ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 4, April 2018

Implementation of Mask Estimation & ANN for Speech Enhancement in Non-Stationary Noise Environment

Suresh Kumar[1], Mr. Santosh Kumar[2]

Computer Science & Engineering Department, Emax Group of Colleges, Ambala, India[1,2]

Abstract: In conventional single-channel speech enhancement, typically the noisy spectral amplitude is modified while the noisy phase is used to reconstruct the enhanced signal. It provides speech enhancement under different noisy conditions. The speech power spectrum varies greatly for different types of speech sound. The energy of voiced speech sounds is concentrated in the harmonics of the fundamental frequency while that of unvoiced sounds is, in contrast, distributed across a broad range of frequencies. To identify the presence of speech energy in a noisy speech signal we have therefore developed two detection algorithms. The first is a robust algorithm that identifies voiced speech segments and estimates their fundamental frequency. The second detects the presence of sibilants and estimates their energy distribution. The use of speech enhancement algorithm removes or reduces the presence of noise. The aim of the noise reduction algorithms is to estimate the clean speech signal from the noisy recordings in order to improve the quality and intelligibility of the enhanced signal. Due to this, it presents a method for speech enhancement using mask estimation iteratively.

Keywords: Speech Enhancement, Speech Processing, Noise Filtering, Sparse Representation etc.

I. INTRODUCTION

Speech enhancement or noise reduction has been one of the main investigated problems in the speech community for a long time. The problem arises whenever a desired speech signal is degraded by some disturbing noise. The noise can be additive or convolutive. In practice, a convolutive noise should be rather considered due to the reverberation. However, it is usually assumed that the noise is additive since it makes the problem simpler and also the developed algorithms based on this assumption lead to satisfactory results in practice. Even this additive noise can reduce the quality and intelligibility of the speech signal considerably. Therefore, the aim of the noise reduction algorithms is to estimate the clean speech signal from the noisy recordings in order to improve the quality and intelligibility of the enhanced signal. There are various applications of speech enhancement in our daily life.

For example, consider a mobile communication where you are located in a noisy environment, e.g., a street or inside a car. Here, a noise reduction approach can be used to make the communication easier by reducing the interfering noise. A similar approach can be used in communications over internet, such as Skype or Google Talk.

Speech enhancement algorithms can be also used to design robust speech/speaker recognition systems by reducing the mismatch between the training and testing stages. In this case, a speech enhancement approach is applied to reduce the noise before extracting a set of features [1]. Speech is the main carrier of human conversation, and speech communication is one of the fastest-growing communication business. With the development of speech signal processing technology, the evaluation of speech quality increases in importance.

The speech quality evaluation has made great achievements in speech coding, speech recognition, speech synthesis, but the research in speech enhancement is not mature. The change of speech quality caused by speech coding essentially differs from by speech enhancement, therefore, the speech quality evaluation system in speech coding field cannot be directly applied to speech enhancement.

Speech quality evaluation measures are classified into subjective and objective methods. Subjective measures best fit human feelings, and can better reflect speech quality, but they are subjected to various test conditions, which influences the reliability of results. Speech signals from the uncontrolled environment may contain degradation components along with required speech components.

The degradation components include background noise, speech from other speakers etc.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 4, April 2018

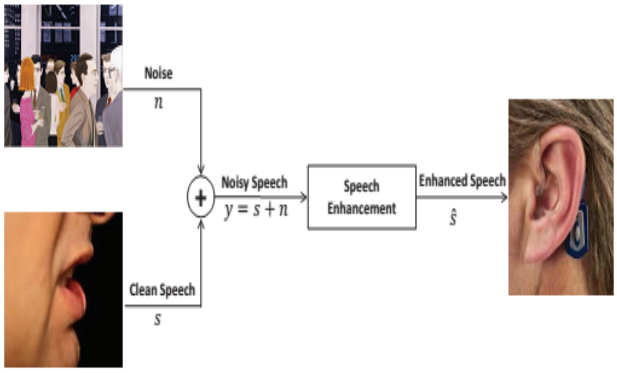

Figure 1: Speech Enhancement System with Corrupted Noise

Speech enhancement has been studied because of its many applications, such as voice communication, voiced –control systems, and the transmitted speech signals. It is a noise suppression technology which has important significance for solving the problem of noise pollution, improving the quality of voice communications, improving speech intelligibility and speech recognition rates, etc.. The objective of speech enhancement is to restore the original signal from noisy observations corrupted by various noises [1]. Speech enhancement techniques have been developed for a single microphone and multiple microphones.

