1-D LOCAL BINARY PATTERNS BASED VAD USED INHMM-BASED IMPROVED SPEECH RECOGNITION

Qiming Zhu, Navin Chatlani and John J. Soraghan

Centre for Excellence in Signal and Image Processing (CeSIP), University of Strathclyde
Royal College Building, 204 George Street, Glasgow
E-mail: qiming.zhu@eee.strath.ac.uk, navin.chatlani@eee.strath.ac.uk, j.soraghan@eee.strath.ac.uk

ABSTRACT

In this paper, 1-D Local binary patterns (LBP) are proposed to be used in speech signal segmentation and voice activation detection (VAD)and combined with hidden Markov model (HMM) for advanced speech recognition. Speech is firstly de-noised by Adaptive Empirical Model Decomposition (AEMD), and then processed using LBP based VAD. The short-time energy of the speech activity detected from the VAD is finally smoothed and used as the input of the HMM recognition process. The enhanced performance of the proposed system for speech recognition is compared with other VAD techniques at different SNRs ranging from 15 dB to a robust noisy condition at -5 dB.

Index Terms –Speech Enhancement using Adaptive Empirical Model Decomposition (AEMD), Local Binary Patterns, Voice Activity Detection, Noise Reduction, Hidden Markov Model

1. INTRODUCTION

Automatic speech recognition (ASR)involves the automatic processing of speech-to-text which converts human voice signals into written words. It has a variety of applications, such as speech-to-speech translation, speech command and control, dialog system, etc.

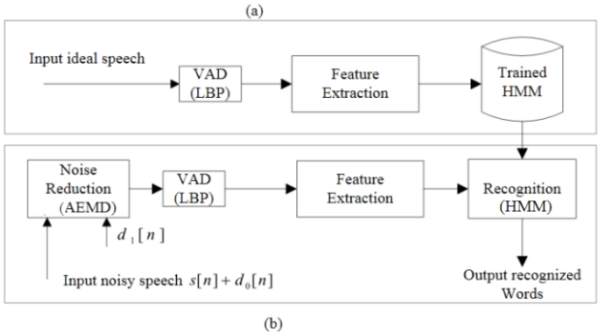

Fig. 1 Noisy speech recognition system
(a) is the training process,(b) is the recognition process

Hidden Markov Model (HMM) based ASR systems have provided high performance with good quality speech input. However, ASR is commonly performed in noisy environments, where a noise reference signal is available. In Fig. 1 (b), such an ASR system is depicted where $d_1[n]$ is a measure of the noise reference. The proposed system illustrated in Fig. 1 shows the voice activity detection (VAD) and noise reduction stages. In Fig. 1,$s[n]$ is the original speech and $d_0[n]$ is the contaminating noise. The composition of these two signals is used to simulate the speech in robust noise background.$d_1[n]$is the reference noise for AEMD process, which could be recorded by an additional microphone or estimated by noise estimation programs.

Noise reduction techniques in ASR systems are used to provide the inputs for recognition process. The main algorithms used include spectral subtraction, cross-correlation, microphone array beam-former and adaptive noise cancellation (ANC). ANC algorithms provide high performance in enclosed spaces speech recognition, such as in cars [22], flights, enclosed rooms, etc. Recently, Empirical Mode Decomposition (EMD) [2] has been applied in speech enhancement such as [3] [4] and [1] [5] [6]. EMD decomposes the signal into data-adaptive functions known as: Intrinsic Mode Functions (IMFs) which can be used to perform the adaptive noise cancellation. Speech enhancement using adaptive EMD (AEMD) [7] was proposed as an improvement to basic ANC. AEMD refines the IMFs by using an adaptive filter and the reference noise $d_1[\mathrm{n}]$. The enhanced IMFs are then used to reconstruct the speech signal which will be the input of the VAD in the system.

The purpose of the VAD is to find the start-points and end-points of a voiced speech segment so that it is distinguished from noise, unvoiced or mute segments. An effective VAD of speech recognition systems can improve the ASR performance. The traditional VAD algorithms are based on the short-time energy levels, short-time average zero-crossing rate (ZCR) which was firstly proposed by Junqua in 1991 [8]. Chatlani et al first proposed local binary patterns (LBP) for 1-D signal processing which was shown to provide improved VAD performance [10].The performance of this VAD algorithm was compared to the

ITU-T standard G. 729 Annex B [9] with favourable results. In this paper EMD and 1-D LBP are combined with a HMM to form a new speech recognition system.

