Received: October 29, 2019

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Discrete Wavelet Denoising into MFCC for Noise Suppressive in Automatic Speech Recognition System

Hay Mar Soe Naing[1]* **Risanuri Hidayat[1]** **Rudy Hartanto[1]** **Yoshikazu Miyanaga[2]**

[1]*Electrical Engineering and Information Technology Department, Gadjah Mada University, Indonesia*
[2]*Graduate School of Information Science and Technology, Hokkaido University, Japan*
* Email: haymarsoenaing@uscy.edu.mm

Abstract: Automatic Speech Recognition (ASR) is a challenging task and the most problematic issues being in presence of background noise and substantial variability in speech. Extracting the noise-robust features adjust for speech degradations due to noise effect retained popular issue in recent years. This paper presented a framework for wavelet denoising scheme and analysed the different wavelet families and proper thresholding rule into feature extraction to enhance the performance of ASR system. Gaussian Mixture Model-based Hidden Markov Model (GMM-HMM) and Deep Neural Network (DNN)-HMM are used as the speech recognizer. The recognition performance shows that the noise-robust features are obtained while combining with the wavelet transform denoising into Mel Frequency Cepstral Coefficient (MFCC) on Aurora2 database. The best accuracy is gained by cross entropy DNN-HMM training using denoising with Coiflet wavelet and Rigrsure threshold, which provides 97.54% in 10dB, 93.13% in 5dB, 75.63% in 0dB and 37.29% in −5dB.

Keywords: Speech recognition, Wavelet denoising, MFCC, GMM-HMM, Cross entropy DNN-HMM.

1. Introduction

Speech is the effectual communication media among humans and speech processing system, consequently becoming a challenging research topic for many researchers and engineers [1]. Among them, automatic speech recognition (ASR) is a trust research topic in speech processing area and it is a multiple class sequential pattern recognition process. The primary intention is to translate from spoken signal into readable word sequences correctly. There are two main processes usually needed to perform the recognition, i.e., feature extraction and building an acoustic model.

After the decades of development, the speech recognition system still needs to be improved and remains far from the desired target as the machine cannot understand any scenarios by any speakers in any environments [2]. Many marketed ASR systems can work adequately with clean speeches. However, when the noise is corrupted in, the performance of ASR system may decrease significantly. Hence, a noise robust speech recognition system is important and necessitated to develop.

In speech recognition system, speech waveform is extracted the discriminative feature vectors which indicates the spectral information and what is being spoken by the speech. Then, a speech recognizer is used these acoustic features for pattern recognition [3]. Without an appropriate feature extractor, the recognition performance can substantially impair to the ASR system. The relevant and good acoustic features can discriminate the different speech classes and tolerate the environmental noise, and variability characteristics of different speakers [4]. Therefore, noise robustness has been a crucial problem in the speech processing area. In literature, mel frequency cepstral coefficients (MFCC) is the most widely spread speech parametrization techniques in speech recognition and speaker identification systems. The calculating of these coefficients takes into account of the properties of human auditory system, that helps to obviously gain the good performance when

employed in speech applications. MFCC based on the cepstral domain analysis, which simulates the human hearing mechanism and can capture the main characteristics of each phoneme from the speech utterances. The main drawback of this technique is a low-level of noise resistance and leads to a sharp deterioration in speech applications [5].

Recently, many researchers have been proposed to increase the noise immunity of MFCC in different perspectives. The implementation of the wavelet transformation denoising combined into the MFCC feature extraction had been presented [6]. They expressed the best denoising parameters to remove the stationary noise and proved the performance in Indonesian isolated digit recognition. In another study [7] is evaluated that the use of psychoacoustic model of frequency masking into the conventional MFCC. They also introduced the transformation mechanism of power spectral density at the multiple fundamental harmonics frequencies to achieve the relative improvement in single words recognition system. However, this technique was decreased the calculation speech and massively required more computational resources. A modified approach of power normalized cepstral coefficient system by utilizing the large time power and minimizing the channel bias was presented [8]. They intended to increase the noise robustness of the system. In their experiments, the highest accuracy was achieved in the car noise condition at SNR -5dB, however, the recognition performance was imperfect in chatter noises like babble, exhibition hall and restaurant environments.

This paper proposed a framework for wavelet transformation (DWT) denoising into conventional MFCC to develop an automatic connected speech recognition system. We analysed the effectiveness of different wavelet families and thresholding rules to remove the real environmental noise from speech signal. This presented denoising procedure was evaluated on the Aurora-2 speech corpus and Kaldi toolkit is employed to build a speaker independent acoustic model. Our proposed work intends to lessen the effect of non-stationary noise specially under the disturbance of environmental noise. The current denoising procedure emphasize the Approximation component is thresholded the noise along with the Detail components. Moreover, the system tolerates the variability of different speaker characteristics using the speaker adaptive training model. With the use of wavelet denoising scheme into the MFCC, dramatically outperformed the recognition accuracy, especially at low SNRs, whereas at high SNR value, the recognition accuracy was almost unchanged. Moreover, the proposed algorithm could not affect the processing speed and did not require more computational resources than the traditional version.

