Sindh Univ. Res. Jour. (Sci. Ser.) Vol.49 (004) 895-898 (2017)

http://doi.org/10.26692/sujo/2017.12.0078

SINDH UNIVERSITYRESEARCH JOURNAL (SCIENCESERIES)

Robust Speech Recognition Using Adaptive Noise Cancellation

M. WAQAS, M. A. A. KHAN, M. NAEEM[*], A.GUL[**], N. AHMAD[+]

Department of Computer Systems Engineering, University of Engineering and Technology, Peshawar

Received 30th July 2016 and Revised 24st March 2017

Abstract- This paper introduces the adaptive noise cancellation technique for the noise reduction in Robust Automatic Speech Recognition. The adaptive noise cancellation is used as front-end stage to enhance the extracted features for speech recognition under noisy conditions. More specifically, the Constrained Stability Least Mean Square (CS-LMS) algorithm which is a member of the family of adaptive filters has been applied. The Hidden Markov Model based Tool Kit (HTK) is used for training and testing the Automatic Speech Recognizer system. The result obtained shows that the application of adoptive filtering at the front-end enhances the performance of the system in noisy conditions while the CS-LMS algorithm gives the most superior performance among the family of LMS algorithms.

Keywords: Automatic Speech Recognition, Robust Speech Recognition, Adaptive Filtering

1. INTRODUCTION

Automatic Speech Recognition (ASR) systems have a wide range of practical applications, however its deployment in many real life application is hindered by a number of issues, such as inter speaker and intra speaker invariability, co articulation and cross talk, and environmental noise (Crowley and Bowern. 2010). By far, the most challenging problem of ASR is noise. Although the ASR systems today are able to achieve a reasonable level of accuracy in controlled environments but its performance degrades drastically with the introduction of noise (Rajnoha and Pollák. 2011). Due to background noise, the features extracted from the test data get significantly different from the similar features extracted from the training data thus rendering inaccurate results (De Wet *et. al.* 2001, Zhang *et. al.* 2006).

Most of the recent research on the ASR is focused mainly to address the issue of speech recognition under noisy conditions. Various techniques have been developed for speech recognition under real-life conditions (Gales and Young. 1993, Hermansky and Sharma. 1999, Choi. 2004). The missing feature detection methods detect the corrupted spectral features and try to either correct these features or otherwise ignore them (Raj and Stern. 2005). The robust feature extraction methods is based on the extraction of features which are inherently robust to environmental noise such as Linear Predictive (LP) Spectral Estimates, Minimum Variance Distortionless Response (MVDR) modeling and RASTA techniques (Kallasjoki *et. al.* 2009). Some models known as compensation model are based on the idea that along with an HMM of speech, an HMM for the noise can be created with an assumption that linear power versions of the speech HMM and noise HMM be orthogonal and thus linearly independent and additive. The combined linear power version of both HMMs is then used in testing (Smith. 2009). A vastly used compensation model is Parallel Combinational Model (PCM) (Gales and Young. 1993a). Some robust speech recognition techniques adhere to the notion that clean speech is corrupted by unwanted signal or noise i.e. humming of engine, another speaker, traffic noise, fan humming etc, and this unwanted noise can be removed from the corrupted speech signal by estimating the noise spectra (Patynen. 2009). More complex techniques use the statistical estimation of the noise for its removal form corrupted speech (Compernolle. 1992). These methods operate on the assumptions of different noise types and therefore the methods like compensation model and feature enhancement methods are sensitive to the noise type and give highly accurate results to selective background noise where as robust features extraction and missing data methods gives fair accurate results in all noise types (Smith. 2009). Single microphone techniques estimates and corrects speech signal with fairly good accuracy, however multiple microphone methods do much accurate noise estimation as compared to single microphone techniques (Patynen. 2009). This method includes adaptive noise cancellation, beam forming, and blind source separation. Adaptive noise cancellation takes into account a noise reference signal and using adaptive filter to estimate noise in the corrupted signal and removes it from the corrupted signal to get the clean signal or speech. Beam forming uses multiple degraded

[+]Corresponding Author email: N. Ahmad, n.ahmad@uetpeshawar.edu.pk

[*]Department of Computer Science, University of Peshawar

[**] Department of Statistics, Shaheed Benazir Bhutto Women University Peshawar

signals instead of noise reference (Compernolle. 1992). This paper addresses the challenge of environmental noise by applying an adoptive noise cancellation technique.

The rest of the paper is organized as follows. The next section explains the working of adoptive noise cancellation approach and the variant algorithms of ANC. Section 3 describes the experimental setup used and discusses the results obtained by using different ANC algorithms. Section 4 concludes the findings of this research and outlines the future line of research.