The remainder of the paper is organised as follows. Section 2 describes the main algorithms used in the proposed speech recognition system. Section 3 provides simulation results. Conclusions are provided in Section 4.

2. ALGORITHMS

2.1 Noise Reduction (AEMD)

The AEMD for signal enhancement is illustrated in Fig. 2. The noisy speech contains the original speech $s[n]$ and the contaminating noise $d_0[n]$. $d_1[n]$ is the reference noise. The noisy speech signal is first decomposed into IMFs: $I_1[n], I_2[n], \ldots I_j[n] \ldots, I_N[n]$, where $I_j[n]$ is the j[th] IMF. These IMFs are mixtures of speech $s_j[n]$ and noise $d_{0,j}[n]$. The resulting IMFs are adaptively filtered using the reference noise. The enhanced speech signal is reconstructed by using the outputs $y_j[n]$ from the adaptive filter, where w_j is the filter coefficient.

It is assumed in Fig. 2 that the contaminating noise is correlated to the reference and the original signal is uncorrelated with these noises. Therefore, the j[th] IMF of the noisy signal can be described below by:

$$I_j[n] = s_j[n] + d_{0,j}[n]. \quad (1)$$

The error signal $e_j[n]$ is given by:

$$e_j[n] = I_j[n] - y_j[n]. \quad (2)$$

The error signal $e_j[n]$ is given by:

$$e_j[n] = I_j[n] - y_j[n]. \quad (2)$$

The mean square error (MSE) is evaluated as:

$$E[e^2{}_j[n]] = E[(I_j[n] - y_j[n])^2]. \quad (3)$$

$$E[e^2{}_j[n]] = E[s^2{}_j[n]] + E[(d_{0,j}[n] - y_j[n])^2] + 2E\big[s_j[n]\big(d_{0,j}[n] - y_j[n]\big)\big]. \quad (4)$$

To minimize the MSE signal, the adaptive filter updates w_j. As $d_{0,j}[n]$ is uncorrelated with $s_j[n]$, the minimum MSE can be presented by:

$$E_{min}[e^2{}_j[n]] = E[s^2{}_j[n]] + E[(d_{0,j}[n] - y_j[n])^2]. \quad (5)$$

During this process, the noise is adaptively filtered to feed into the system, to perform an uncorrelated error signal. Hence, the enhanced signal $\hat{s}[n]$ which is reconstructed by the enhanced IMF $\widehat{I_j}[n]$ is shown is Equation (6):

$$\hat{s}[n] = \sum_{j=1}^{N} \widehat{I_j}[n]. \quad (6)$$

This reconstructed signal will then be used as the input of the VAD in the system.

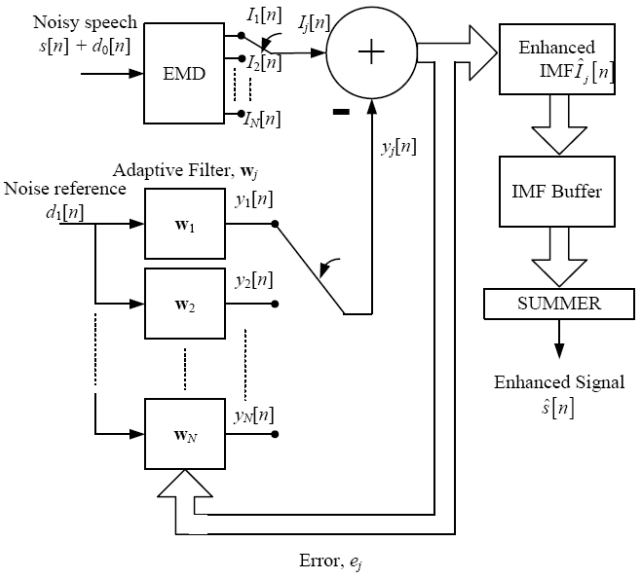

Fig. 2. AEMD model

2.2 Voice Activity Detection (1-D LBP)

The aim of voice activity detection is to make judgments between unvoiced sound and voice based on the different characteristics [11]. As the characteristics of unvoiced sound and noise are very similar, the noise is usually treated as unvoiced sound in speech recognition. Therefore, the unvoiced/voiced sound detection algorithm used commonly in normal environment. A generalized VAD procedure is summarized in Fig.3 [12]. Features of the input speech signal are calculated and intermediate decisions are made following the VAD rules that are described below.