The outline of this paper is organized as follows. The overview of the acoustic feature extraction is described in Section 2. The detail explanation of discrete wavelet denoising framework and feature recognition schemes are deeply explained in Section 3 and Section 4. Then, the speech materials usage, experimental results and discussion parts are shown in Section 5. In Section 6, we will conclude our proposed work.

2. Acoustic feature extraction

Mel Frequency Cepstral Coefficients (MFCC) is a commonly used technique in the automatic speech recognition system. MFCC aims to mimic the behaviour of the human auditory system and capture the main characteristic of phones in speech. MFCC has low complexity, but recognition performance in clean conditions is high. However, the background noise and contaminated situation affect and hamper the recognition performance [9]. To overcome this problem, the authors enforced the discrete wavelet denoising into the conventional MFCC approach. The detail front-end process is illustrated in Fig. 1. The discrete wavelet denoising is worked on the input speech signal, and the denoised output signal is chunked as the multiple short time frames. Each frame is done with hamming windowing to keep the continuity at the boundary.

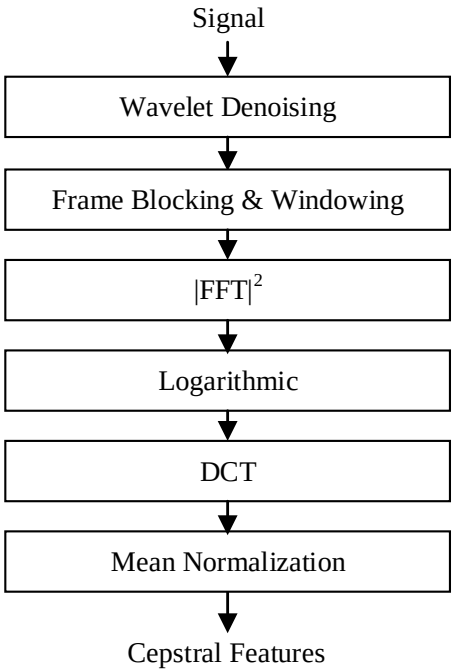


Figure. 1 Block diagram of MFCC feature extraction with wavelet denoising process

The Fast Fourier transform (FFT) is performed to convert from time domain to frequency domain. After that, Mel filterbank analysis is used to extract the spectral envelope, which is constituted by how much energy exist in different frequency regions. After finishing the filterbank analysis, the discrete cosine transforms (DCT) is applied to decorrelate the filterbank coefficients and provide a compressed representation of the filterbank. Finally, the cepstral mean normalization is processed to compensate the speech variability in cepstral domain. This technique is effective in reducing the effects of additive environmental noise and talking style variance [10].

3. Discrete wavelet denoising

Wavelet thresholding is one of the most popular and widely used technique in the signal denoising process. The discrete wavelet transform tends to separate the signal into approximation subband and detailed subband. Most of the noise components focus on the high frequency subband while the main information concentrates on the low frequency sub-band [11]. To select a proper wavelet function is essential in denoising process as the inappropriate wavelet functions may lead to the wavelet transform are complex and unmanageable [12]. Therefore, this study investigated the different wavelet functions into connected digits speech recognition.

3.1 Orthogonal wavelet families

There are many wavelet families, such as Haar, Daubechies, Coiflets, and Symlets etc. [13-15]. An example of mother wavelets that yield in Matlab are illustrated in Fig. 2.

3.1.1. Haar wavelet

Haar wavelet is the first mother wavelet and has the shortest length of support among all orthogonal wavelets. Haar wavelet transform is discontinuous and resembles the step function. It has only one vanishing moment and unsuitable for approximation of smooth functions. However, it is computationally simple and fast. Moreover, it can easily detect time localized information and is a memory efficient.

3.1.2. Daubechies wavelet

Daubechies wavelets are a family of orthogonal wavelets and characterized by a maximal number of vanishing moments for some given support. The Db1 wavelet is the same as Haar wavelet. Like the Haar transform, the Daubechies wavelet transform is implemented as a succession of decompositions. The only difference is that the filter length is more than two. So, it is more localized and smooth. The scaling signals and wavelets have slightly longer support.

3.1.3. Symlets wavelet

The Symlet wavelets are also known as the Daubechies least asymmetric wavelet and are very compactly supported, orthogonal and continuous. Construction of symlet wavelet is very similar to the construction of Daubechies wavelet. However, the symmetry is stronger than that of Daubechies.

