Improved Voice Activity Detection for Speech Recognition System

Siew Wen Chin, Kah Phooi Seng, Li-Minn Ang, King Hann Lim
School of Electrical and Electronic Engineering
The University of Nottingham
Jalan Broga, 43500 Semenyih, Selangor Darul Ehsan, Malaysia.
{keyx8csw, Jasmine.Seng, Kenneth.Ang}@nottingham.edu.my

Abstract— **An improved voice activity detection (VAD) based on the radial basis function neural network (RBF NN) and continuous wavelet transform (CWT) for speech recognition system is presented in the paper. The input speech signal is analyzed in the form of fixed size window by using Mel-frequency cepstral coefficients (MFCC). Within the windowed signal, the proposed RBF-CWT VAD algorithm detects the speech/ non-speech signal using the RBF NN. Once the interchange of speech to non-speech or vice versa occurred, the energy changes of the CWT coefficients are calculated to localize the final coordination of the starting/ending speech points. Instead of classifying the speech signal using the MFCC at the frame-level which easily capture lots of undesired noise encountered by the conventional VAD with the binary classifier, the proposed RBF NN with the aid of CWT analyzes the transformation of the MFCC at the window-level that offers a better compensation to the noisy signal. The simulation results shows an improvement on the precision of the speech detection and the overall ASR rate particularly under the noisy circumstances compared to the conventional VAD with the zero-crossing rate, short-term signal energy and binary classifier.**

Keywords-voice activity detection; continuous wavelet transform; mel frequency cepstral coefficient; radial basis function

I. Introduction

Voice activity detection (VAD) is denoted as a crucial speech detector for most of the speech communication system, for instance, automatic speech recognition [1, 2], speech coding [3] to increase the bandwidth efficiency, speech enhancement [4] and telephony [5]. The failure of the VAD to accurately capture the presence of speech signal would directly affect the subsequent performance of the aforementioned applications. Therefore, a robust VAD for speech/non-speech detection, particularly under the noisy circumstances has become an essential topic and investigation throughout the decades [6].

There are numerous kind of VAD methodologies been proposed. The zero-crossing rate and short-term signal energy [7, 8] due to their simplicity, are normally used as the acoustic feature for the VAD. Nevertheless, the detection capability would be dramatically degraded under the noisy environments. Furthermore, the sound from the surrounding environment gives the same zero-crossing rate and energy as the speech signal, has yield the difficulty of accurate speech detection for the VAD [9]. Other than the abovementioned methods, the available VAD algorithm includes low-variance spectrum estimation [10], higher order statistic in the linear predictive coding (LPC) residual domain [11] and pitch detection [12].

The recent research tends to incorporate the statistical model into the VAD approach to improve its performance [13-15]. J.H. Chang *et al.* [13] presented the VAD algorithm based on multiple statistical models, which incorporate the complex Laplacian and Gamma probability density functions for the statistical properties analysis. Besides, the VAD approaches which integrates the neural network and hidden Markov model post-processing to work under the presence of breathing noise is proposed in [15]. In [14], the VAD using support vector machine (SVM) that employs the likelihood ratios (LRs) computed in each frequency bin as the elements of the feature vector is demonstrated. Furthermore, the VAD using mel-frequency cepstral coefficients (MFCC) and SVM is presented by Tomi Kinnunen *et al.*[16] .

In this paper, the VAD based on RBF NN and the CWT, in short named as RBF-CWT VAD is presented. Instead of utilizing the MFCC at the frame-level as the training material for the binary SVM classifier demonstrated in [16], the RBF NN dealt with the MFCC at the window-level and the computation of CWT energy change is proposed to improve the VAD performance especially under the noisy circumstances. This is due to the reason that the frame-level MFCC which used as the training material for the speech/non-speech classifier [16] would easily capture lots of undesired noise, though some of the post-processing, e.g., median filtering is applied to smooth out the VAD output, it is however part of the noisy signal are still be detected as the false detection. Alternatively, by evaluating the input audio signal with the window-level MFCC and the sliding information as proposed in this paper, the characteristic provided by the MFCC itself within the fixed window frame and the subsequent analyzes of the energy change of the CWT coefficients would offer a sequence of signal observation for the classifier and reduce the detection error.