Fig. 3 VAD procedure

LBP have been widely used in 2-D image processing [13][14]which demonstrate it as a simple, discriminative descriptor of texture in [15]. Chatlani et al [10][21]proposed 1-D LBP with application to VAD.

The 1-D LBP obtains the neighbour data samples form a signal $x[i]$. After thresholding the centre samples against the neighbouring samples, an LBP code will be allocated to each of them. The 1-D LBP operating of a sample value $x[i]$ can be defined as:

$$LBP_P(x[i]) = \sum_{r=0}^{\frac{P}{2}-1} \Big\{ s\Big[x\Big[i + r - \tfrac{P}{2}\Big] - x[i]\Big] 2^r + s\Big[x[i + r + 1] - x[i] 2^{r+\frac{P}{2}}\Big] \Big\}. \quad (7)$$

where the Sign function is:

$$S[x] = \begin{cases} 1, & \text{for } x \geq 0 \\ 0, & \text{for } x < 0 \end{cases}. \quad (8)$$

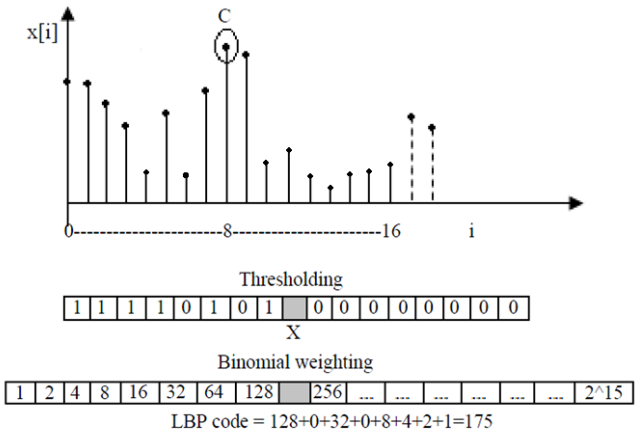

Fig. 4 Computation of 1-D local binary pattern for P=16 (LBP_{16})

the signal $x[i]$ of length N for $i = [\frac{P}{2} : N - \frac{P}{2}]$ is used to form the P neighbour samples from the centre sample. The Sign function performs a P-bit binary code for the differences.

Fig. 4 illustrates a P=16 1-D LBP operator with the circled centre sample C is given. From (7), a binary code of 11110101_00000000 is produced by the 16 neighbouring samples after thresholding against the centre sample C. The LBP code of 175 is formed by summing the binary code multiplying by the binomial weights. The LBP code can describe the data using the differences between a sample and its neighbours locally [10]. The distribution of the LBP codes also can be used to describe the local patterns of $x[i]$:

$$H_k = \sum_{\frac{P}{2} \leq i \leq N - \frac{P}{2}} \delta(LBP_P(x[i]), k). \quad (9)$$

where k=1...n and n is the number of histogram bins and each bin corresponds to an LBP code. $\delta(i, j)$ is the Kronecker delta function.

As described in [10], for a constant or slowly varying signal, the differences between a sample and its neighbours cluster near zero. At peaks and troughs, especially at edges these differences will be larger. This provides a possibility for using LBP to detect the voiced and unvoiced signal.

The performance of the LBP based VAD has been compared to the G.729B VAD standard [10]. However, to make this VAD less complex with less noisy input, a short-time energy and short-time zero crossing rate (ZCR) based algorithm is applied to the VAD smoothing before the decision is made. This VAD algorithm is applied to the HMM based speech recognition system.

2.3 Speech Feature Extraction

Unlike Linear prediction coefficient (LPC), Mel frequency cepstral coefficient (MFCC) does not suffer from variations in the amplitude of the speech signal due to noise [3]. MFCC generates the training vectors by transforming the signal into frequency domain and is less prone to noise. MFCC is preferred as the feature method in this paper over the LPC in capturing the significant acoustic information [16].