2. ADAPTIVE NOISE CANCELLATION

The Adaptive Noise Cancellation (ANC) uses an adaptive filter for estimating noise using a referenced noise signal which is statistically similar to the additive noise contained in noisy speech (Górriz *et. al.* 2009). Two microphones are required to capture reference noise and noise corrupted speech signal. Adaptive filter updates its filter coefficients at every incoming signal sample using a feedback mechanism. This feedback mechanism uses a weight update equation for computing new filter coefficients. A working of a typical ANC mechanism is depicted in Figure 1**.**

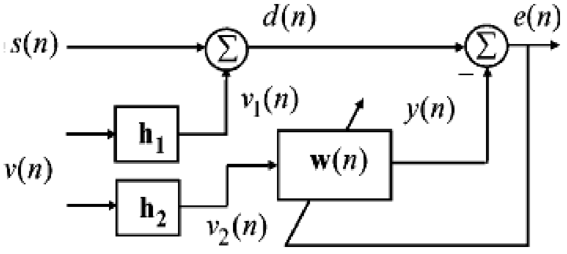

Figure 1: ANC (Adaptive Noise Canceller)

Here $\mathrm{s(n)}$ is the clear or primary speech and $\mathrm{d(n)}$ is noisy speech signal corrupted by $\mathrm{v_1(n)}$; a statistically similar noise to $\mathrm{v_2(n)}$ where $\mathrm{v_2(n)}$ is reference noise input to the adaptive filter. $\mathrm{e(n)}$ is error signal and feedback for adaptive filter. A typical FIR adaptive filter, as shown in Figure 1 can be expressed as,

$$y(n) = \sum_{l=0}^{p} \omega_n(l) v_2(n-l) \qquad (1)$$

In terms vector notation,

$$y(n) = \boldsymbol{\omega}_n^T \boldsymbol{v_2(n)}$$

Where $\omega_n(\mathrm{l})$ is the value of l^th adaptive filter coefficient at time n,

$$\boldsymbol{\omega}_n = [\omega_n(0), \omega_n(1) \ldots . \omega_n(l)]^{\boldsymbol{T}}$$

$$\boldsymbol{v_2}(n) = [v_2(n), v_2(n-1) \ldots . v_2(n-l)]^{\boldsymbol{T}}$$

The design of adaptive filter is much more difficult because of its shift-variant nature. Due to its adaptive nature, its coefficients i.e. ω_n are not fixed and keep on changing. At every iteration the coefficients set is updated with a new set of optimum filter coefficients. The filter coefficients are selected such that it minimizes the mean square error, $\xi(n) = E\{|e(n)|^2\}$.

Where

$$e(n) = d(n) - y(n) = d(n) - \boldsymbol{\omega}_n^T \boldsymbol{v_2(n)} \qquad (2)$$

Replacing $d(n) = s(n) + v_1(n)$ in eq. (2) we get

$$e(n) = s(n) + v_1(n) - \boldsymbol{\omega}_n^T \boldsymbol{v_2(n)} \qquad (3)$$

As can be seen in Figure 1, the input to adaptive filter is $v_2(n)$, a reference noise which is correlated with $v_1(n)$. The FIR adaptive filter estimates $v_1(n)$ and is then subtracted from $d(n)$ to get the clear speech (n) . So,

$$y(n) = \widehat{v_1} = \boldsymbol{\omega}_n^T \boldsymbol{v_2(n)} \qquad (4)$$

$$e(n) = s(n) + v_1(n) - \widehat{v_1} \qquad (5)$$

$$e(n) = \widehat{s(n)} \ldots .. (6) \qquad (6)$$

There are a number of adaptive algorithms with the very basic one being the LMS adaptive algorithm/filter. The weight update equation of LMS adaptive filter is:

$$\omega(\mathrm{n}+1) = \omega(\mathrm{n}) + \mu\mathrm{e(n)v_2}^*(\mathrm{n})$$

For good tracking ability and convergence to mean, the step size μ should be $0 < \mu < \frac{2}{\lambda_{\max}}$. Where $\lambda_{\max}$ is the maximum eigenvalue of auto correlation matrix,$\mathrm{R_{v_2}}$, of $\mathrm{v_2}$. Adaptive filters are compared and characterized in terms of their miss-adjustments and EMSE (Excess Mean Square Error). Decreasing the miss-adjustments and EMSE minimizes mean square error, and thus improves the performance of adaptive filter.

The problem with using LMS algorithm is that, $\lambda_{\max}$ and $\mathrm{R_{v_2}}$are not known and are therefore estimated. To cope up with this problem another variant of LMS Algorithm known as Normalized LMS (NLMS) is used. The weight update equation of NLMS is given by:

$$\boldsymbol{\omega(n+1)} = \boldsymbol{\omega(n)} + \boldsymbol{\beta} \frac{\boldsymbol{v_2}^*\boldsymbol{(n)}}{\boldsymbol{\epsilon} + ||\boldsymbol{v_2(n)}||} \boldsymbol{e(n)}$$

Where $0 < \beta < 2$ is the normalized step size and ϵ is a small positive number.