3.1.4. Coiflets wavelet

The Coiflet wavelet is obtained by enforcing the varnishing moment condition on both scaling and wavelet functions. It is causing more coefficients. Coiflet is more symmetrical than the Daubechies wavelet and has *3p-1* support size instead of *2p-1* in case of Daubechies.

3.2 Noise thresholding rules

There are four kinds of thresholding rule mostly utilized by different researchers in signal denoising process [16].

Fixed Threshold: This is known as the universal or global threshold, and it is computed as:

$$\lambda = \sigma\sqrt{2\,log(N)} \tag{1}$$

where $\sigma = MAD/0.6745$, MAD is the median value of wavelet coefficients, λ means the representation of threshold and N is the total number of wavelet coefficients.

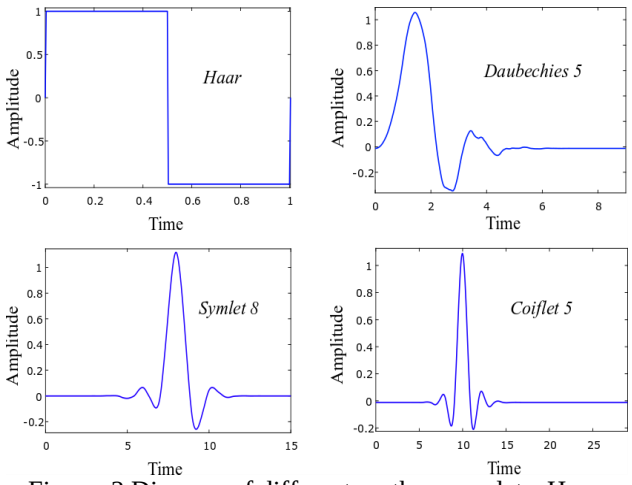


Figure. 2 Diagram of different mother wavelets: Haar, Daubachie-5, Symlet-8 and Coiflet-5

Minimax Threshold: This is also utilized the fixed threshold and yields the minmax performance for the Mean Square Error value (MSE) against an ideal procedure.

$$\lambda = \sigma \begin{cases} 0.3936 + 0.1829\, log_2 N & N \geq 32 \\ 0 & N < 32 \end{cases} \quad (2)$$

Steins unbiased risk estimator (SURE) – Rigrsure: This is an adaptive thresholding method and based on Stein's unbiased likelihood estimation principle. Let $W = [w_1, w_2, .., w_N]$ be a vector consists of the squared of wavelet coefficients and arranged in the order of small to large. Then, the risk vector R is calculated according to the vector W, and select the minimum value r_b from risk vector, given as:

$$R = r_i = \left[\frac{N - 2i + (N-i)w_i + \sum_{k=1}^{l} w_k}{N}\right] \quad (3)$$

where $i = 1,2,..,N$. The selected threshold represents $\lambda = \sigma\sqrt{w_b}$ where, w_b is the b^{th} squared wavelet coefficient at minimum risk chosen from wavelet vector W.

Heursure: This is the combination of SURE and global threshold method. If the signal-to-noise ratio of the signal is minimal, SURE method estimation will have more amounts of noises. In this situation, the fixed thresholding method gives better threshold estimation. λ_1 is the global threshold method and threshold obtained from Rigrsure is λ_2. Then, the Heuristic SURE is defined by,

$$\lambda = \begin{cases} \lambda_1 & A < B \\ min(\lambda_1, \lambda_2) & A \geq B \end{cases} \quad (4)$$

where $A = \frac{s-N}{N}$ and $B = (log_2 N)^{\frac{3}{2}}\sqrt{N}$. N is the length of wavelet coefficients and s is the squared of wavelet coefficients.

3.3 Overview of wavelet denoising process

First of all, k^{th} level discrete wavelet transforms (DWT) is taken to the input signal by decomposing it into k detail components and k^{th} approximation component. This study used 5 level decomposition schemes. Then, the noise estimation is performed for each detail components using median absolute deviation (MAD). The noise threshold for each detail subband is calculated by four thresholds: "rigrsure", "minimaxi", "sqtwolog" and "heursure". If the threshold is too large, the signal will lose overmuch detail and if the threshold is too small, the noise will remove not clean enough [17]. The thresholding function is applied to remove the low frequency components using the selected noise threshold. This study applied the soft thresholding function in the wavelet denoising process.

$$\widetilde{w}_{j,i} = \begin{cases} sgn(w_{j,i})(|w_{j,i}| - \lambda_j) & |w_{j,i}| \geq \lambda_j \\ 0 & |w_{j,i}| < \lambda_j \end{cases} \quad (5)$$

where $w_{j,i}$ is the wavelet coefficients at j^{th} level decomposition and i^{th} position of detail coefficient. λ_j is the selected threshold value. The basic idea is that the coefficient is squeezed to zero when the amplitude is lower than the threshold value and preserved or shrank any factors when the amplitude is higher than this threshold [11].