2.4 Speech Recognition Algorithm (HMM)

Since its introduction in the 1980s, Hidden Markov Model (HMM) has become the dominant statistical modelling tool used for ASR. Rabiner et al [17] provides a description of HMM based speech processing while other researchers improved the technology to achieve the higher performance [18][19]. Researchers [20] developed the Hidden Markov Model Toolkit (HTK), which is widely applied is ASR.

As proposed by Rabiner et al [17], a HMM is characterized by the following:

- The number of states in the model N. And the states can be denoted as $S = \{s_1, s_2, \ldots s_N\}$, the state at time t is q_t.
- The number of observation symbols per state M. The observation symbol can be presented as $V = \{V, V_2, \ldots V_M\}$.
- The state transition probability $A = \{a_{ij}\}$, where

$$a_{ij} = P\{q_{t+1} = s_j | q_t = s_i\};\ 1 \leq i, 1 \leq j. \quad (10)$$

- The observation probability in stated j, $B = \{b_j(k)\}$, where

$$b_j(k) = P\{O_t = V_k | q_t = s_i\};\ 1 \leq j \leq N, 1 \leq k \leq M. \quad (11)$$

- The initial state probability $\pi = \{\pi_i\}$, where

$$\pi_i = P[q_1 = s_i];\quad i \leq j \leq N. \quad (12)$$

There are three main parts of an HMM that involves evaluation, decoding and learning [17]. Assume a HMM $\lambda = (\pi, A, B)$ and an observed sequence $O = \{O_1, O_2, \ldots O_T\}$ has been given. 'Evaluation' means to calculate the probability of this observed sequence from the model; 'decoding' means to computer the maximum likelihood (ML) and the maximum likely state from the model; 'learning' means to adjust the probability matrix 'A' and 'B' so as to best describe how a given observation sequence comes about.

HMM in speech recognition is possible given the solution to these three problems. Learning and decoding can be used to estimate the suitable HMM for the given training data, evaluation can be applied to recognize the test data from the HMM. HMM based system uses the saved HMMs as the reference library, input the test speech and calculate the probabilities of the test speech from each of the HMMs, choose the maximum one as the recognized speech.

Left-to-right HMM, which means the states could only be transmitted from the previous states, had been verified to be the most suitable HMM for speech recognition. It is selected to build up HMM reference temple which is shown

in Fig. 1 (a) and used as the recognition method which is shown in Fig. 1 (b).

3. EXPERIMENTAL RESULTS AND DISSCUSSION

The numbers from 0 to 9 were spoken and recorded a total of 10 times by 5 males with the sampling frequency *fs*=16 kHz for HMM training. Additionally 5 records from 0 to 9 by the same 5 people were used to test the isolated word recognition performance. 12 recorded sentences which contain 7 different words are used to test the continuous speech recognition performance. All these speech signals are used to perform new noisy speech by adding babble noise at different signal noise rates (SNR). A sampling frequency of 16 *kHz* babble noise from the Noisex-92 database is used as the reference noise.

A high-pass FIR filter of order 40 and a normalised lower cut-off frequency of 0.05 is designed to filter the reference noise $d_1[n]$ in order to generate the contaminating noise $d_0[n]$. A 41 tap FIR adaptive filter with a standard NLMS learning algorithm was used as the adaptive filter in AEMD. The length of the segments for AEMD is W=10ms, which is the same as the length for VADs segment (LBP, G.729, short-time energy and zero crossing rates).

The blocking feature for MFCC is N=256 and M=128, which means the signal is blocked into frames of 256 samples with the adjacent frames being separated by 128.

The HMM is initialized with the state number N=6.The initial state probability $\pi_i = [1\ 0\ 0\ 0\ 0\ 0]$, the initial transition probability A is:

$$A = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Fig. 5 shows the VAD result of a continuous speech signal with the SNR at 0 dB. The signal is first input to the AEMD to reduce the noise. The reconstructed signal is then processed by LBP. LBP makes the VAD decision after smoothing, the solid lines denote the start-point of the voiced segments and the dotted lines denote the end-point of the voiced segments. As Fig. 5 shows, LBP based VAD picks out the speech segments successfully.

