Noise Robust Automatic Speech Recognizer for Voice Controlled Micro Air Vehicles

Chaitra K N[*], Veena S[**], Rahul D K[***], Sandhya Lakshmi R[****] and Anjan Kumar B S[*****]
[*]PG Student, Bangalore Institute of Technology, Bengaluru
chaitra.chethu321@gmail.com
[**]Principal Scientist, CSIR- National Aerospace Laboratories, Bengaluru
veenas@nal.res.in
[***]Project Assistant, CSIR- National Aerospace Laboratories, Bengaluru
[****]PG Student, RV College of Engineering, Bengaluru
[*****]Assistant Prof, Bangalore Institute of Technology, Bengaluru

Abstract: Voice controlled MAVs are an attractive feature to popularize MAV operation among the prospective users. However, its success depends on achieving good recognition rates in the presence of noise. This paper proposes Joint Lattice predictor for prediction of voice command in the presence of noise at the MAV operational environment. The developed algorithm is integrated with the Mission Planner, an open source Ground Control Station (GCS) platform for controlling of MAVs.

Keywords: MATLAB, Mission Planner, speech controlled MAV, speech enhancement, speech predictor, joint lattice structure.

Introduction

The **Micro Air Vehicles** (MAV) find applications in both civilian and military arena. Typically MAVs are remote controlled through a ground control station (GCS). Traditional GCS has a menu based graphical front end with keyboard and/or mouse as user interface. The graphical content on the page requires a certain level of understanding and requires some training. But to popularize MAV applications, untrained personnel should also be able to fly and control the MAV operations with ease. Also, in certain applications such as flood situations, the user may find difficulty in finding a suitable place for comfortably operating the GCS. This demands GCS to be user friendly, with minimal usage of GCS hardware. In fact, it would be desirable to have the GCS in a backpack, so that the user can be hands free. These requirements could be met if speech can be used as the mode for GCS command activation. Here, the user utters the desired action to be performed by MAV using speech and the GCS takes care of understanding the speech and generating the corresponding MAV command. The success of this approach depends on how well the GCS can interpret the uttered speech.

First step in achieving this is the development of Automatic Speech Recognition (ASR) techniques to recognize the uttered speech. Even though, ASR is a well established area, its usage to MAV applications in relatively recent. Works on ASR from MAV-GCS perspective are sighted in [1] [2] [3]. These works mainly focus on extraction of speech features and developing efficient speech models to achieve accurate recognition. The laboratory performance of such systems is very good and highly encouraging to deploy them for MAV applications. However, since MAV is a typical field application, the ASR has to operate in the presence of ambient noise like traffic, wind etc. It is a well known fact that speech recognition gets affected by the ambient noise [4]. Therefore for the success of this application, recognition of speech even in the presence of noise is very much essential.

There are several ways to increase the robustness of speech recognition in noisy environments [5]. The first is to use of sophisticated voice activity detection algorithms like Cepstral-based or GMM-based detectors. The second way to increase robustness is to use special noise-adapted acoustic models, as well as robust features for the parameterization of noisy speech. Another alternative is to use some noise-reduction algorithms, which is simple and the most popular. There are several approaches like spectral estimation, filtering and model based methods. However, a simple, efficient and computationally light scheme is the requirement of real-time applications.

In this paper, a gradient lattice adaptive predictor is proposed for improving the SNR of speech corrupted with ambient noise. The proposed algorithm is developed in C# and integrated with the Mission Planner GCS and is tested in real-time with various noise types simulated in the laboratory.

This paper is organized as follows: Section – II describes the operation of voice controlled MAV, Section – III gives details of the joint lattice predictor proposed for speech enhancement. Section IV gives the performance of ASR with the proposed technique.

Voice Controlled MAV

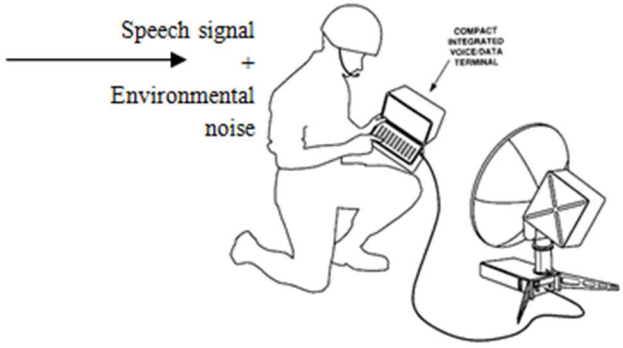

Figure1. User operation of MAV-GCS

Fig. 1 shows the user operation of a GCS with an option for voice based control. A typical speech based command would be like, "**TAKE OFF** FROM HOME, **TURN LEFT** AND MOVE **100 METERS**". This sentence has to be recognized and keywords marked in highlighted text have to be picked up. Therefore the ASR should accomplish Key Word Spotting (KWS), which is the capability to identify specific words or phrases from a list of items in the language being spoken. Once the key words are recognized, they have to be converted into relevant action commands for the MAV. This requires development and integration of speech interface into the GCS.

