	72.31	69.73	**72.94**	66.85
	5	73.26	71.12	**74.78**	64.93
	10	74.22	70.78	**74.06**	65.17
	15	73.53	70.94	**74.78**	66.77
Babble	-5	51.63	51.71	**53.07**	50.26
2	0	59.08	59.64	**58.68**	54.75
	5	64.21	62.92	**62.92**	61.24
	10	68.21	67.81	**69.89**	63.48
	15	72.06	68.85	**71.98**	67.17
F16	-5	50.91	53.47	**53.39**	55.71
3	0	58.36	57.55	**59.64**	57.55
	5	63.32	62.04	**65.33**	63.72
	10	67.89	66.37	**68.37**	65.33
	15	69.01	67.81	**71.52**	66.53
Factory1	-5	48.66	48.90	**50.14**	48.82
4	0	54.51	55.15	**55.39**	54.43
	5	59.64	60.68	**61.86**	59.24
	10	65.65	64.04	**67.17**	62.84
	15	68.61	68.13	**71.58**	65.65
Hf channel	-5	56.03	53.55	**56.91**	56.35
5	0	62.28	57.71	**62.52**	59.32
	5	64.13	62.12	**62.36**	62.84
	10	65.57	65.01	**68.29**	63.64
Leopard	-5	66.93	68.13	**68.61**	63.08
6	0	70.62	69.41	**70.13**	65.41
	5	71.42	70.94	**73.18**	67.17
	10	71.82	70.94	**73.56**	66.53
	15	73.42	71.02	**74.06**	66.85

Bold values are significent

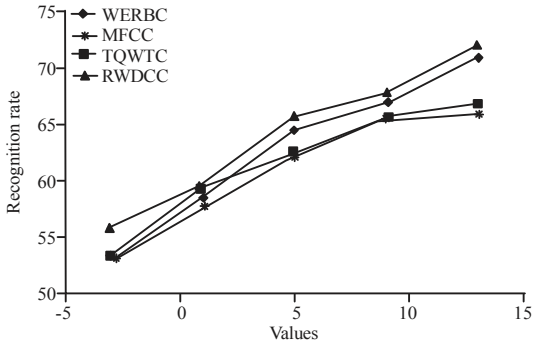

Fig. 8: Phonemes recognition rate for studied features and suggested features of RWDCC in the case of white noise for different noise levels

Figure 8 shows a comparison of the performance of the three studied features compared with the proposed RWDCC features in the case of white noise. We note that the proposed features show the best performance in terms of recognition accuracy. Table 3 shows the phoneme recognition accuracy rate using different acoustic feature compared to the propose RWDCC features in case of six different non-stationary
noise types taken from. Results show that the proposed features outperform all other feature for all SNR level. That in the case of Volvo noise and factory noise, the proposed feature extraction algorithm outperforms the studied features for all noise levels. In addition, the performance in factory noise is the worst of all studied features of all noise types. Performance in the case of Volvo noise and tank noise is close in the low noise levels (10 and 15 dB)while the performance is worse in case of tank noise at higher noise levels (-5 dB).

CONCLUSION

In this study, we proposed a new speech parameterization method based on two wavelet decompositions one using high Q wavelet filters and another using low Q wavelet filters. The tow decomposition is concatenated to have one feature vectors. Results show that the proposed features outperform the classical MFCC feature and two other wavelet-based features (WERBC and TQWTC) in clean environment and in the presence of white noise and other 6 non stationary noises in terms of phoneme recognition rate which make the proposed features suitable for ASR in noisy environment.

REFERENCES

Biswas, A., P.K. Sahu and M. Chandra, 2016. Admissible wavelet packet sub-band based harmonic energy features using ANOVA fusion techniques for Hindi phoneme recognition. IET. Signal Process., 10: 902-911.

Biswas, A., P.K. Sahu, A. Bhowmick and M. Chandra, 2015. Admissible wavelet packet sub-band-based harmonic energy features for Hindi phoneme recognition. IET Signal Process., 9: 511-519.

Choueiter, G.F. and J.R. Glass, 2007. An implementation of rational wavelets and filter design for phonetic classification. IEEE. Trans. Audio Speech Lang. Process., 15: 939-948.

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.

Farooq, O. and S. Datta, 2001. Mel filter-like admissible wavelet packet structure for speech recognition. IEEE Signal Process. Lett., 8: 196-198.

Farooq, O. and S. Datta, 2003. Phoneme recognition using wavelet based features. Inf. Sci., 150: 5-15.

Farooq, O. and S. Datta, 2004. Wavelet based robust sub-band features for phoneme recognition. IEE. Proc. Vision Image Signal Process., 151: 187-193.

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

Hung, J.W. and H.T. Fan, 2009. Subband feature statistics normalization techniques based on a discrete wavelet transform for robust speech recognition. IEEE. Signal Process. Lett., 16: 806-809.

Palo, H.K. and M.N. Mohanty, 2017. Wavelet based feature combination for recognition of emotions. Ain Shams Eng. J., 9: 1799-1806.

Pavez, E. and J.F. Silva, 2012. Analysis and design of wavelet-packet cepstral coefficients for automatic speech recognition. Speech Commun., 54: 814-835.

Rabinar, L.R. and B.H. Juang, 1993. Fundamentals of Speech Recognition (Prentice Hall Signal Processing Series). Prentice Hall PTR., USA., ISBN: 9780130151575, Pages: 507.

Rahim, M.G., B.H. Juang, W. Chou and E. Buhrke, 1996. Signal conditioning techniques for robust speech recognition. IEEE. Signal Process. Lett., 3: 107-109.

Sahu, P.K., A. Biswas, A. Bhowmick and M. Chandra, 2014. Auditory ERB like admissible wavelet packet features for TIMIT phoneme recognition. Eng. Sci. Technol. Int. J., 17: 145-151.

Selesnick, I.W., 2011a. Resonance-based signal decomposition: A new sparsity-enabled signal analysis method. Signal Process., 91: 2793-2809.

Selesnick, I.W., 2011b. Wavelet transform with tunable Q-factor. IEEE. Trans. Signal Process., 59: 3560-3575.

Upadhyaya, P., O. Farooq and M.R. Abidi, 2018. Mel scaled M-band wavelet filter bank for speech recognition. Int. J. Speech Technol., 21: 797-807.

Vignolo, L.D., H.L. Rufiner and D.H. Milone, 2016. Multi-objective optimisation of wavelet features for phoneme recognition. IET. Signal Process., 10: 685-691.

Wang, S.S., P. Lin, Y. Tsao, J.W. Hung and B. Su, 2017. Suppression by selecting wavelets for feature compression in distributed speech recognition. IEEE. ACM. Trans. Audio Speech Lang. Process., 26: 564-579.

Xu, J. and G. Wei, 2000. Noise-robust speech recognition based on difference of power spectrum. Electron. Lett., 36: 1247-1248.

Yuo, K.H. and H.C. Wang, 1999. Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences. Speech Commun., 28: 13-24.