Figure 1 illustrates the overview of the proposed RBF-CWT VAD algorithm. The audio input signal is processed within a fixed length of non-overlapped window. Under the supervised learning practice, the MFCC with delta and double delta are extracted as the most representative audio features from a number of audio signals under different environment conditions. Subsequently, the RBF NN is trained by the extracted window-level MFCC to classify the speech and non-speech signal under different level of signal-to-noise ratio. On the other hand, during the VAD operation, the CWT and its energy calculation is triggered once the frame where the interchange of speech to non-speech or vice versa occurred. The starting/ending points of the speech are localized according to the energy threshold predefined by the user and finally the detected speech signal would be sent for further ASR system to obtain final recognized speech. This paper organized as: Section 2 introduces the proposed RBF-CWT VAD algorithm and briefly discusses about the MFCC feature extraction and the process of the RBF NN. Some simulation results and analysis are demonstrated in Section 3 following by the conclusion in Section 4.

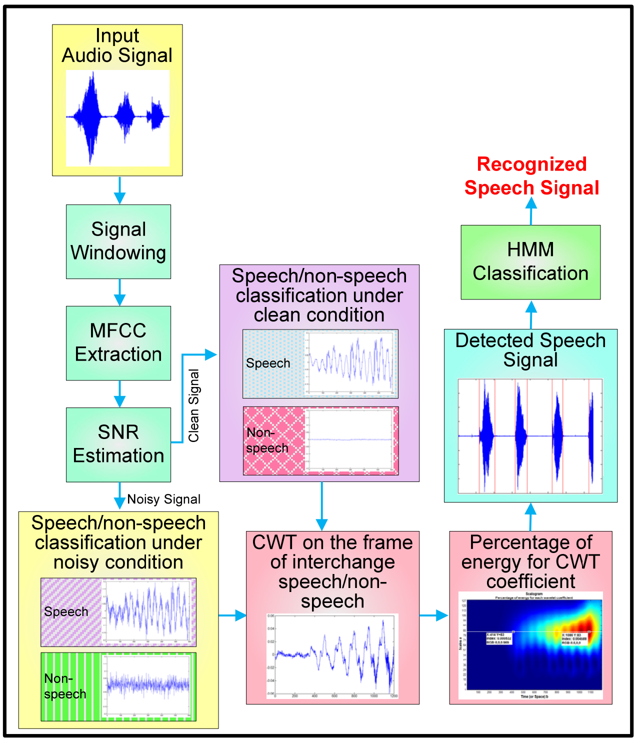

Input
Audio Signal
Signal Windowing
MFCC Extraction
SNR Estimation
Clean Signal
Noisy Signal
Speech/non-speech classification under clean condition
Speech
Non-speech
Speech/non-speech classification under noisy condition
Speech
Non-speech
CWT on the frame of interchange speech/non-speech
Percentage of energy for CWT coefficient
Detected Speech Signal
HMM Classification
Recognized Speech Signal

Figure 1. The overview of the proposed RBF-CWT VAD approach for ASR system

II. Proposed Voice Activity Detection based on Radial Basis Function Neural NEtwork and Continuous Wavelet Transform

The proposed RBF-CWT VAD algorithm contains three main parts: (i) feature extraction from the audio signal (ii) speech/non-speech classification and (iii) the energy calculation of CWT coefficients for speech localization. The respective components are discussed in details as in the sub-sections below.

A. *Mel Frequency Cepstral Coefficent- Feature Extraction*

In this paper, MFCC is used as the most representative features of the speech signal for the coming RBF classification. MFCC, which possesses the characteristic of human ear's non-linear perceptional condition (basically known as the logarithm relation) has been broadly applied in the ASR system. The general process to obtain the MFCC is depicted as in Figure 2.