For efficient operation, the accuracy of the system must be good irrespective of the operational environment. This calls for a robust ASR, implying that it is capable of handling any mismatch in training and testing, arising due to environmental condition [6] and can maintain good recognition performance even if this mismatch exists. Since noise is the main concern, if the noisy speech is cleaned before feeding it to the ASR system, then no changes in the recognition system are necessary to make it robust. This paper follows such an approach.

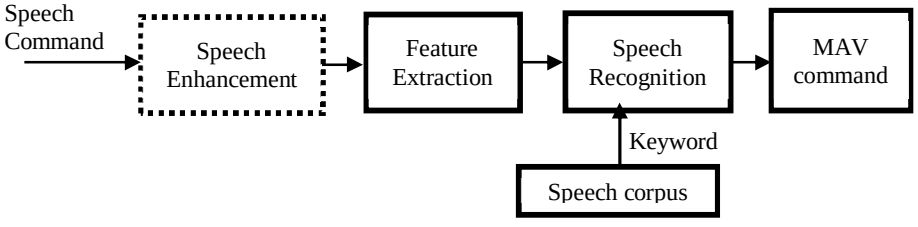

Figure 2. Block diagram of ASR system

Fig. 2 shows the basic block diagram description of the tasks that need to be carried out. Each block of ASR system is explained here. In an ASR system the speech uttered is fed to a Feature extractor. This extracts compact coefficients from the speech signal by preserving temporal and/or spectral characteristics of speech. [7] describes several methods and among the available methods, Cepstral analysis (MFCC) has been used extensively for feature extraction in speech recognition. The feature vector obtained from this block measures the similarity between an input speech and a reference pattern/a model (obtained during training) and determines a reference/ a model, which best matches the input speech. Hidden Markov Model (HMM), a statistical model is the most popular model used for speech recognition.HMM is very powerful mathematical tool for modelling time series and is based Markov Chain. A natural extension of Markov chain is Hidden Markov Model (HMM), the extension where the internal states are hidden and any state produces observable symbols or evidences. For the uttered word, HMM selects the most likely word from speech corpus with the largest probability.

Here speech corpus is the database created for training and testing HMM. It contains speech audio files and corresponding text transcriptions and is known as speech corpus. The MAV commands fall into the category of isolated words or connected words and simple sentences.

Usually the training environment is a noise free laboratory, but since the testing/operational environment is usually outdoors, the command uttered is corrupted by environmental noise. If this signal is fed to feature extractor, the feature extracted is not accurate and this affects the recognition. [8] discusses the effect of noise on the MFCC coefficients. The proposed paper uses a speech enhancement system as a pre processor to suppress the noise in a noisy speech signal and fed to the speech recognizer. As this is the core component of this paper, more details regarding this is given in Section – III.

Proposed Speech Enhancement Method

Noise reduction or speech enhancement means the improvement of the quality or intelligibility of speech signals by reducing or removing background interference, noise or speech distortion, and hence, the improvement of the SNR of the contaminated speech. Many types of additive noise as well as most channel distortions vary slowly compared to the variations in speech signals. Filters that remove variations in the signal that are uncharacteristic of speech (including components with both slow and fast modulation frequencies) improve the recognition accuracy significantly. Literature shows the usage of spectral subtraction (SS), Wiener filter, Kalman filter and adaptive filter techniques for speech enhancement [5].

This paper uses a linear predictor based on joint Lattice structure [9] to obtain clean speech for ASR. Even though lattice predictor has been extensively for speech feature extraction [10], using the joint Lattice Structure for the obtaining noise free speech signal is being proposed for the first time in this paper.

Linear prediction is a classic signal processing technique that provides estimates of the input based on the measurements made at the past. This technique is generally useful in separating signals from noise based on the differences in the correlation between signal and noise.