In literature, some proposed a novel multi-channel speech enhancement method by combining the wiener filtering and subspace filtering with a convex combinational coefficient. It investigated a multi-channel de-noising auto-encoder (DAE)-based speech enhancement approach. In recent years, deep neural network (DNN)-based monaural speech enhancement and robust automatic speech recognition (ASR) approaches have attracted much attention due to their high performance. It also proposed a sparse hidden Markov model (HMM) based single-channel speech enhancement method that models the speech and noise gains accurately in non-stationary noise environments. Some presented a harmonic phase estimation method relying on fundamental frequency and signal-to-noise ratio (SNR) information estimated from noisy speech. The proposed method relies on SNR-based time-frequency smoothing of the unwrapped phase obtained from the decomposition of the noisy phase.

The paper is ordered as follows. In Section II, It defines the description of speech level estimation system. Section III describes the problem definition & description of proposed system. Results are explained in section IV. Finally, conclusion is explained in Section V.

II. SPEECH ESTIMATION SYSTEM MODEL

In this work we study the time-delay estimation (TDE) problem, where we want to estimate the Time Delay 'D', i.e. the problem of estimating the time delay and the correlation function between two received signals is presented [1]. A mathematical model for the two signals is introduced. We are interested in the estimation of the time-delay that the signal suffers due to the differing spatial locations of the distinct receiver from the source.

1. System Model

It considers a multi-path environment where one source and two sensors are presented; the two sensors are located at different distances from the same source. The received signal at the two microphones can be modelled as:

$$r_1(t) = s(t) + n_1(t), \qquad 0 \le t \le T \qquad r_2(t) = s(t - D) + n_2(t)$$

Where $r_1(t)$ and $r_2(t)$ are the outputs of the two microphones that are separated spatially, $s(t)$ is the source signal, $n_1(t)$ and $n_2(t)$ are representing the additive noises.'T', the observation interval ,and 'D',the time delay between the two received signals. The signal and noises are assumed to be uncorrelated having zero-mean and Gaussian distribution. Our objective is to estimate this 'D' and thus the problem 'Time Delay Estimation'.

Since Time Delay Estimation is an important technique for identifying, localizing and tracking radiation sources. Because of its central significance, accuracy and precision are of critical importance to the TDE algorithms. Since now there exist various methods and algorithms to estimate the time delay. Here we are considering the comparative study of only four methods for TDE, viz. the Cross-correlation Function (CCF) method, the Phase Transform (PHAT) Method falling under the Generalized Cross-correlation method and the Average square Difference Function (ASDF) method and the adaptive least mean square filter (LMS) methods are discussed and compared for the estimation of the

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 4, April 2018

time delay. Their simulation results are compared in terms of computational complexity, hardware implementation, precision, and accuracy.

Since the performances of the TDE methods are considerably degraded by the signal-to-noise ratio (SNR) level, this factor has been taken as a prime factor in benchmarking the different methods. The CC method cross-correlates the microphone outputs and considers the time argument that corresponds to the maximum peak in the output as the estimated time delay. To improve the peak detection and time delay estimation, various filters, or weighting functions, have been suggested to be used after the cross correlation [6]. The estimated delay is obtained by finding the time-lag that maximizes the cross-correlation between the filtered versions of the two received signals. This technique is called generalized cross-correlation (GCC). The GCC method, proposed by Knapp and Carter in 1976, is the most popular technique for TDE due to their accuracy and moderate computational complexity. The role of the filter or weighting function in GCC method is to ensure a large sharp peak in the obtained cross-correlation thus ensuring a high time delay resolution.

There are many techniques used to select the weighting function; such as the Phase Transform (PHAT), that is based on maximizing some performance criteria. These correlation-based methods yield ambiguous results when the noises at the two sensors are correlated with the desired signals. To overcome this problem, higher-order statistics methods were employed. There are also some other algorithms used to estimate the time-delay. Algorithms based on minimum error: Average Square Difference Function (ASDF) seeks position of the minimum difference between signals $r_1(t)$ and $r_2(t)$ [6]. Adaptive algorithms such as LMS can also be introduced into the TDE [8]. In these algorithms, the delay estimation process is reduced to a filter delay that gives minimal error.

