IJIREEICE

ISSN (Online) 2321-2004
ISSN (Print) 2321-5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2018

Study on Mask Estimation & Artificial Neural Network for Speech Enhancement: A Review

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

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

Abstract: Speech enhancement technology can extract the speech signal in noise environment and enhance the recognition rate. As the current speech enhancement algorithms can give results for improved speech audibility only. So, our main motivation is to develop speech enhancement algorithm that would improve performance in speech. It is difficult to understand the speech signal under presence of noise from background areas. The human speech and hearing system is inherently sensitive to interfering noise. 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 review on method for speech enhancement using mask estimation iteratively. In this work, it will provide the concept of optimization of cost function by iterative method. This will help for reducing the noise from signal. All simulations will be implemented in MATLAB.

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

I. INTRODUCTION

During conversation, both hearing and speaking adapt to the background noise in a noisy environment. It is therefore possible to have a conversation in quite disturbing background noise environments. However, when the conversation takes place over the telephone disturbances are more annoying. The disturbances are a problem since the brain will not get the extra visual and other background information when interpreting the speech. The speech signal transmitted to the other party is picked up by a microphone connected to the telephone. The microphone signal contains both speech and noise at some ratio (Speech to Noise Ratio, SNR) depending on, for example, how far the microphone is mounted from the speaker's mouth.

In vehicles, it is common to use a telephone hands free accessory. The main motivation for this is to facilitate both hands on the steering wheel when driving. The drawback is that the telephone microphone is situated at a farther distance (50 cm) from the speaker's mouth as compared to handheld telephony, in a high noise level environment. To increase the SNR and to allow for the listener to grasp the speech clearly, a speech enhancement method should be applied. The use of systems involving speech-based communication technology is now ubiquitous; such systems include mobile phones, hearing aids and video-conferencing technology. The perceived quality, and in more severe cases the intelligibility, of the speech signal in these systems is reduced when they are used under the adverse noise conditions encountered in real environments such as offices, crowded public spaces, or railway stations [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.

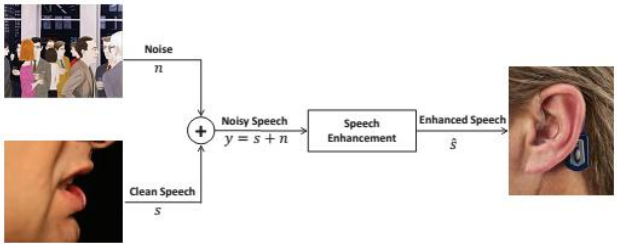

Figure 1: Speech Enhancement System with Corrupted Noise

IJIREEICE

ISSN (Online) 2321-2004
ISSN (Print) 2321-5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2018

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.

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

II. RELATED WORK

Meng Sun et. al. [1] presented a deep auto encoder (DAE) for accurately modelling the clean speech spectrum. In the subsequent stage of speech enhancement, an extra DAE was introduced to represent the residual part obtained by subtracting the estimated clean speech spectrum (by using the pre-trained DAE) from the noisy speech spectrum. By adjusting the estimated clean speech spectrum and the unknown parameters of the noise DAE, one can reach a stationary point to minimize the total reconstruction error of the noisy speech spectrum. The enhanced speech signal was thus obtained by transforming the estimated clean speech spectrum back into time domain. The above proposed technique was called separable deep auto encoder (SDAE). Given the under-determined nature of the above optimization problem, the clean speech reconstruction was confined in the convex hull spanned by a pre-trained speech dictionary

Shoko Araki et. al. [2] 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. Although multichannel speech enhancement usually outperforms single channel approaches, there has been little research on the use of multi-channel processing in the context of DAE. In this paper, they explored the use of several multi-channel features as DAE input to confirm whether multi-channel information can improve performance.

Zheng Gong et. al. [3] developed two embedded hearing aid systems with noise reduction, respectively using Kalman filter and Wiener filter techniques. Next, they gave a comparative study on the two speech enhancement-based hearing aid systems by testing subjective auditory in noise environment. The comparative result showed that the hearing aid system based on the Kalman filter-based speech enhancement can increase the rate of speech recognition and the hearing comfort of hearing impaired persons in a noisy environment, compared with the hearing aid system based on the Wiener filter-based speech enhancement.

Feng Deng et. al. [4] 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. Autoregressive models were employed to describe the speech and noise in a unified framework and the speech and noise gains are modelled as random processes with memory. The likelihood criterion for finding the model parameters is augmented with a regularization term resulting in a sparse autoregressive HMM (SARHMM) system that encourages sparsity in the speech- and noise- modelling. In the SARHMM only a small number of HMM states contribute significantly to the model of each particular observed speech segment.

Pejman Mowlaee et. al. [5] 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. To incorporate the uncertainty in the estimated phase due to unreliable voicing decision and SNR estimate, they proposed a binary hypothesis test assuming speech-present and speech-absent classes representing high and low SNRs. The effectiveness of the proposed phase estimation method is evaluated for both phase-only enhancements of noisy speech and in combination with an amplitude-only enhancement scheme.