The LBP based VAD for speech recognition results will be compared with short-time energy and short-time zero crossing rates (refer to VAD1), G.729B (refer to G.729) at different SNRs from 15 dB to a robust noisy condition at -5 dB.

The comparison results are shown in Table I and Table II. Table I provides the isolated word recognition results which contains the noisy speech and noiseless speech recognition. It can be observed in Table I that LBP has the highest performance in HMM based noisy speech recognition and it also provides a high performance for noiseless speech at the recognition rate of 98%.It should be noticed that LBP still provides the best recognition rate in low SNR, the recognition rate is 94% at -5 dB.

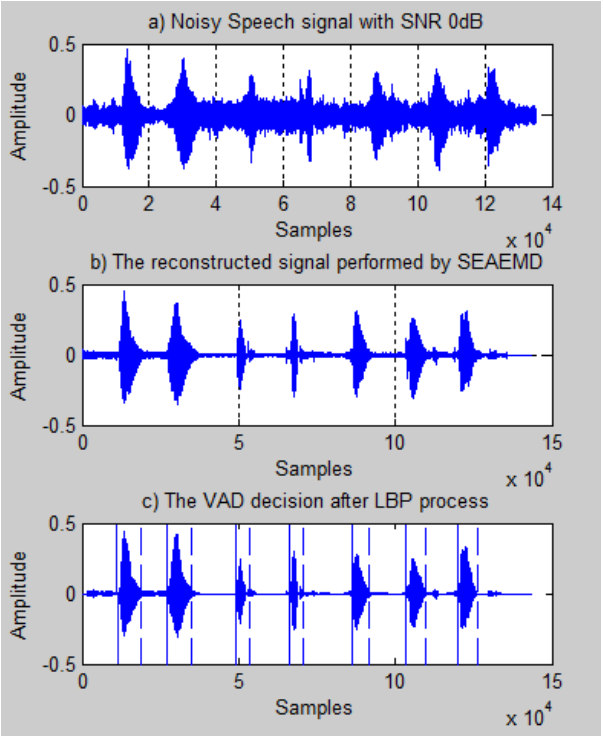

Fig. 5 The LBP VAD results
(a) the input noisy speech signal, (b) the reconstructed signal performed by AEMD, (c) the VAD decision after LBP process

TABLE I
Isolated word recognition rate

SNR(dB)	-5	0	5	10	15	No noise
AEMD+LBP	94 %	96 %	98 %	98 %	98 %	98%
AEMD+VAD1	84 %	90 %	92 %	94 %	96 %	98%
AEMD+G.729	90 %	92 %	96 %	96 %	98 %	98%

TABLE II
Continuous speech recognition rate

SNR(dB)	-5	0	5	10	15	No noise
AEMD+LBP	92.8 %	94.0 %	94.0 %	95.2 %	96.4 %	96.4 %
AEMD+VAD1	76.2 %	79.8 %	82.1 %	85.7 %	89.2 %	94.0 %
AEMD+G.729	79.8 %	83.3 %	86.9 %	91.7 %	92.8 %	97.6 %

The continuous speech recognition results are shown in Table II. From this table it is seem that, LBP provides the highest performance for noisy speech input and the recognition rate at –5 dB is 92.8%. LBP still provides a good recognition rate of 96.4% with the noiseless input.

4. CONCLUSION

This paper proposed a new speech recognition system that combines AEMD, 1-D LBP and HMM. The output de-noised signals from AEMD is used as the input of the 1-D LBP based VAD. The detected voiced segments are recognized by HMM.

The experimental results show that 1-D LBP can distinguish the voiced and unvoiced components of speech signals [10]. The LBP is shown to be superior to the G.729 VAD and short-time energy VAD by comparison results in HMM based noisy speech recognition. It performs higher recognition rates for robust noisy speech input.

However, the experiments are based on a known noise reference. To make the algorithm more efficient, an independent noise estimation program should be incorporate. This will be the focus in the future work.