Now since in real time problems such as room reverberation, acoustic background noise and the short observation interval exists, thus collectively they can be combined into a case where the signal-to-noise ratio (SNR) is low. And the low SNR affect the performances of these methods. Since the SNR plays an important role in TDE, a SNR threshold is considered as a distinguishable standard between the high and low SNR. So the low SNR aspects are also considered.
Let $x_i(n) = 1,2$ denote the i^{th} microphone signal:

$$x_{i=} \propto_i s(n - \tau_{i)}) + b_i(n)$$

Where $\propto_i$ is an attenuation factor due to propagation effects, $\tau_{i)}$ is the propagation time from the unknown source s(n) to microphone i, and b_i (n) is an additive noise signal at the i^{th} microphone. The relative delay between the two microphone signals 1 and 2 is defined as:

$$\tau_{12} = \tau_1 - \tau_2$$

Unfortunately, in a real acoustic environment we must taken into account the reverberation of the room and the ideal model no longer holds. A more complicate but more complete model for the microphone signals, $x_i(n), i = 1,2;$ can be expressed as follows:

$$x_i(n) = h_i * s(n) + b_i(n)$$

Where* denotes convolution and hi is the acoustic impulse response between the source s(n) and the i^{th} microphone. The reverberation model for single source can be viewed as single-input multiple-output (SIMO) system.

2. Speech Enhancement

The majority of the energy in a speech signal is concentrated in the voiced intervals. In the time-frequency domain, most of the voiced speech energy is located in a small number of harmonic peaks that remain detectable even at poor SNRs. In this section, we propose a method to estimate the speech active level at low SNRs from the energy of the harmonic peaks during voiced intervals. Intelligibility and pleasantness are difficult to measure by any mathematical algorithm. Usually listening tests are employed.

However, since arranging listening tests may be expensive, it has been widely studied how to predict the results of listening tests. The central methods for enhancing speech are the removal of background noise, echo suppression and the process of artificially bringing certain frequencies into the speech signal. First of all, every speech measurement performed in a natural environment contains some amount of echo.

Echoless speech, measured in a special anechoic room, sounds dry and dull to human ear. In most cases the background random noise is added with the desired speech signal and forms an additive mixture which is picked up by microphone. It can be stationary or non stationary, white or colored and having no correlation with desired speech signal.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 4, April 2018

III. DESCRIPTION OF PROPOSED SYSTEM

Speech enhancement or noise reduction has been one of the main investigated problems in the speech community for a long time. The problem arises whenever a desired speech signal is degraded by some disturbing noise. The noise can be additive or convolutive. Even this additive noise can reduce the quality and intelligibility of the speech signal considerably. Therefore, the aim of the noise reduction algorithms is to estimate the clean speech signal from the noisy recordings in order to improve the quality and intelligibility of the enhanced signal. Due to this, it proposes a method for speech enhancement using mask estimation iteratively. The main focus is to improve the cost of system.

Speech enhancement or noise reduction has been one of the main investigated problems in the speech community for a long time. The problem arises whenever a desired speech signal is degraded by some disturbing noise. The noise can be additive or convolutive. Even this additive noise can reduce the quality and intelligibility of the speech signal considerably. Therefore, the aim of the noise reduction algorithms is to estimate the clean speech signal from the noisy recordings in order to improve the quality and intelligibility of the enhanced signal. Due to this, it proposes a method for speech enhancement using mask estimation iteratively.

Echo suppression is needed in big halls to enhance the quality of the speech signal, especially if the distance between the microphone and the speaker is large. When the background noise is suppressed, it is crucial not to harm or garble the speech signal. Another thing to remember is that quiet natural background noise sounds more comfortable than more quiet unnatural twisted noise. If the speech signal is not intended to be listened by humans, but driven for instance to a speech recognizer, then the comfortless is not the issue. It is crucial then to keep the background noise low. Background noise suppression has many applications. Using telephone in a noisy environment like in streets of in a car is an obvious application.

If the background noise is evolving more slowly than the speech, i.e., if the noise is more stationary than the speech, it is easy to estimate the noise during the pauses in speech. Finding the pauses in speech is based on checking how close the estimate of the background noise is to the signal in the current window. Voiced sections can be located by estimating the fundamental frequency. Both methods easily fail on unstressed unvoiced or short phonemes, taking them as background noise. On the other hand, this is not very dangerous because the effect of these faint phonemes on the background noise estimate is not that critical.