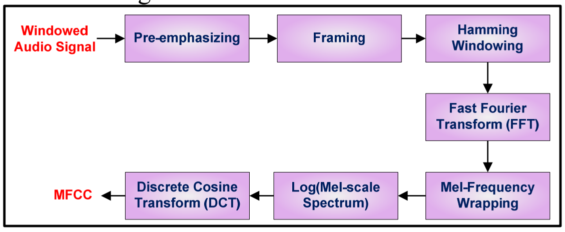


Figure 2. The process of MFCC audio signal feature extraction [17]

Before the MFCC feature extraction, the signal is first windowed with the fixed length dimension. The windowed audio signal is then goes for the pre-emphasizing process using the finite impulse response (FIR) filtering for the purpose of flattening the spectrum. After that, the pre-emphasized signal is split into the over-lapping fixed-length frames and Hamming windowing is subsequently applied onto each frame to smooth out the signal. By applying the Fourier fast transform (FFT), each of the frames is transformed from time to frequency domain and mapped onto the Mel scale [18]. The relationship between the frequency and the Mel-frequency is equated as [19]:

$$Mel(f) = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right) \tag{1}$$

With the Mel-frequency filter bank, the frequency wrapping is done and the log-scale would be then transformed back to the time domain with the Discrete Cosine Transform (DCT). The MFCC, its delta and double delta coefficients are computed as formulated in [20]. The total number of 39 coefficients is extracted as the features for the RBF NN speech/non-speech classification as discussed in the next sub-section.

$$\tilde{M}_t = \sqrt{\frac{2}{N}} \sum_{n=1}^{N} m_n \cos\left(\frac{\Pi t}{N}(n - 0.5)\right) \tag{2}$$

Where N is denoted as the number of bandpass filter while m_n is known as the log bandpass filtering output amplitude

B. *Speech/Non-speech Classification*

Referring to Figure 1, the output from the previous sub-section which is the MFCC of the windowed signal would be passed to the present discussed RBF NN for speech/non-speech classification. Before the process of classification, a simple SNR estimation is made by calculating the SNR of the current windowed signal to the fixed clean windowed non-speech signal. A SNR threshold is set as to define the clean and noisy signal. After the estimation, the extracted MFCC would be then sent to the appropriate classifier. The architecture of the RBF NN which used as the speech/non-speech classifier is briefly discussed as following. A general three-layer RBF NN as depicted in Figure 3 is adopted as the classifier to differentiate speech and non-speech signal.

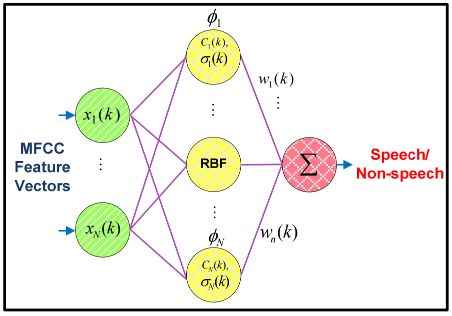


Figure 3. The architecture of the RBF NN speech/non-speech classifier

The first layer would be the input vector, $X(k)=[x_1(k),x_2(k),\ldots,x_n(k)]^T$. In this paper, the input vector would be the extracted MFCC from the previous subsection. Furthermore, the centre layer of the RBF NN includes the parameters of RBF centre, C_n and the Gaussian width, σ_n of each RBF unit which represent its corresponding subclass. On the other hand, the third layer is known as the output of the RBF NN as equated below [21]:

$$y(k)=W^T(k)\Theta(k) \tag{5}$$

Where $W(k)=[w_1(k),w_2(k),\ldots,w_N(k)]^T$

$$\Theta(k)=[\phi_1(k),\phi_2(k),\ldots,\phi_N(k)]^T$$

The $\Theta(k)$ is known as the Gaussian function of the RBF. The above expression is equivalent as following:

$$y(k)=\sum_{n=1}^{N} w_n(k)\phi_n(k) \tag{6}$$

Where ϕ_n could be mathematically represented as:

$$\phi_n(k)=\exp\left(-\frac{\|X(k)-C_n\|}{\sigma_n^2}\right) \qquad n=1,2,\ldots,N \tag{7}$$

The classification using RBF NN is based on the distances between the input and the centre values of each subclass. Therefore, from (7), $\|\cdot\|$ is denoted as the Euclidean norm of the input, while C_n is the RBF centre and the σ_n is the Gaussian width of the n^{th} RBF unit.