Swati Pawar et. al. [6] presented an algorithm for improving speech intelligibility. Various speech enhancement algorithms were developed but only some of them can be used for real time hearing aid applications. This proposed algorithm can be used for practical hearing prosthetic devices. Implementation of the binary masking algorithm uses a bank of band-pass filters to perform mapping of signals. Also, classification is performed with a signal-to-noise (SNR) estimate and a comparator. This includes spatial filtering method, classification of signals such as original and noisy signal. After this based on SNR threshold level signals are recombined to obtain reduced noise level in speech signal.

Xia Yousheng et. al. [7] proposed a novel multi-channel speech enhancement method by combining the wiener filtering and subspace filtering with a convex combinational coefficient. Because of using both the advantage in noise reduction of the subspace speech enhancement technology and the stable characteristic of the Wiener filtering technology, the proposed multi-channel speech enhancement method had a better performance in robustly removing colored noise from noisy speech signals. Simulation examples confirmed that under different colored noise, the proposed multi-channel

IJIREEICE

ISSN (Online) 2321-2004
ISSN (Print) 2321-5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2018

speech enhancement method can obtain better speech recovery results than the traditional subspace multi-channel speech enhancement method.

Zhang Jie et. al. [8] discussed the suitability of speech quality evaluation measures under various noise environments in the application of spectral subtraction speech enhancement. It took three kinds of typical noise and evaluated comprehensively the speech quality under the standard of global signal-to-noise ratio of noisy speech. It took six kinds of quality measures which include mean opinion score, perceptual evaluation of speech quality, segmental signal-to-noise ratio, weighted spectral slope, log-likelihood ratio and log spectral distance. Then appropriate evaluation algorithms were chosen for speech enhancement based on spectral subtraction. The simulation results showed that in the application of speech enhancement, the suitability of speech quality evaluation algorithms is limited to the SNR of noisy speech, recording people, recording content and background noise environment.

Atsunori Ogawa et. al. [9] proposed a fast segment search method for corpus based speech enhancement. It was mainly based on two techniques derived from speech recognition technology. The first was a search like segment evaluation function for accurately finding the longest matching segments. The second was a tree and linear connected search space for efficiently sharing the segment likelihood calculations. In the experiments for non-stationary noisy observations using the 26 multi-condition TIMIT parallel speech corpus, the proposed search method found the segments almost in real-time without degrading the quality of the enhanced speech.

A. Prasanna et. al. [10] presented a Codebook-based speech enhancement (CBSE) employing trained speech and noise codebooks for handling non-stationary noise. However, the high compute intensive nature of this technique rendered it inapplicable in real-time speech enhancement scenarios by introducing a significant delay in speech transmission. In this work, this problem was addressed by providing an efficient, parallel CBSE algorithm. The proposed parallel CBSE algorithm achieved significant speedup and reduced execution time, resulting in a speech transmission delay which is well within the limits of realizing real-time speech enhancement. The proposed parallel CBSE algorithm was then used as a basis to provide a novel cloud based framework to achieve real-time speech enhancement in mobile communication as a proof-of concept.

III. 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 modeled 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'.

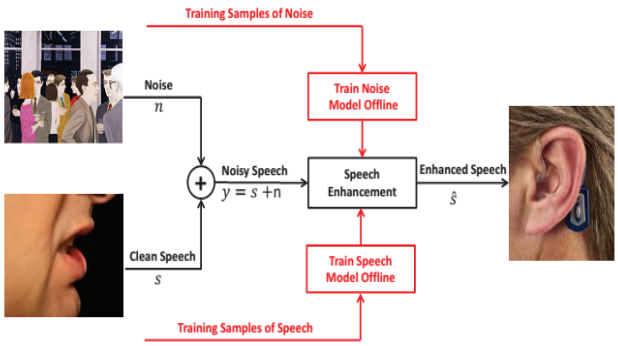

Figure 2: Supervised Speech Enhancement System

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.

IJIREEICE

ISSN (Online) 2321-2004
ISSN (Print) 2321-5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2018

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 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 ith microphone signal:

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

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 ith 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 ith 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.

IJIREEICE

ISSN (Online) 2321-2004
ISSN (Print) 2321-5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2018

IV. PROBLEM DEFINITION

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.

V. CONCLUSION

Speech signals can be degraded in many ways during their acquisition in noisy environments and they can also be further degraded in the electronic domain. This paper provides a review on speech enhancement method based on sparse modelling in noisy environment. This method will provide a noise reduction procedure which functions and gives low residual noise, high quality speech. The main parameters are BER, noise sigma and cost of function. For this, it will use the concept of sparse modelling and PCA vectors. The low variability of the gain function during stationary input signals will give an output with less tonal residual background noise, thus low noise distortion. After this, it will use the concept of ANN for minimizing the error.

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.