A working VAD (voice activity detection) in hand, giving values of zero and one as indicators of the voice activity in each frame, enables us to update the estimate of the background noise spectrum during the frames that have zero V AD, using the formula

$$N(\omega, n) = \lambda\, |N(w, n-1)|^2 + (1-\lambda)\, |X(w, n)|^2$$

In the binary mask approach to speech enhancement, a binary-valued gain mask is applied to the speech in the time-frequency domain and the signal is then transformed back into the time-domain. This procedure is similar to that used in conventional approaches such as spectral subtraction or MMSE estimators except that, in the latter cases, a continuously variable gain function is applied. The principal advantage of the binary mask approach over other state-of-the-art algorithms operating in the time frequency domain is that the problem of enhancement is changed from one of gain estimation to one of classification.

To ensure that classification is independent of the signal input level, the first step of the system is the power normalization of the speech component of the noisy speech signal, $y(\tau)$. The power normalization is performed over the entire duration of the utterance. If the input signal was long enough to include changes in the speech active level, the signal could be divided up into segments to perform this stage.

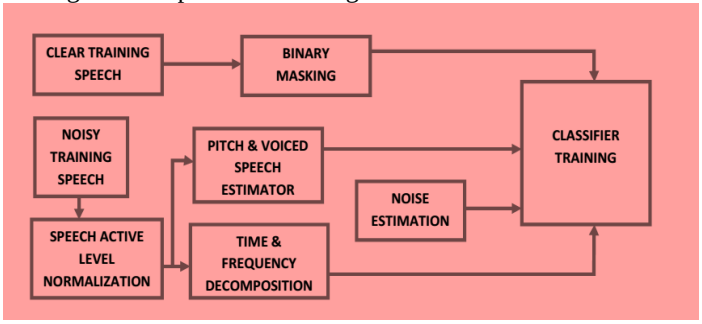

Figure 2: Proposed System Model

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 4, April 2018

Most voiced speech energy is concentrated within the fundamental frequency and its harmonics. Therefore, identifying voiced speech segments and estimating their fundamental frequency makes it possible to locate high speech energy regions. The algorithm provides a fundamental frequency estimate at every time-frame, together with a probability of each time-frame containing voiced speech. Identifying time-frames which contain sibilant phones is important for the preservation of periodic speech energy at high frequencies. Furthermore, an estimation of the power spectrum of the sibilant phone would also help identifying the frequency bands containing most of the sibilant speech energy.

It process the noisy signal in overlapping frames and the energy of the harmonics is spread over a range of frequencies by the effects of the analysis window and the rate of change of f0.To extract the energy of these harmonics, we need to identify the voiced speech intervals and, within these, estimate the value of f0.

The inclusion of the normalized noisy speech periodogram and the noise estimation as parameters aids the mask estimation algorithm by providing information about the energy distribution across frequency of both speech and noise. Classifier is a procedure that constructs a binary decision tree for predicting the output response or class from a set of input parameters taking discrete or continuous values. Each internal node compares one of the input parameters to a threshold and continues to a sub-branch of the tree according to the binary output. This process continues until a terminal node is reached, where prediction is performed by aggregating or averaging all the training data points which reach that node. A visual example of how a binary tree operates is shown. The classifier approach can either be used for classification or regression. Classification trees provide a categorical value at each terminal node while regression trees provide a continuous output. For each internal node of the tree, the training process selects a feature to test and a threshold against which it is compared. These choices are made in order to minimize the average value of a misclassification function. The cross-correlation can be modelled by:

$$R_{r1r2}(\tau) = E[r_1(t)r_2(t-\tau)] \qquad D_{CC} = ar_{\tau}^{g}max[R_{r1r2}(\tau)]$$