One of the outputs of the RBF NN speech/non-speech classification is demonstrated as in Figure 4 below. The windowed signal is marked as zero for the non-speech signal while noted as one for the speech signal. Additionally, if the interchange of zero-to-one, meaning from non-speech to speech or vice versa is triggered, the interchange windows which is within the boxes labeled in Figure 4 would be sent for further signal starting/ending point localization as discussed in the coming sub-section. The zoom in versions of the interchange signal labeled as starting and ending in Figure 4 are shown as Figure 5 and Figure 6 respectively.

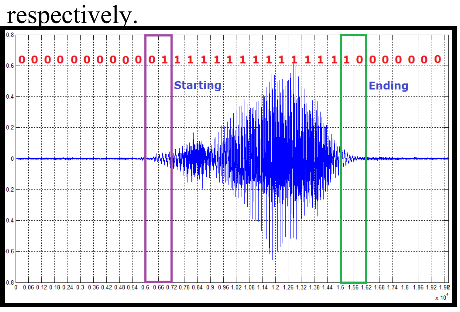


Figure 4. The output of the RBF NN speech/non-speech classification

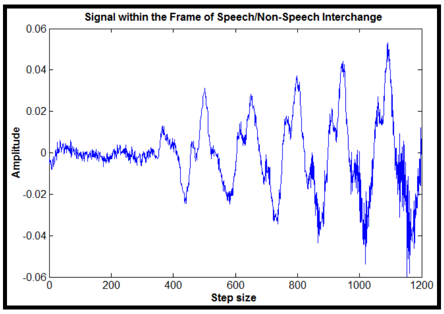


Figure 5. The interchange signal labelled as starting in Figure 4.

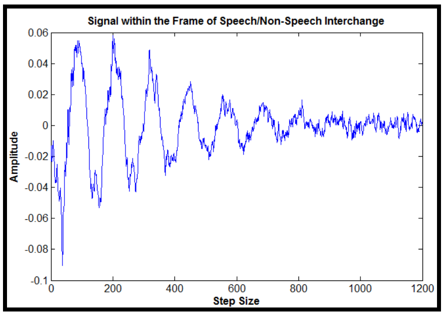


Figure 6. The interchange signal labelled as ending in Figure 4.

C. *Continuous Wavelet Transfrom Speech Localization*

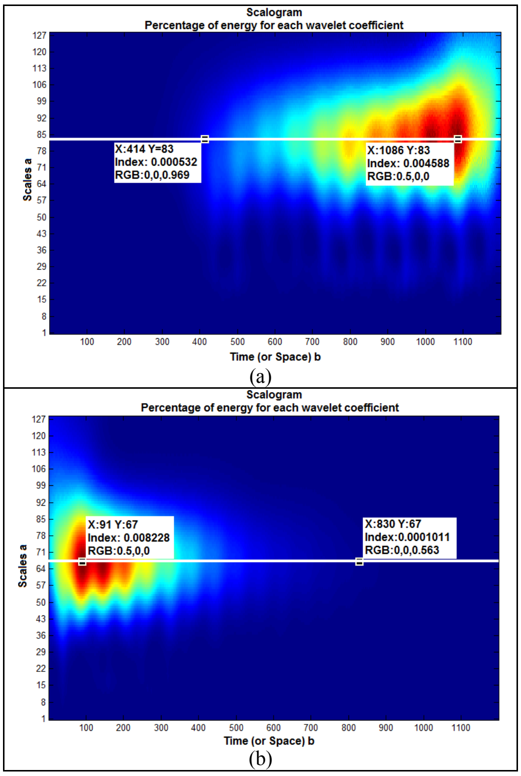
After obtaining the speech and non-speech classification of the windowed signal, the frame which has the interchange of speech to non-speech will be captured for the energy calculation as discussed in this sub-section. At first, the interchange frame is sent for the CWT to obtain its coefficients. The CWT possesses the characteristic of illustrating the frequency contents of the sound source as a function of time. Generally, the CWT would be defined as the sum over all time of the signal, $f(t)$ multiplied by the scaled and shifted versions of the wavelet function ψ [22]:

$$Coeff\,(scale,translatio\,n)=\int_{-\infty}^{+\infty} y(t)\psi(scale,translatio\,n,t)dt \tag{8}$$