The approximation component focuses on the primary information of the signal, however, it can also contain the low-frequency noises [18]. Thus, it needs to consider removing these low-frequency noises of the approximation components. As the same manner of detail subband, the noise threshold is computed for the approximation component and removed the noise using the thresholding function. At last, the inverse wavelet discrete transform (IDWT) is taken onto the k detail components and k^{th} approximation component. The detail denoising process is illustrated in Fig. 3. This denoising scheme is performed prior to the feature extraction process.

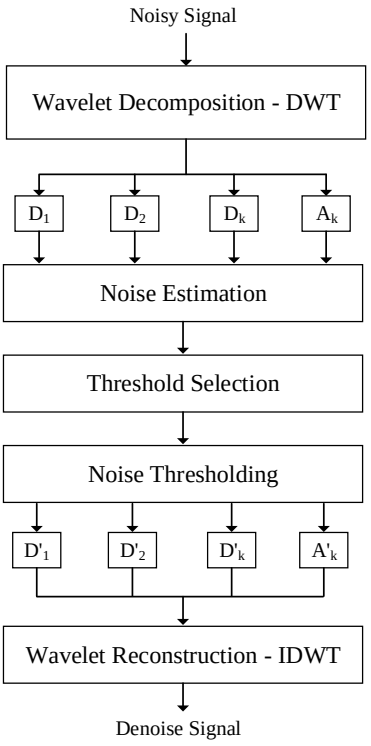


Figure. 3 The block diagram of discrete wavelet transform denoising scheme

4. Feature recognition engine

In the ASR system, most of the acoustic model building process uses the Gaussian Mixture Model based Hidden Markov Model (GMM-HMM) act as the sequential structure of speech. Each HMM state utilizes a mixture of Gaussian to model a spectral representation of the speech. This model established the alignment between the acoustic feature vectors and phoneme units. Fig. 4 illustrates the stage of speaker adaptive GMM-HMM model training for English connected digits recognition system using the Kaldi speech recognition toolkit.

4.1 GMM-HMM based acoustic model

4.1.1. Monophone training

This is the first training and does not focus on any contextual information from previous and next phonemes. This model is trained from a flat start by using 13 MFCC features.

4.1.2. Triphone training (MFCC+ Δ+ΔΔ)

The triphone model concentrates the contextual information between the left and right phonemes. In general, delta and delta-delta features represent the first order derivatives (Δ) and the second order derivatives (ΔΔ) features to support MFCCs. The feature of delta is calculated on the window of the original features, while the features of delta-delta are calculated on the window of the delta features.

4.1.3. LDA-MLLT triphone training

The linear discriminant analysis (LDA) uses the feature components and reduces feature dimension for all data to create HMM states. The maximum likelihood linear transform(MLLT) uses the reduced feature dimension from LDA and applies the linear transformation to get a significant transformation for each speaker [4].

4.1.4. Speaker adaptive training (SAT)

Speaker adaptive training performs the speaker and noise normalization by adapting to each speaker with a precise data transform. The primary goal of the speaker adaptation is to modify the parameters to match the features of the actual acoustic. Thus, the feature-space maximum likelihood linear regression (fMLLR) which is the most common techniques in ASR field is built for speaker adapted GMM model.

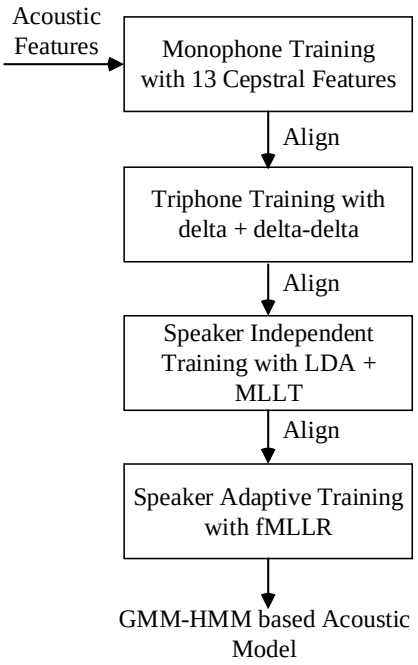


Figure. 4 The training process flow of GMM-HMM model using Kaldi toolkit

4.2 DNN-HMM based acoustic model

The DNN-HMM training has the advantage of strong learning power of DNN and the sequential modeling ability of HMM to outperform the existing GMM-HMM. At first, the GMM-HMM model is learnt using the maximum likelihood estimation and utilized for hybrid DNN-HMM model building that contains the state prior and transition probabilities. Then, the forced alignment is generated by matching the acoustic feature vectors with the corresponding frame-level state labels from Viterbi algorithm of GMM-HMM model. Finally, the DNN are trained the estimation of the posterior probabilities of HMM states from given observation sequences. The weight of DNN training is done by minimizing the cost function of cross entropy criterion [19].

$$C = -\sum_{i=1}^{Q} q_i \, log \, p_i \tag{6}$$

where C is the cross-entropy cost function, p_i is the estimating probabilities, where $i \in \{1..Q\}$ of each class. q_i are the state labels of DNN output layer.