5. REFERENCE

[1] N. Chatlani, J. J. Soraghan "EMD-based filtering (EMDF) of low-frequency noise for speech enhancement", *IEEE Transactions on Audio, Speech, and Language Processing, vol. 20,* pp. 1158-1166, 2012.

[2] N. E. Huang et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Royal Society London A*, vol. 454, pp. 903-995, 1998.

[3] A. O. Boudraa, J-C.Cexus, "EMD-based signal filtering", *IEEE Transactions on Instrumentation and Measurement*, vol.56, no.6, pp.2196-2202, Dec. 2007.

[4] X. Zou, X. Li, R. Zhang, "Speech enhancement based on Hilbert-Huang transform theory", *IEEE CS Proceeding of the First International Multi-Symposium of Computer and Computational Sciences (IMSCCS|06)*, vol. 1, pp. 208-213, June 20-24, 2006.

[5] P. Flandrin, P. Goncalves and G. Rilling, "Detrending and denoising with empirical mode decompositions", *Proc. EUSIPCO 2004*, pp. 1581-1584, 2004.

[6] Z-F Liu, Z-P Liao, E-F Sang, "Speech enhancement based on Hilbert-Huang transform", *Proceedings of 2005 International Conference on Machine Learning and Cybernetics*, vol. 8, pp. 4908-4912, 18-21 Aug. 2005.

[7] N. Chatlani, J. J. Soraghan, "Speech enhancement using adaptive empirical mode decomposition",*16th International Conference on Digital Sig. Proc (DSP 2009)*, Santorini, Greece, pp. 1-6, July 2009.

[8] J. C. Junqua, B. Mark, B. Reaves, "A study of endpoint detection algorithms in adverse conditions: incidence on a DTW and HMM recognize," *Proc. Euro speech 1991*, pp. 1371–1374, 1991.

[9] ITU-T, "A silence compression scheme for G.729 optimized for terminals conforming to ITU-T V.70 Annex B," *ITU-T Recommendation G.729-Annex B*,1996.

[10] N. Chatlani, J. J. Soraghan, "Local binary patterns for 1-D signal processing", *18th European Signal Proc. Conference (EUSIPCO)*, Aalborg, Denmark, pp. 95-99 Aug. 2010.

[11] X. Yang, B. Tan, J. Ding, J, Zhang and J. Gong, "Comparative study on voice activity detection algorithm," *International Conference on Electrical and Control Engineering*, pp. 599-602, 2010.

[12] J. C. Junqua, B. Mark, B. Reaves, "A robust algorithm for word boundary detection in the presence of noise", *IEEE Transactions on Speech and Audio Processing*, 2, pp. 406–412, 1994.

[13] T. Ojala, M. Pietikainen, "Unsupervised texture segmentation using feature distributions", *Pattern Recognition 32*, pp. 477-486, 1999.

[14] S. He, J. J. Soraghan et al, "Quantitative analysis of facial paralysis using local binary patterns in biomedical videos", *IEEE Transactions on Biomedical Engineering, vol. 56(7)*, pp. 1864-1870, Jul 2009.

[15] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multi-resolution grayscale and rotation invariant texture analysis with local binary patterns", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24(7), pp. 971-987, 2002.

[16] Davis, S. B. and Mermelstein, P. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. On Acoustic, Speech and Signal Processing*, ASSP-28, No. 4, 1980.

[17] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proceedings of IEEE*, Vol. 77, No.2, pp. 257-286, Feb. 1989.

[18] X. D. Huang, Y. Ariki and M. A. Jack, "Hidden Markov Models for speech recognition," *Edinburgh University Press*, 1990.

[19] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Simultaneous modelling of spectrum, pitch and duration in HMM-Based speech synthesis," *Proc. EUROSPEECH-99*, pp. 2374-2350, Sep. 1999.

[20] S. Young, G. Everman, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, "The HTK Handbook" *Cambridge University Engineering Department*, 2009.

[21] N. Chatlani, "Advanced signal enhancement techniques with application to speech and hearing,"*University of Strathclyde PhD Thesis*, 2011.

[22] Y. Cho and H. Ko, "Speech enhancement for robust speech recognition in car environments using Griffiths-Jim ANC based on two-paired microphones", *IEEE International Symposium on Consumer Electronics,*pp. 123-127, 2004.