According to the equation shown above, it is denoted as the inner multiplying of a family of wavelet $\psi_{\vartheta,\varphi}(t)$ with the signal, $y(t)$ as following:

$$Coeff(\vartheta,\varphi)=\int_{-\infty}^{+\infty}\frac{1}{\sqrt{\vartheta}}\overline{\psi}\left(\frac{t-\varphi}{\vartheta}\right)y(t)dt \qquad (9)$$

Where ϑ and φ are the respective parameter for scale and translation, $\overline{\psi}$ known as the complex conjugate of the ψ, and the $Coeff$, denoted as the wavelet coefficients are the output of the CWT.



(a)

(b)

Figure 7. The Scalogram of the (a) non-speech to speech and (b) speech to non-speech interchange

Subsequently, with the CTW coefficients, the percentage of energy for each of the coefficient is computed as below:

$$P = abs(coeff \times coeff) \qquad (10)$$

$$EnergyPercentage = \frac{P}{sumP} \times 100\% \qquad (11)$$

The aforementioned percentage of the energy for the CWT coefficients could be graphically represented by the scalogram as in Figure 7(a) and (b). The respective scalograms represent the energy changes of the interchange for signal starting and ending point from Figure 5 and Figure 6 in the previous sub-section.

From the scalograms shown, the position where the maximum energy occurred is first found. Subsequently, according to the energy which gradually reduces as the contour spreading from the maximum point, the coordination where the energy reaches the predefined threshold located at the same horizontal position as the maximum energy is noted as the starting/ending point speech signal. The starting and ending points are plotted on the respective scalograms. Finally, the resultant of the speech signal localization is demonstrated as in Figure 8 below. The coordination obtained from the energy percentage of the CWT coefficients as shown in Figure 7 (a) and (b) would be the respective starting and ending points after the conversion to the time step base.

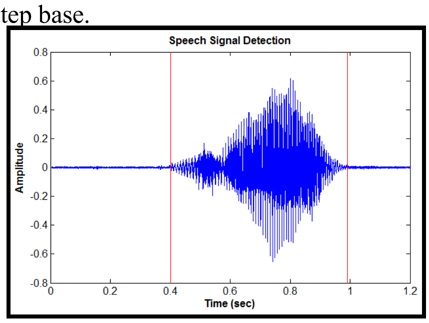
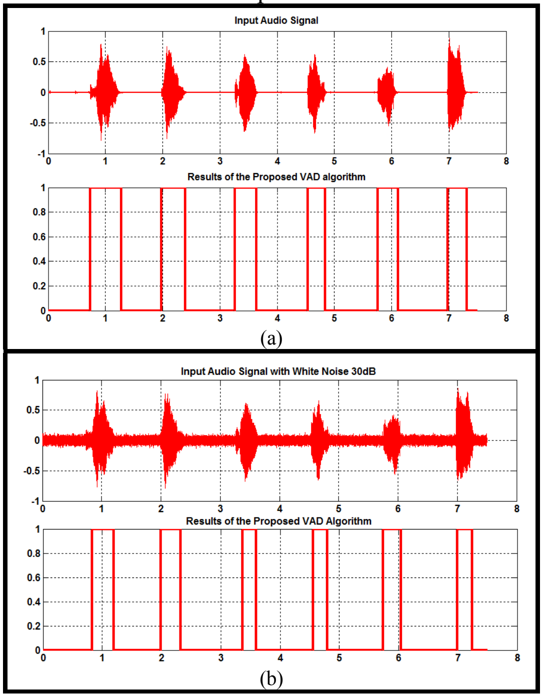