Kaldi toolkit is used to implement the acoustic model training process for this study. Kaldi is a free, open-source software that is designed for the speech recognition research written in C++ language and is freely released by the Apache License v2.0 [20].

5. Results and discussion

The AURORA-2 is the noisy speech database, designed by the European Telecommunications Standards Institute (ETSI) to evaluate robust feature extraction for a distributed recognition framework. AURORA-2 is based on original English connected digits dataset TIDigits (as available from LDC) [21]. Additionally, the different noise signals have been artificially added into the clean speech data. This dataset is considered two training modes: training on clean data and multi-condition training on noisy data. Multi-condition training corresponds to the noise added at several SNRs (20dB, 15dB, 10dB, 5dB and clean). Four types of noise like recording inside a subway, babble noise, recording in a car and an exhibition hall are used. In total, the data from 20 different noise conditions are taken as the input for multi-condition training mode [22]. 1,001 utterances are utilized for evaluation of different noise situations at the SNRs of 10dB, 5dB, 0dB and −5dB. Summary of detail data usage was described in Table 1.

In this paper, connected digits recognition has been carried out to evaluate the noise robustness of the proposed features extraction technique. The feature extraction part is implemented using the MATLAB software. Then, the extracted acoustic feature vectors are passed through the Kaldi speech recognition toolkit to build the speaker adaptive acoustic model using GMM-HMM. The multi noise condition training model was built on 8,440 speech utterances and evaluated the performance with the word recognition (%) on 1,001 speech utterances. Table 2 shows the word recognition accuracy (%) using the conventional MFCC feature extraction approach under different types of noise conditions and SNR levels.

Table 1. The summary of Aurora-2 database

Category	**Description**	
Vocabulary	continuous digits sequences (0-9, ‘oh’)	
Sampling	44.1kHz, 16 bits, mono	
Male	111 speakers	21-70 ages
Female	114 speakers	17-59 ages
Training	8,440 utterances	Multi-condition
	Subway, Babble, Car, Exhibition hall	
	20dB, 15dB, 10dB, 5dB and clean	
Testing	1,001 utterances	
	Subway, Babble, Car, Exhibition hall	
	10dB, 5dB, 0dB, −5dB	

Table 2. Recognition accuracy (%) of conventional MFCC feature extraction approach

Noises	10dB	5dB	0dB	-5dB
Subway	96.68	91.56	69.17	18.48
Babble	96.61	86.94	48.19	17.62
Car	95.56	87.06	41.54	10.74
Exhibition	93.37	84.45	60.81	18.82
Average	95.5	87.5	54.9	16.4

Although the conventional MFCC function well in higher SNR level, the recognition results significantly degrade under lower SNR value particularly at 0dB and -5dB. Recognition accuracy (%) was 96.68% in SNR of 10dB under subway noise, 96.61% under babble noise, 95.56% under car noise and 93.37% under exhibition hall noise. In terms of SNR -5dB, the recognition performance was degraded rapidly to 18.48% in subway noise, 17.62% in babble noise, 10.74% in car noise and 18.82% in exhibition noise.

Although MFCC are renowned and widely used in speech recognition area, still present the low robustness in noise-corrupted speech signal, since all MFCC are varied by noisy signal when at least one frequency band is distorted. Apart from this, it is inherently assumed that each frame contains only one phoneme information at a time, whereas it may be the case that a frame of speech contains two consecutive phonemes information in continuous speech. To increase the immunity of noise effect into the standard MFCC, we combined the discrete wavelet transformation (DWT) denoising process into MFCC without affecting the undistorted signal and decreasing in algorithm speed than traditional version. The DWT was successfully employed for denoising task as a result of the better frequency resolution at lower frequencies region and better localised time of transient phenomena in the time domain. The idea of combining DWT and MFCC is to gain the benefit of two methods.

The wavelet denoising dramatically depends on the selection of threshold value because the poor choice of threshold leads to pretend the unremoved noise or distorted the signal. Thus, the four accessible rules; Rigrsure, Sqtwolog, Minimaxi, and Heursure are determined in this study. Each rule was applied with soft thresholding on the input signal because the soft thresholding is function well in when the detailed coefficients contain both the signal and noise components. Besides, the choice of wavelet family can result the ability of signal denoising and different mother wavelet consist of

how scaling the signal, and the wavelets are defined. Also, the choice of wavelet determines the final shape of the signal wave. For this reason, four different wavelet families, Haar, Daubechies (db5), Symlet (sym8) and Coiflets (coif5) were analysed in denoising process. Tables 3, 4, 5, and 6 illustrate the effectualness of selecting wavelet families in the accuracy of speech recognition system.