A relatively new and improved LMS algorithm is proposed in (Górriz *et. al.* 2009) known as Constrained Stability Least Mean Square (CS-LMS) algorithm/filter is given by:

$$\omega(n+1) = \omega(n) + \mu \frac{\delta v_2{}^*(n)}{\epsilon + \left|\left|\delta v_2(n)\right|\right|} \delta e(n)$$

where $\delta e(n) = e(n) - e(n-1)$, and

$$\delta v_2(n) = v_2(n) - v_2(n-1)$$

As explained in (Górriz *et. al.* 2009), CS-LMS and NLMS adaptive filters converge to optimal wiener solution, and for similar step size CSLMS shows good filtering results than NLMS by improving EMSE and miss-adjustments.

3. EXPERIMENTS AND RESULTS

The experimental carried out in research consist of recognizing speech in the noisy environment while using adaptive filters for noise filtering. The setup used is shown in Figure 2.

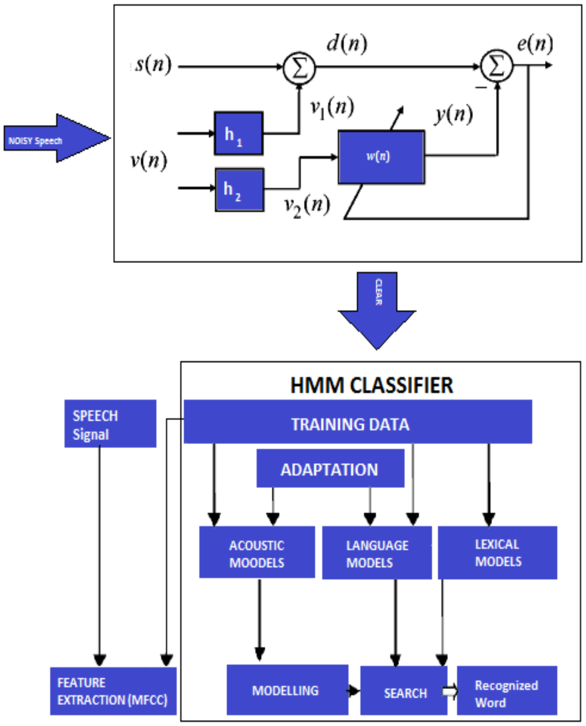

Figure 2: Proposed Robust ASR System

The input to the system is a noisy speech and output is recognized words. The experimental setup is divided into two modules, i.e. the adaptive noise cancellation module which estimates clean speech from the noisy speech, and the ASR module which converts the cleaned speech to recognized text.

VidTIMIT (Sanderson and Paliwal. 2002) dataset is used for training and testing. The dataset consist of the pre-recorded 10 sentences of English language for each speaker. Out of these 10 sentences 2 sentences are same for each while the remaining 8 sentences are different for each speaker. There are 43 speakers in the dataset resulting in a total of 430 sentences.

Additive White Gaussian Noise available in MATLAB is used to make the noisy speech in the range from +60 DB to -60 DB.

HTK, a Hidden Markov Model (HMM) based ASR toolkit is used for feature extraction, and training and testing of the classifier. A five state left right HMM model with 3 emitting state is trained using MEL-Frequency Cepstral Coefficient (MFCC) features from the speech signals after noise removal.

Three different weight update equations i.e. LMS, NLMS and CS-LMS are used for noise removal. Using $v(n)$, a primary noise source, two different noise patterns namely $v_1(n)$ and $v_2(n)$ are created which are statically similar to $v(n)$ but have different parameters. Where $v_1(n)$ is added to the clear speech and $v_2(n)$ is passed to adaptive filter for estimation of noise contained in noisy speech. For practical reasons the impulse response of h_1 and h_2 are:

$$H_1^{-1}(z) = 1 - 0.3z^{-1} - 0.1z^{-2}$$

$$H_2^{-1}(z) = 1 - 0.2z^{-1}$$

The FIR adaptive filter is modeled using 3 different LMS algorithm with 5 filter taps using step size of 0.01. The percentage of correctly recognized words using the three LMS algorithms at different noise level is shown in Figure 3.