Figure 8. The final speech localization

III. Simulations and Analysis

The system evaluations are mainly focus on the speech/non-speech discrimination at different SNR levels. Furthermore, the influence of the VAD decision on the performance of the subsequent ASR system would also been analyzed. The proposed RBF-CWT VAD algorithm is evaluated by applying CUAVE database from Clemson University [23]. From the CUAVE database, 36 speakers utter continuous, connected and isolated digits are available. For the speech/non-speech classification training material, the MFCC is extracted from 20 speakers with the total speech length of 260s. In order to generate the noisy environment, the white noise with the SNR range from 30dB to 0dB is added to the clean speech data.



(a)

(b)

Figure 9. The results of the proposed VAD algorithm under (a) the clean condition (b) the white noise 30dB (c) the white noise 10dB

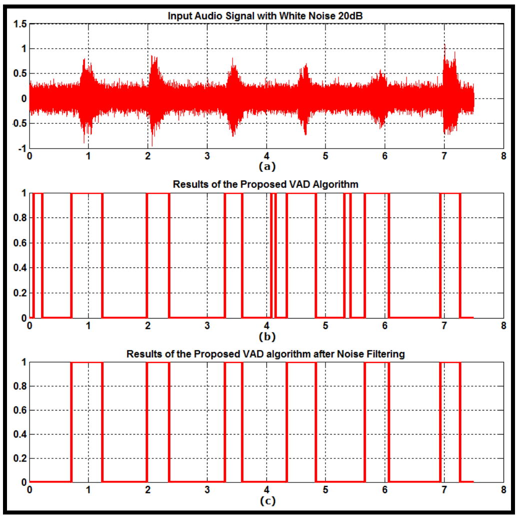


Figure 10. (a) Input Audio Signal with the white noise 20dB (b) the results of the proposed VAD algorithm with some undesired false detection (c) the final results of the proposed VAD algorithm after noise filtering

Some of the simulation results of the proposed RBF-CWT VAD approach in different level of SNR are presented. Figure 9 (a) demonstrates the RBF-CWT VAD speech detection results under the clean condition while Figure 9(b) shows the VAD results under the distorted signal with white noise 30dB. Another set of RBF-CWT VAD results are shown in Figure 10. In Figure 10(b), it is noticed that some of the false speech detection occurred, however the undesired false detection could be eliminated by applying the simple noise filtering, whichever detected signal is not exceeded to a fixed signal length, it would be considered as the non-speech signal as illustrated in Figure 10(c).

As to analyze the performance of the proposed RBF-CWT VAD algorithm in terms of speech/non-speech discrimination, the clean uttered digit from the CUAVE database is used as the reference decision. For the reference decision, each utterance is manually hand-marked as either the speech or non-speech frames. The performance of the speech detection with respect to the function of SNR is investigated in terms of the speech hit-rate (SHR) and the non-speech hit-rate (NSHR) [6]. The respective hit-rate is computed as below [6]:

$$SHR = \frac{N_{speech}}{N_{speech}^{ref}} \tag{12}$$

$$NSHR = \frac{N_{non-speech}}{N_{non-speech}^{ref}} \tag{13}$$

Where N_{speech} and $N_{non-speech}$ are the respective number of speech and non-speech frames which been correctly classified; while N_{speech}^{ref} and $N_{non-speech}^{ref}$ are the number of reference speech and non-speech respectively.

The SHR and NSHR for the proposed RBF-CWT VAD versus the (i) SVM binary classifier [16] and (ii) the conventional zero-cross rate and energy [7] are plotted in Figure 11 and Figure 12 respectively. From the plotted graphs illustrated, the proposed RBF-CWT VAD algorithm possesses a higher capability of detecting the speech signal. On the other hand, the false detection of the non-speech signal from the non-speech hit-rate graph is reduced by the proposed RBF-CWT VAD algorithm compared to others approaches. An obvious improvement of the system performance could be observed under the noisy circumstances, i.e. under the lower SNR condition.