Table 3. Accuracy (%) of wavelet denoising with Db5 family and different thresholds (Db5-MFCC)

SNR	Minimax	Fixed	Rigrsure	Heursure
10dB	94.37	93.87	95.25	93.87
5dB	85.84	85.5	87.33	85.5
0dB	58.33	58.85	58.3	58.85
−5dB	20.62	19.01	17.27	19.01
Avg.	64.79	64.31	64.54	64.31

Table 4. Accuracy (%) of wavelet denoising with Haar family and different thresholds (Haar-MFCC)

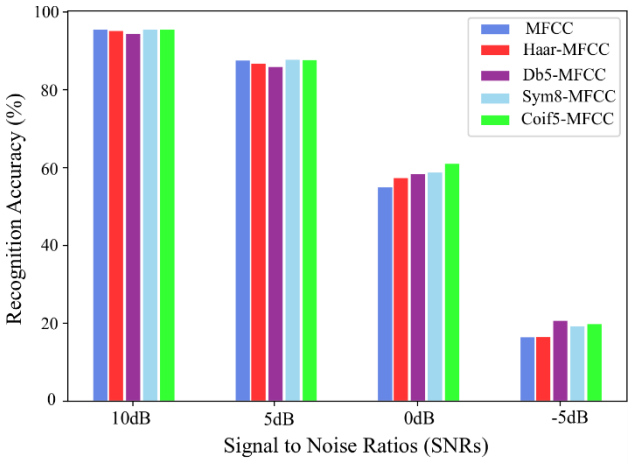
SNR	Minimax	Fixed	Rigrsure	Heursure
10dB	92.87	93.22	95.1	93.22
5dB	84.23	83.95	86.68	83.95
0dB	53.53	54.94	57.3	53.86
−5dB	15.43	16.88	16.48	16.81
Avg.	61.52	62.25	63.89	61.96

Table 5. Accuracy (%) of wavelet denoising with Sym8 family and different thresholds (Sym8-MFCC)

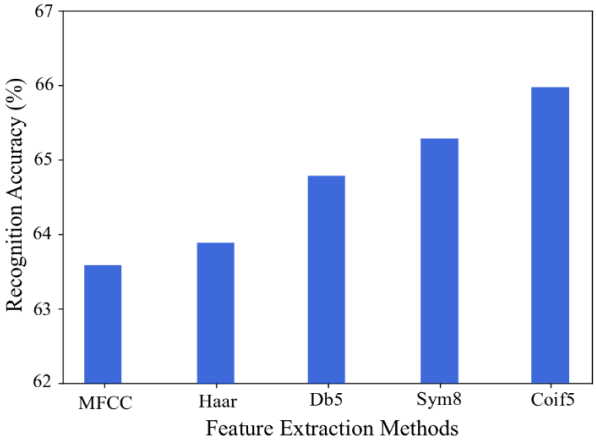
SNR	Minimax	Fixed	Rigrsure	Heursure
10dB	95.15	94.33	95.49	94.33
5dB	86.49	85.6	87.7	85.6
0dB	57.59	57.14	58.73	57.14
−5dB	19.85	19.35	19.22	19.35
Avg.	64.77	64.11	65.29	64.11

Table 6. Accuracy (%) of wavelet denoising with Coif5 family and different thresholds (Coif5-MFCC)

SNR	Minimax	Fixed	Rigrsure	Heursure
10dB	95.25	95.09	95.5	94.66
5dB	86.88	85.96	87.64	85.76
0dB	58.89	54.94	61.00	56.86
−5dB	19.23	17.85	19.78	18.88
Avg.	65.06	63.46	65.98	64.04



(a)



(b)

Figure. 5 Recognition results for comparison target: (a) accuracy (%) of MFCC and DWT-MFCC on each SNRs and (b) overall accuracy (%) of MFCC and DWT-MFCC

According to these experimental results, the combination of Haar, Symlet, and Coiflets families with Rigrsure thresholding rule provides the best recognition performance. However, the Daubechies wavelet achieved a higher average accuracy with the Minimaxi thresholding rule. Fig. 5 (a) illustrates the recognition accuracy (%) for the comparison target of conventional MFCC and proposed DWT denoising using each wavelet family into MFCC under different SNRs. From this experiment, we can clearly see that the proposed denoising approach outperformed than the MFCC specifically at low SNR and without affecting the performance for undistorted signals. Among them, the DWT with Coiflet wavelet achieve the same accuracy at 10dB and increases the accuracy from 87.5% to 87.64% at 5dB, from 54.93% to 61.00% at 0dB and from 16.42% to 19.78% at SNR of −5dB when comparing with the standard MFCC. Fig. 5 (b) shows the overall accuracy of MFCC and DWT-MFCC under overall SNR levels and different noise conditions. The average accuracy was 63.58% in standard

MFCC, 63.89% in denoising with Haar family, 64.79% with Db5, 65.29% with Symlet and 65.98% with Coif5. It can be proved that the proposed wavelet denoising method using Coif5 family is outperformed when comparing with other wavelet families.

