A REAL TIME SPEECH TO TEXT CONVERSION SYSTEM USING BIDIRECTIONAL KALMAN FILTER IN MATLAB

Neha Sharma
Dept. Electronics and Communications
Chandigarh University
Gharuan, Mohali, Chandigarh, India

Shipra Sardana
Dept. Electronics and Communications
Chandigarh University
Gharuan, Mohali, Chandigarh, India

***Abstract*— A real time speech to text conversion system converts the spoken words into text form exactly in the similar way that the user pronounces. We created a real time speech recognition system that was tested in real time noiseous environment. We used the design of a bidirectional nonstationary Kalman filter to enhance the ability of this Real time speech recognition system. Bidirectional Kalman filter has been proved to be the best noise estimator in nonstationary noiseous environment. Real time speech to text conversion system introduces conversion of the uttered words instantly after the utterance. The purpose of this project was to introduce a new speech recognition system that is computationally simple and more robust to noise than the HMM based speech recognition system. We have used our own created database for its flexibility and TIDIGIT database for its accuracy comparison with the HMM based speech recognition system. MFCC features of speech sample were calculated and words were distinguished according to the feature matching of each sampled word. System was tested in different noise conditions and we obtained overall word accuracy of 90%**.

Keywords*— *Natural language processing; Speech recognition; speech enhancement; bidirectional Kalman filter.

I. INTRODUCTION

The field of Natural language processing has always been a good research area from past years. There are numerous applications of Natural Language Processing. Speech recognition is one of the most important applications of Natural Language Processing. Speech has always been the most important part of our day to day communication. We express our ideas through a specific language. Computers understand our language (natural language) by speech recognition. Speech or word by word recognition is the process of extracting the attributes of speech and classifying the same attributes with the prerecorded datasets. To recognize a word, word must be passed on to higher-level software for syntactic and semantic analysis. It is a technique of pattern matching, where audio signals are tested and framed into phonetics (number of words, phrases and sentences) [1]. To perform such task one needs to record a voice sample and then convert this voice sample into wav format. Spectrum based parameters are obtained when a word is recognized.

Various statistical methods are used for the analysis of words which give some specific value of words. Words fluctuate between in its bounded range of occurrence. In the improvement of word recognition process, one of the important tasks is to find the most informative parameters of speech signal. To perform such tasks some of the techniques are used Linear Predicted coefficients (LPC) and Mel Frequency Cepstrum Coefficients (MFCC) [2]. By using such techniques new spectrum is obtained that is different from the previous spectrum of spoken words.

Speech intensification is the process of amelioration in comprehensibility or quality of a speech sample when it depraved. Speech enhancement is not only to reduce noise from a speech sample but to de-reverberate and separate the unconstrained signals. It is desirable to enhance the speech because when speech is processed through any of the tool in lab it get influenced with the noise (background noise or otherwise) and individuality of the speech changes with time which affects the whole recognition process.

So, it has become very difficult task for the boffins to find contrivances that really work in different practical environments. However this criterion plays an important role in justifying the performance of the algorithm with reference to quality and comprehensibility. [3]

Kalman Filter is a state estimator that produces an optimal estimate and minimizes the mean square error. Kalman filtering is an effective approach to remove nonstationary noise form background or otherwise. It is a state space model that always distillates the adorable information from the signal which is going to be processed [4]. In its contrivance a system model is first selected and model parameters are estimated from its previous state [5]. The parameters of real time model are selected first and there are so many unknown parameters that are really hard to select. There are so many proposed algorithms on Kalman filtering that shows that it a best parameter estimator in the world [6, 7], Mathe et al. [8] have used the Kalman filter for speech enhancement purpose. The [9] uses the Bidirectional Kalman filter for a robust speech recognition system. In this paper we have used the Bidirectional Kalman filter for the intensification of our speech recognition system in background noise.

II. PROCESS DESIGN OF SPEECH TO TEXT CONVERSION SYSTEM

The process of speech recognition system is divided in two stages, first is training stage and second is the testing stage.

978-1-5090-2029-4/16/$31.00 @2016 IEEE

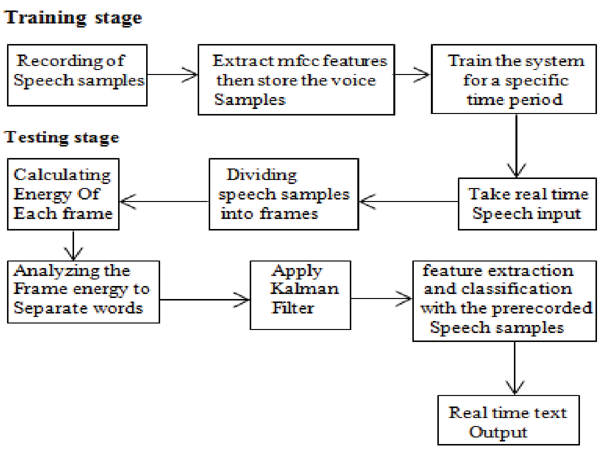

Training stage
Recording of Speech samples
Extract mfcc features then store the voice Samples
Train the system for a specific time period
Testing stage
Calculating Energy Of Each frame
Dividing speech samples into frames
Take real time Speech input
Analyzing the Frame energy to Separate words
Apply Kalman Filter
feature extraction and classification with the prerecorded Speech samples
Real time text Output

Figure 1: Speech Recognition Process

III.TRAINING STAGE

In this stage a database is created by recording some speech samples by the user. Then the recorded speech samples are stored into .wav format in Matlab. After this stage it is necessary to train the speech recognition system.