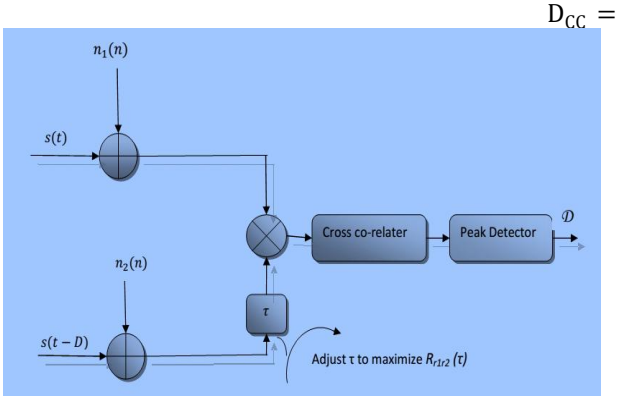

Figure 3: The Cross-Correlation Processor

IV. RESULTS & DISCUSSION

Active noise suppression is a method in which the idea is to produce anti-noise into the listener's ear to cancel the noise. The delay must be kept very small to avoid producing more noise instead of cancelling the existing noise. In this section, we propose a method to estimate the speech active level at low SNRs from the energy of the harmonic peaks during voiced intervals. First of all, every speech measurement performed in a natural environment contains some amount of echo. Echoless speech, measured in a special anechoic room, sounds dry and dull to human ear. In most cases the background random noise is added with the desired speech signal and forms an additive mixture which is picked up by microphone.

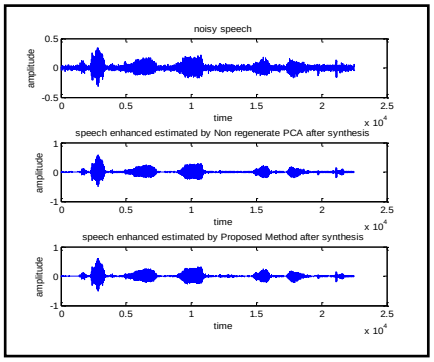

Figure 4: Speech Enhancement by Actual & Proposed Method

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 4, April 2018

Both speech enhancement methods aimed at suppressing the background noise are (naturally) based in one way or the other on the estimation of the background noise. If the background noise is evolving more slowly than the speech, i.e., if the noise is more stationary than the speech, it is easy to estimate the noise during the pauses in speech.

Finding the pauses in speech is based on checking how close the estimate of the background noise is to the signal in the current window. Voiced sections can be located by estimating the fundamental frequency. Both methods easily fail on unstressed unvoiced or short phonemes, taking them as background noise.

In a digital transmission, BER is the percentage of bits with errors divided by the total number of bits that have been transmitted, received or processed over a given time period. The rate is typically expressed as 10 to the negative power.

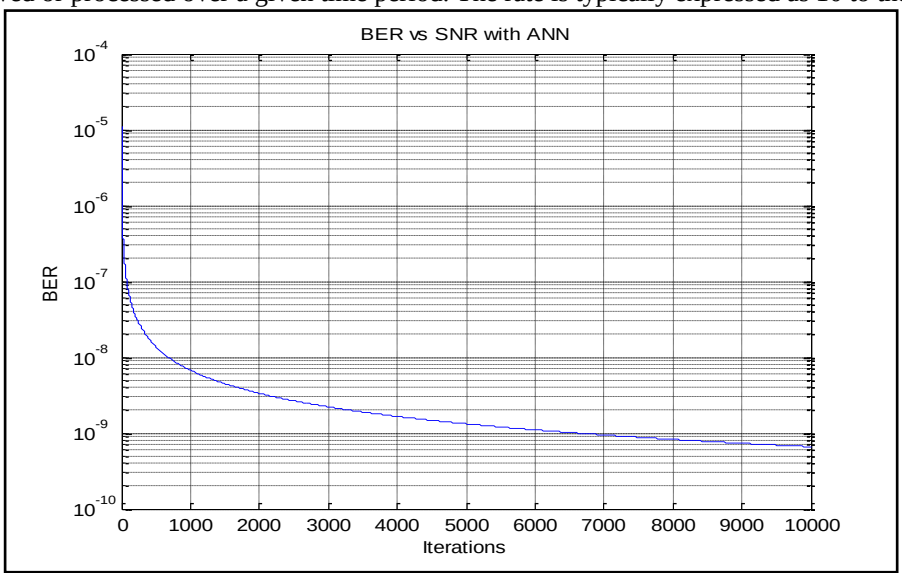

Figure 5: BER Response of Proposed Method

Table 1 shows the MSE response of system after ANN. In this work, it provides speech enhancement under different noisy conditions. After this, it provides the performance comparison with non regenerative method in terms of cost and rank of matrix. The error is minimized by ANN method and proves the system performance better in terms of error reduction.