In addition, we trained the learning power of the cross entropy based Deep Neural Network (DNN-HMM) based on the GMM-HMM model to boost the performance of the system. Table 7 shows the recognition accuracy (%) of DNN training using the DWT-MFCC feature extraction technique with the Coiflet wavelet and Rigrsure thresholding rule. When applying a cross entropy based DNN training, the accuracy was achieved up to 98.04% in a SNR of 10dB under subway noise, 97.64% under babble noise, 97.79% under car noise and 96.7% under exhibition hall noise. Besides, in term of −5dB, recognition accuracy was obtained 44.55% under subway noise, 33.86% under babble noise, 25.05% under car noise and 45.7% under the exhibition hall noise.

The average recognition accuracy over all SNR levels was carried out to determine how the results that will be yielded on different types of noise situations. The recognition in accuracy (%) was shown in Fig. 6. According to these results, the proposed wavelet denoising algorithm achieved the performance in 79.54% under subway noise, 73.16 under babble noise, 73.12% under car noise and 77.78% under exhibition noise.

Table 7. Recognition accuracy (%) of DWT-MFCC using cross entropy based DNN training

SNR	10dB	5dB	0dB	−5dB
Subway	98.04	95.36	80.2	44.55
Babble	97.64	91.38	69.7	33.86
Car	97.79	93.23	76.4	25.05
Exhibition	96.7	92.56	76.2	45.7
Average	97.54	93.13	75.63	37.29

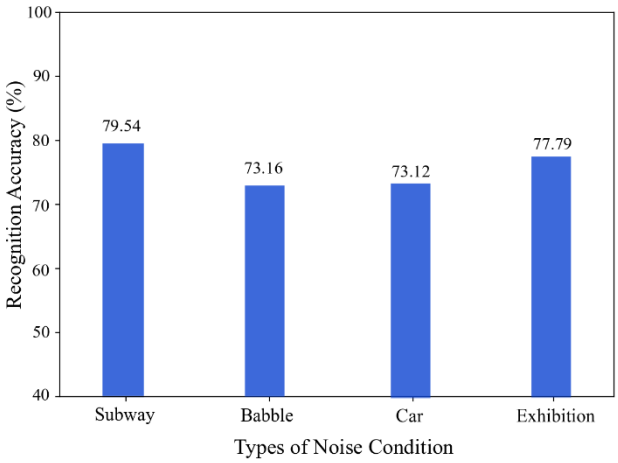


Figure. 6 Average accuracy over all SNR separated for different noise types using cross-entropy based DNN

6. Conclusion

This paper presented the wavelet transform denoising into MFCC feature extraction to develop a noise robust automatic speech recognition system. We analysed the selection of appropriate wavelet families and thresholds to remove the noise effect. The Coiflet family with Rigrsure thresholding rule achieves superior performance in all SNR levels compared to the MFCC with the use of conventional GMM-HMM model. By applying the cross entropy based DNN-HMM training, the accuracy reaches up to 97.54% in 10dB, 93.13% in 5dB, 75.63% in 0dB and 37.29% in −5dB. The future effort will assess the denoising on real environment large vocabulary data and reduce the unwanted speech degradations.

Acknowledgments

We would like to express our gratitude to ASEAN University Network/Southeast Asia Engineering Education Development Network (AUN/SEED-Net) Project JICA for their financial support for this research. This study is supported in parts by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B) (18H0321) and Fund for the Promotion of Joint International Research, Fostering Joint International Research (B) (18KK0277). This study is also supported in parts by Ministry of Internal Affairs and Communications for SCOPE Program (185001003).

References

[1] X. Huang, A. Acero, H. W. Hon, and R. Foreword By-Reddy, "Spoken Language Processing: A Guide to Theory, Algorithm, and System Development", *Prentice hall PTR*, 2001.

[2] D. Juafsky and J. H. Martin, "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition", *2nd Edition, Prentice Hall*, 2008.

[3] A. Kuamr, M. Dua, and T. Choudhary, "Continuous Hindi Speech Recognition using Gaussian Mixture HMM", In: *Proc. of IEEE Students' Conference on Electrical, Electronics and Computer Science*, pp.1-5, 2014.