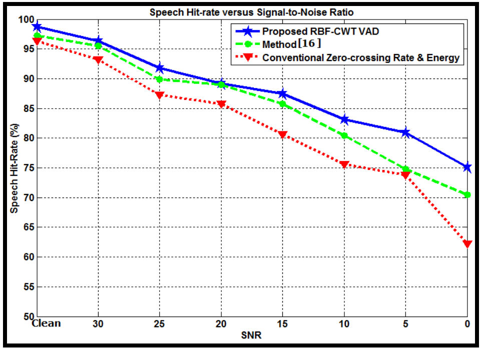


Figure 11. Speech Hit-rate comparison among the proposed VAD algorithm and other recent reported VAD approaches

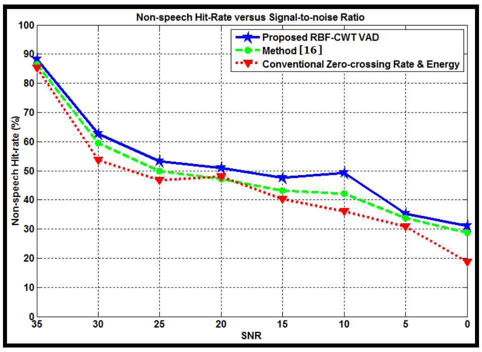


Figure 12. Non-speech Hit-rate comparison among the proposed VAD algorithm and other recent reported VAD approaches

Furthermore, the evaluation of the ASR system with various kind of VAD approach is as well presented. For the experiment setup, the training data contains 1500 utterances while testing data consists of 300 utterances. For the ASR back-end processing which consists of the speech classification, hidden Markov model (HMM) [24] is employed as the classification tool as illustrated in Figure 1 to model each word. The following ASR performance analysis is based on the output of the front-end processing (i) proposed RBF-CWT VAD (ii) zero-crossing rate and energy [7] and (iii) the SVM binary classifier presented in [16]. The average word accuracy under different SNR levels is tabulated in Table 1. From the tabulated results, it is noticed that the performance of the pre-processing, which is, in this case VAD algorithm would have an impact on the performance of the subsequent ASR. Compared to others VAD algorithms, the improved RBF-CWT VAD approached presented in this paper has demonstrated a better speech recognition rate either under the clean or noisy circumstances.

TABLE I. PERFORMANCE COMPARISON OF THE PROPOSED RBF-CWT VAD TO OTHER RECENTLY REPORTED ALGORITHMS

SNR (dB)	Proposed VAD (%)	SVM binary classifier (%)	Zero-crossing rate and energy (%)
Clean	95.72	95.10	94.58
30	93.81	91.92	91.28
25	89.66	84.37	83.72
20	84.02	82.63	83.04
15	82.38	79.85	77.66
10	67.57	61.29	60.51
5	28.13	23.74	20.95
0	15.24	10.82	9.63

IV. CONCLUSION

An improved VAD with the RBF NN and the CWT for ASR is proposed in this paper. The improved RBF-CWT VAD successfully detects the speech signal from the input audio signal even with the presence of noisy background. With the aid of the window-level MFCC as the training material for the RBF NN and the computation of the energy changes of the CWT coefficients, the proposed RBF-CWT VAD algorithm offers a better noise compensation and provides more precise speech/non-speech detection compared to the conventional frame-level binary classification approach. Furthermore, the proposed RBF-CWT VAD approach is also shows a more accurate detection compared to the zero-crossing rate and short-term signal energy algorithms not even under the clean condition but as well the noisy circumstances. With a higher capability of speech/non-speech detection offers, the final ASR system has shown a higher recognition rate with the integration of the proposed RBF-CWT VAD algorithm.

REFERENCES

[1]J. Ramirez, *et al.*, "Improved Voice Activity Detection Using Contextual Multiple Hypothesis Testing for Robust Speech Recognition," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 15, pp. 2177-2189, 2007.

[2]J. M. Górriz, *et al.*, "Improved likelihood ratio test based voice activity detector applied to speech recognition," *Speech Communication,* vol. 52, pp. 664-677.