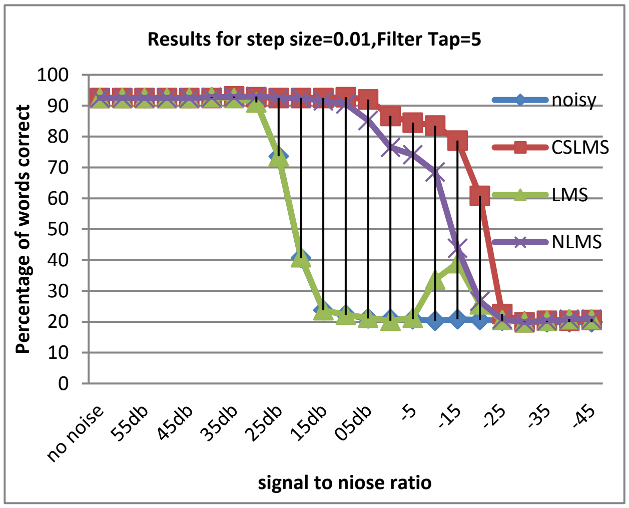

Figure 3: Recognition results

The results in Fig. 3 shows that for the noisy speech the percentage of correctly recognized words drops rapidly at a fairly high SNR level i.e. bellow 25dB. At negative SNR levels, the words correct percentage remains around 21%. Using LMS algorithm, the performance is almost the same as that of the noisy speech. This is probably because the determination of step size $\boldsymbol{\mu}$ is dependent on the max eigen value of the auto correlation matrix of the reference input to the adaptive filter i.e. $\boldsymbol{v_2(n)}$, and also due to the weak tracking ability of LMS algorithm. The NLMS algorithm shows fairly good results even at -10 DB noise level whereas the CSLMS shows the best results with a much higher accuracy of 60% at a SNR of -20 dB. This is because; the CSLMS has better tracking ability due to minimum EMSE and miss-adjustments as compared to LMS and NLMS.

4. CONCLUSION

This paper presented Robust Speech recognition using adaptive noise cancellation. ANC using LMS, NLMS and CSLMS algorithms were investigated and the word recognition performance compared for a range of SNR levels. This work showed that CSLMS works better than the other algorithms due to its efficient weight update mechanism. For recognition HMM based classifier was used with MFCC features. The noise type used in this work is Additive White Gaussian Noise which is the more generic noise type; however it can be tested on specific noise types such as car noise, canteen noise, fan noise etc. Various other types of classifier like LDA, PCA, Hybrid HMM and ANN etc can be used with other features like PLP, WT and LPC.

REFERENCES:

Crowley, T., and C. Bowern, (2010). An introduction to historical linguistics. Oxford University Press.

Choi, E., (2004). Noise robust front-end for ASR using spectral subtraction, spectral flooring and cumulative distribution mapping. 10th Australian International Conference on Speech Science and Technology: 451-456.

Compernolle, D. V. (1992). "DSP techniques for speech enhancement", ESCA Workshop on Speech Processing in Adverse Conditions, 21–30.

De Wet, F., B. Cranen, J. de Veth, and L. Boves, (2001). A comparison of LPC and FFT-based acoustic features for noise robust ASR. INTERSPEECH: 865-868.

Gales, M. J. F., and S. J. Young, (1993). Parallel model combination for speech recognition in noise. University of Cambridge, Department of Engineering.

Gales M. J. F. and S. J. Young, (1993a). "Cepstral parameter compensation for HMM recognition in noise", Speech Communication, 12, 231–239.

Górriz, J. M., J. Ramírez, S. Cruces-Alvarez, C. G. Puntonet, E. W. Lang, and D. Erdogmus, (2009). A novel LMS algorithm applied to adaptive noise cancellation. IEEE Signal Processing Letters, 16(1), 34-37.

Hermansky, H. and S. Sharma, (1999). Temporal patterns (TRAPS) in ASR of noisy speech. 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. 1: 289-292.

Kallasjoki, H., K., Magi, C., Alku, P. and Kurimo, M., 2009. Noise robust LVCSR feature extraction based on stabilized weighted linear prediction. 13th International Conference on Speech and Computer (SPECOM). 221-225.

Patynen, J. (2009). Feature Enhancement in Automatic Speech Recognition, in *Studies on Noise Robust Automatic Speech Recognition*, K. J. Palomaki, U. Remes and M. Kurimo (Eds.): 35-44.

Raj, B. and R. M. Stern, (2005). Missing-feature approaches in speech recognition. IEEE Signal Processing Magazine, 22(5): 101-116.

Rajnoha, J., and P. Pollák, (2011). ASR systems in noisy environment: Analysis and solutions for increasing noise robustness. Radio engineering, 20(1): 74-83.

Smit P., (2009). Model Compensation for Noise Robust Speech Recognition. in *Studies on Noise Robust Automatic Speech Recognition*, K. J. Palomaki, U. Remes and M. Kurimo (Eds.): 45-52.

Sanderson, C., and K. K. Paliwal, (2002). Polynomial features for robust face authentication.2002 International Conference on Image Processing. 3: 997-1000.

Zhang, C., J. van de Weijer, and J. Cui, (2006). Intra- and inter-speaker variations of formant pattern for lateral syllables in Standard Chinese. Forensic science international, 158(2): 117-124.