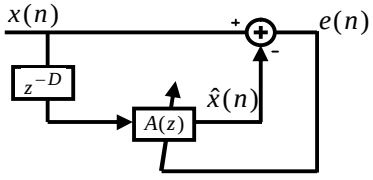

Figure3. Linear Predictor

Fig. 3 gives the block diagram of linear predictor. z^{-D} refers to De-correlation delay. Here, the value of D is chosen to differentiate between speech and noise. The adaptive filter $A(z)$ is adjusted to match delayed version of the input to match the present value by performing Mean Square Error (MSE) minimization. The signal estimate $\hat{x}(n)$ is gives the clean speech signal as the predictor output is the signal with correlation greater than D and noise is having a correlation lesser than D.

$$\hat{x}(n) = \sum_{i=1}^{L-1} a_i x(n-i-D)$$

and the error is given as

$$e(n) = x(n) - \hat{x}(n)$$

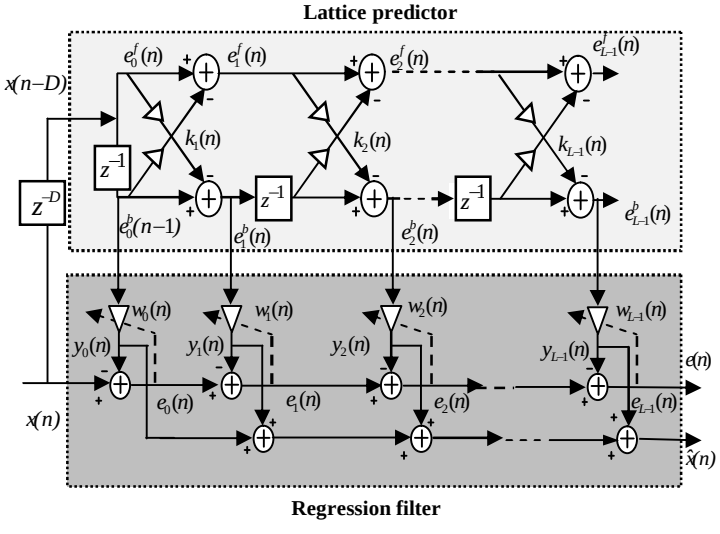

Figure 4. Joint Lattice Filter structure based predictor for speech enhancement

Least Mean Square (LMS) algorithm is one of the popular algorithms for MSE minimization owing to its simplicity [11]. However, the Convergence of LMS algorithm is slow for narrowband signals due to large eigenvalue spread in the input

autocorrelation matrix. Since speech is a narrowband signal, performance of LMS predictor is poor. To address this, an adaptive Joint Lattice structure is available in literature. Lattice Predictor is one of the most effective structures for generating simultaneously the forward and backward prediction errors [9]. The uncorrelated backward errors are weighed to achieve the predictor output $\hat{x}(n)$.

Fig. 4 shows the structure of predictor based on Joint Lattice Structure. It consists of two parts, first part is lattice predictor, which converts input samples to mutually uncorrelated backward prediction errors and the second part is the linear combiner, which produces the output of the filter by the linear combination of the backward prediction errors.

The adaption equations are as follows

$$k_l(n+1) = k_l(n) + \alpha_l(n)\left[e_l^f(n)e_{l-1}^b(n-1) + e_l^b(n)e_{l-1}^f(n)\right]$$

where $\alpha_l(n) = \dfrac{\alpha}{E_{l-1}(n)}, 0 < \alpha < 1$

$$E_l(n) = \gamma\ E_l(n-1) + (1-\gamma)\left[f_l^2(n) + b_l^2(n-1)\right], 0 < \gamma << 1$$

$\alpha_l(n)$ and $E_l(n)$ are the step size and the power estimate of the input, for the l^{th} stage and at n^{th} time instant, respectively. The prediction estimate can be achieved by summing the scaled backward prediction errors by adaptive filter coefficients,

$$\hat{x}(n) = \sum_{l=0}^{L-1} w_l(n)e_l^b(n)$$

where, $w_l(n+1) = w_l(n) + \dfrac{\mu_u}{\hat{G}_l(n)} b_l(n)e_l(n)$

Then, $e(n) = x(n) - \hat{x}(n)$

$\hat{x}(n)$ is the estimate of the speech signal. This noise free signal is fed to the feature extraction block of Fig.2.

Simulation Results

Initially the algorithms were implemented and the simulations were carried out on MATLAB platform to validate the effectiveness of the adaptive noise cancellation algorithms for different noise signals based on SNR improvement and Recognition. Noise signals from various MAV operating scenarios were downloaded from internet.