Table 1: MSE Response after ANN

zIterations	MSE
1	0.4999
1000	0.4998
5000	0.00000719
10000	0.00000457

V. CONCLUSION

Although the most approaches aim to estimate the clean speech by applying a continuous gain. The original goal of binary mask estimation was to identify the regions where the SNR was higher than 0 dB. In addition we have developed an algorithm for estimating the active level of a speech signal even when high levels of noise are present. The noise can be additive or convolutive. Even this additive noise can reduce the quality and intelligibility of the speech signal considerably. Therefore, the aim of the noise reduction algorithms is to estimate the clean speech signal from the noisy recordings in order to improve the quality and intelligibility of the enhanced signal. Due to this, it proposes a method for speech enhancement using mask estimation iteratively. In this work, it provides speech enhancement under different noisy conditions. After this, it provides the performance comparison with non regenerative method in terms of cost and rank of matrix. The error is minimized by ANN method and proves the system performance better in terms of error reduction.

Future work to improve the algorithm could include the application of temporal continuity constraints to the voicing probability estimate.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 4, April 2018

REFFERENCES

[1] Meng Sun, Xiongwei Zhang, Hugo Van hamme, "Unseen Noise Estimation Using Separable Deep Auto Encoder for Speech Enhancement", IEEE/ACM Transactions On Audio, Speech, And Language Processing, Vol. 24, No. 1, January 2016.
[2] Shoko Arakit, Tomoki Hayashi, "Exploring Multi-Channel Features for Denoising-Auto-encoder-Based Speech Enhancement", IEEE 2015.
[3] Zheng Gong and Youshen Xia, "Two Speech Enhancement-Based Hearing Aid Systems and Comparative Study", IEEE International Conference on Information Science and Technology, April 24-26, 2015.
[4] Feng Deng, Changchun Bao, "Sparse Hidden Markov Models for Speech Enhancement in Non-Stationary Noise Environments", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 23, No. 11, November 2015.
[5] Pejman Mowlaee and Josef Kulmer, "Harmonic Phase Estimation in Single-Channel Speech Enhancement Using Phase Decomposition and SNR Information", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 23, No. 9, September 2015.
[6] Swati R. Pawar, Hemant kumar B. Mali, "Implementation of Binary Masking Technique for Hearing Aid Application", IEEE International Conference on Pervasive Computing, 2015.
[7] Xia Yousheng, Huang Jianwen, "Speech Enhancement Based on Combination of Wiener Filter and Subspace Filter", IEEE 2014.
[8] Zhang Jie, Xiaoqun Zhao, Jingyun Xu, "Suitability of Speech Quality Evaluation Measures in Speech Enhancement", IEEE 2014.
[9] Atsunori Ogawa, Keisuke Kinoshita, Takaaki Hori, "Fast Segment Search For Corpus-Based Speech Enhancement Based On Speech Recognition Technology", IEEE International Conference on Acoustic, Speech and Signal Processing, 2014.
[10] AN.SaiPrasanna, Iyer Chandrashekaran, "Real Time Codebook Based Speech Enhancement with GPUs", International Conference on Parallel, Distributed and Grid Computing, 2014.
[11] Zawar Shah, Ather Suleman, Imdad Ullah, "Effect of Transmission Opportunity and Frame Aggregation on VoIP Capacity over IEEE 802.11n WLANs", IEEE 2014.
[12] Lee Ngee Tan, Abeer Alwan, "Feature Enhancement Using Sparse Reference And Estimated Soft-Mask Exemplar-Pairs For Noisy Speech Recognition", IEEE International Conference on Acoustic, Speech and Signal Processing, 2014.
[13] Seung Yun, Young-Jik Lee, and Sang-Hun Kim, "Multilingual Speech-to-Speech Translation System for Mobile Consumer Devices", IEEE Transactions on Consumer Electronics, Vol. 60, No. 3, August 2014.
[14] Christian D. Sigg, Tomas Dikk, "Speech Enhancement Using Generative Dictionary Learning", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 6, August 2012.
[15] H. Veisi H. Sameti, "Hidden-Markov-model-based voice activity detector with high speech detection rate for speech enhancement", IET Signal Processing, 2012.