[4] S. Sadhu, R. Li, and H. Hermansky, "M-vectors: Sub-band based Energy Modulation Features for Multi-stream Automatic Speech Recognition", In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6545-6549, 2019.

[5] K. M. Indrebo, R. J. Povinelli, and M. T. Johnson, "Minimum Mean-Squared Error Estimation of Mel-Frequency Cepstral Coefficients using a Novel Distortion Model", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 16, No. 8, pp. 1654-1661, 2008.

[6] R. Hidayat, A. Bejo, S. Sumaryono, and A. Win ursito, "Denoising Speech for MFCC Feature Extraction Using Wavelet Transformation in Speech Recognition System", In: *Proc. of 10th International Conf. on Information Technology and Electrical Engineering (ICITEE)*, pp.280-284, 2018.

[7] K. K. Tomchuk, "Spectral Masking in MFCC Calculation for Noisy Speech", In: *Proc. of Wave Electronics and its Application in Information and Telecommunication Systems*, St. Petersburg, Russia, pp. 1-4, 2018.

[8] M. Tamazin, A. Gouda, and M. Khedr, "Enhanced Automatic Speech Recognition System Based on Enhancing Power-Normalized Cepstral Coefficients", *Applied Sciences*, Vol.9, No.10, pp.1-13, 2019.

[9] S. Narang and M. D. Gupta, "Speech Feature Extraction Techniques: A Review", *Internation al Journal of Computer Science and Mobile Computing*, Vol.4, No.3, pp.107-114, 2015.

[10] D. Grozdic, S. Jovicic, D. S. Pavlovic, J. Galic, and B. Markovic, "Comparison of Cepstral Normalization Techniques in Whispered Speech Recognition", *Advances in Electrical and Computer Engineering*, Vol.17, No.1, pp. 21-27, 2017.

[11] Z. S. Al-Timime, "Signal Denoising Using Double Density Discrete Wavelet Transform", *Al-Nahrain Journal of Science*, Vol. 20, No.4, pp. 125-129, 2017.

[12] S. G. Mihov, R. M. Ivanov, and A. N. Popov, "Denoising Speech Signals by Wavelet Transform", *Annual Journal of Electronics*, pp. 712-715, 2009.

[13] R. L. Jyothi and A. Rahiman, "Comparative Analysis of Wavelet Transforms in the Recognition of Ancient Grantha Script", *International Journal of Computer Theory and Engineering*, Vol. 9, No. 4, 2017.

[14] C. Stolojescu, I. Railean, S. Moga, and A. Isar, "Comparison of wavelet families with application to WiMAX traffic forecasting", In: *Proc. of 12th International Conference on Optimization of Electrical and Electronic Equipment*, pp.932-937, 2010.

[15] A. Dogra, "Performance Comparison of Different Wavelet Families Based on Bone Vessel Fusion", *Asian Journal of Pharmaceutics (AJP): Free full text articles from Asian J Pharm*, Vol. 10, No. 4, 2017.

[16] N. Verma, "Performance Analysis of Wavelet Thresholding Methods in Denoising of Audio Signals of Some Indian Musical Instruments", *International Journal of Engineering Science and Technology*, Vol. 4, No. 5, pp.2040-2045, 2012.

[17] A. Stanley Raj, D. H. Oliver, Y. Srinivas, and J. Viswanath, "Wavelet denoising algorithm to refine noisy geoelectrical data for versatile inversion", *Modeling Earth Systems and Envi ronment*, Vol. 2, No. 1, pp.1-11, 2016.

[18] M. Srivastava, C. L. Anderson, and J. H. Freed, "A New Wavelet Denoising Method for Selecting Decomposition Levels and Noise Thresholds", *IEEE Access*, Vol. 4, pp. 3862–3877, 2016.

[19] V. V. R. Vegesna, K. Gurugubelli, H. K. Vydana, and B. Pulugandla, "DNN-HMM Acoustic Modeling for Large Vocabulary Telugu Speech Recognition", In: *Ghosh A., Pal R., Prasath R. (eds) Mining Intelligence and Knowledge Exploration, Lecture Notes in Computer Science*, Vol. 10682. Springer, Cham, 2017.

[20] D. Povey, A. Ghoshal, G. Boulianne, and L. Burget, "The Kaldi speech recognition toolkit", In: *Proc. of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, No. Conf. IEEE, Signal Processing Society*, 2011.

[21] R. Leonard, "A Database for Speaker-Independent Digit Recognition", In: *Proc. of International Conf. on Acoustics, Speech, and Signal Processing*, Vol. 9, No. 1, pp. 328–331, 1984.

[22] D. Pearce and H. G. Hirsch, "The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions", In: *Proc. of 6th International Conference on Spoken Language Processing ICSLP*, 2000.