The speech recognition system was trained and tested for isolated-words at a sampling frequency of 16kHz. The performance of the proposed algorithm was compared with that of Spectral Subtraction (SS) [12], Kalman filter [13] and LMS predictor. The results obtained from the different algorithms are summarized in the above Table - 1, it shows that lattice structure based adaptive noise cancellation has very good recognition compared to other algorithms. The input SNR value is the value at which the recognition failed. For each of the algorithms, SNR improvement achieved is clearly indicated. As seen from the Table -1, even though the SNR improvement is comparable to the LMS algorithm, speech recognition couldn't be achieved as the quality of estimated speech was not comparable to clean speech. Joint Lattice Predictor was capable of achieving good SNR improvement along with clear speech.

Table 1. Performance of speech enhancement algorithms

Type of Noise	Input SNR in dB	Improved SNR				Recognition
		SS	Kalman	LMS	Lattice	
White	9	12	12	58	56	√ for all
Traffic	6	7	7	23	26	√ for all
Wind	7	10	8	32	32	√ for lattice
Train	7	6	8	41	36	√ for lattice
Air Plane	12	13	14	55	54	√ for lattice

This ascertained the effectiveness of the proposed algorithm. This algorithm was developed in C# language and integrated with the speech recognition module developed in [3] for the Mission Planner Ground Control Station software. The real-time testing was done with the setup as shown in Fig. 5. The speech was uttered with the noise playing in the background and a check was made for the recognition of Mission Planner menu item. The results correlated with MATLAB result.

Figure 5 Laboratory setup for evaluating the performance of voice activated GCS in the presence of noise

Conclusion

This paper proposed a Joint Lattice predictor based speech enhancement technique to achieve robust speech recognition for a Voice activated MAV. The proposed algorithm was integrated with Mission Planner, a commercially available open source MAV-GCS. Its performance was evaluated in the laboratory by uttering MAV commands in the presence of noise obtained from possible MAV operational environments.

Acknowledgment

This work was supported in part by a grant from SIGMA panel, AR & DB.

References

[1] Mallesh Babu S, Mr. Lokesha H, Mrs.Veena S, Mr. Jayantkumar.A.Rathod, “Controlling the Micro Air Vehicle through voice instructions”, International Journal of Computer Engineering and Technology (IJCET), ISSN 0976-6367(Print),ISSN 0976 - 6375(Online), Volume 6, Issue 4, April (2015), pp. 21-27

[2] Kavitha S, Veena S, R Kumaraswamy, “Development of Automatic Speech Recognition system for voice activated Ground Control System”, *Trends in Automation, Communications and Computing Technology (I-TACT-15), 2015 International Conference on.* Vol. 1. IEEE, 2015.

[3] Rahul D K., et al. "Development of Voice Activated Ground Control Station." *Procedia Computer Science* 89 (2016): 632-639.

[4] Prodeus, A. M. "Performance measures of noise reduction algorithms in voice control channels of UAVs." *Actual Problems of Unmanned Aerial Vehicles Developments (APUAVD), 2015 IEEE International Conference.* IEEE, 2015.

[5] Ortega-García, Javier, and Joaquín González-Rodríguez. "Overview of speech enhancement techniques for automatic speaker recognition." *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on.* Vol. 2. IEEE, 1996.

[6] Paliwal, Kuldip K., and Kaisheng Yao. "Robust speech recognition under noisy ambient conditions." *Human-centric interfaces for ambient intelligence. Academic Press, Elsevier* (2009).

[7] NawelSouissi, AdnaneCherif. “Speech recognition system based on short-term cepstral parameters, Feature reduction method and Artificial Neural Networks.”,2nd International conference on Advanced Technologies for signal and image processing,2016: 667 – 671.

[8] De La Torre, Angel, et al. *Speech recognition under noise conditions: Compensation methods*. INTECH Open Access Publisher, 2007.

[9] Veena, S., and S. V. Narasimhan. "Improved active noise control performance based on Laguerre lattice." Signal processing 84.4 (2004): 695-707.

[10] Vaseghi, Saeed V. Advanced digital signal processing and noise reduction. John Wiley & Sons, 2008.

[11] Haykin, Simon S. *Adaptive filter theory*. Pearson Education India, 2008.

[12] Loizou, Philipos C. *Speech enhancement: theory and practice.* CRC press, 2013.

[13] Paliwal, K., and Anjan Basu. "A speech enhancement method based on Kalman filtering." *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87.*. Vol. 12. IEEE, 1987.