[3]M. W. Hoffman, *et al.*, "GSC-based spatial voice activity detection for enhanced speech coding in the presence of competing speech," *Speech and Audio Processing, IEEE Transactions on,* vol. 9, pp. 175-178, 2001.

[4]S. M. ERINNOVIAR, HAYASHI S, "Speech Enhancement with Voice Activity Detection in Subbands," *Bulletin of Science and Engineering, Takushoku University,* vol. 7, pp. 49-54, 2000.

[5]K. Srinivasan and A. Gersho, "Voice activity detection for cellular networks," in *Speech Coding for Telecommunications, 1993. Proceedings., IEEE Workshop on*, 1993, pp. 85-86.

[6]J. R. a. J. M. G. a. J. C. Segura, Ed., *Voice Activity Detection. Fundamentals and Speech Recognition System Robustness* (Robust Speech Recognition and Understanding. 2007, p.^pp. Pages.

[7]L. R. Rabiner and M. R. Sambur, "Algorithm for determining the endpoints of isolated utterances," *The Journal of the Acoustical Society of America,* vol. 56, p. S31, 1974.

[8]R. G. Bachu, *et al.*, "Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy," in *Advanced Techniques in Computing Sciences and Software Engineering*, K. Elleithy, Ed., ed: Springer Netherlands, 2010, pp. 279-282.

[9]K. Ishizuka, *et al.*, "Noise robust voice activity detection based on periodic to aperiodic component ratio," *Speech Communication,* vol. 52, pp. 41-60, 2010.

[10]A. Davis, *et al.*, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 14, pp. 412-424, 2006.

[11]E. Nemer, *et al.*, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *Speech and Audio Processing, IEEE Transactions on,* vol. 9, pp. 217-231, 2001.

[12]J. Ramírez, *et al.*, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication,* vol. 42, pp. 271-287, 2004.

[13]C. Joon-Hyuk, *et al.*, "Voice activity detection based on multiple statistical models," *Signal Processing, IEEE Transactions on,* vol. 54, pp. 1965-1976, 2006.

[14]Q. H. Jo, *et al.*, "Statistical model-based voice activity detection using support vector machine," *Signal Processing, IET,* vol. 3, pp. 205-210, 2009.

[15]J. W. Shin, *et al.*, "Voice activity detection based on statistical models and machine learning approaches," *Computer Speech & Language,* vol. 24, pp. 515-530, 2010.

[16]E. C. Tomi Kinnunen, M. Tuononen,P.Franti, H. Li,, "Voice Activity detection Using MFCC Features and Support Vector Machine," in *SPECOM* 2007.

[17]K. P. S. Siew Wen Chin, Li-Minn Ang, King Hann Lim, "Fractional Cepstral Normalization (FCN) for Robust Speech Recognition System," presented at the Int. Conf. on Embedded Sys. and Intelligent Tech. (ICESIT2010), 2010.

[18]J. V. S. S. Stevens, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *Journal of the Acoustical Society of America,* vol. 8, no.3, pp. 185-190, 1937.

[19]J. K. I. Shah, A.N. , "Robust voice/unvoiced classification using novel featuresand Gaussian Mixture Model," Temple University, Philadelphia, USA2004.

[20]S. Young, "The HTK book: for HTK version 2.1," ed: Cambridge, England: Cambridge University Press., 1997.

[21]P. Seng Kah and L. M. Ang, "Adaptive RBF Neural Network Training Algorithm For Nonlinear And Nonstationary Signal," in *Computational Intelligence and Security, 2006 International Conference on*, 2006, pp. 433-436.

[22]S. Sinha, Routh, P. S., Anno, P. D., and Castagna, J. P, "Spectral decomposition of seismic data with continuous-wavelet transforms," *Geophysics 70,* pp. 19-25, 2005.

[23]E. Patterson, *et al.*, "CUAVE: a new audio-visual database for multimodal human-computer interface research," in *Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on*, 2002, pp. 2017-2020.

[24]"HMM toolbox, available at http://www.freewebs.com/lvtaoran/HMMall.rar."