A. FEATURE EXTRACTION

Every signal is composed of some features/attributes. According to its features we can classify the signal characteristics. In case of speech we extract some of its attributes. Attributes extraction is one of the easiest ways to recognize the speech. Speech is a time varying signal and to deal with such a time varying signal is a difficult task. So attributes of speech play an important role in recognition. To deal with a large sequence of speech is short time attributes are taken on Mel scale (melody of speech). So we decided to extract short time features of speech which are MFCC.

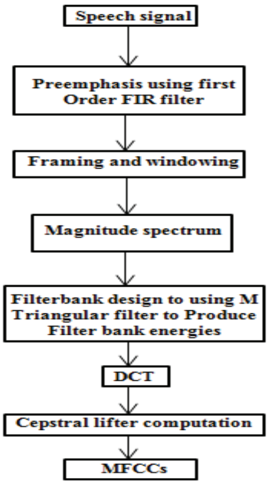

Speech signal
Preemphasis using first Order FIR filter
Framing and windowing
Magnitude spectrum
Filterbank design to using M Triangular filter to Produce Filter bank energies
DCT
Cepstral lifter computation
MFCCs

Figure 2: MFCC Feature Extraction

We have used MFCC attributes for recognition of each word for this system. The word 'Mel' in the MFCCs represents the melody of a speech signal.

MFCC features are based on the human ear perception that means human's ear's critical bandwidth frequencies filters the spaced linearity between the high frequency and low frequency of each word uttered by the user. The human understanding for different frequency ranges of the uttered word is shown on a nonlinear scale. Pitch period of every word is measured with a Mel scale. [11].

B. CREATING THE DATABASE AND TRAINING OF VOICE SAMPLES

To recognize the uttered word of the speaker, a database is created to resemble the pronounced word. To create such database, we first recorded some speech samples.

We have trained this system for 100 words in English Language and three separate samples of each word were taken.

IV. EXPERIMENTAL TESTING

Our speech recognition system was a speaker dependent system. So it was dependent on the user's voice only. In the testing of this system we created a database of nine words. After the training of this system, a real time speech input was given to it through a good quality microphone. The system divided the real time speech sample into small segments of frames or continuous groups of samples. After that the energy of each frame segment was calculated using simple energy formula:

$$E_x = \int_{-\infty}^{\infty} x^2\, dx \qquad (1)$$

Energy calculated was then analyzed by a speech detection algorithm to separate the words.

A. SPEECH DISCLOSURE ALGORITM

The speech disclosure algorithm is applied to detect each word by processing the stored speech samples in database (derived or self-created) frame by frame with a simple loop operation performed using MATLABs. We divided the whole frame into the segment of 160 samples and each of the samples was detected by the system. For the detection of each frame we used a combination of signal energy and a zero crossing rate. This calculation became very simple with the MATLAB mathematical and logical operators.

B. ACOUSTICAL MODEL

It is very important to create an acoustical model for the detection of each uttered word. So we created an acoustical model. It is known that different sounds are produced by human vocal cord and different sounds can have different frequencies. To predict the different frequencies it power spectral density measure can be a better way. So we find out the frequencies by power spectral densities measure.

Speech can be termed as short term stationary so MFCC features were again extracted and word pronounced by the user was detected. The output speech signal was compared with the prerecorded clean speech signal with the correlation figure where we had taken the three voice samples for each of the word training. Correlation figure was calculated using [9]:

$$\rho(w_o, w_r) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} w_{oij} * w_{rij}}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} w_{oij}^2}\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} w_{rij}^2}} \qquad (2)$$

C. *APPLYING BIDIRECTIONAL NONSTATIONARY KALMAN FILTER ALGORITHM*

After the process of word separation Kalman filter removes the unwanted noise and gives the filtered output. The proposed bidirectional Kalman filter algorithm is given in the figure- 2.

Real time speech's (k)' get interrupted with some unwanted noise v(k) while processing.

$$y(k) = s(k) + v(k) \qquad (3)$$

As we know the nature of noise is nonstationary and cannot be measured. In that case Kalman filter is applied to minimize the mean square error of the clean speech signal and the noiseous speech signal to enhance the speech quality.

Equations of the bidirectional Kalman filter algorithm are divided into two parts: *forward run* and *backward run*. In the forward run first we set the initial state estimate P to '0'. Then the second stage is the error prediction e(k) that can be calculated with the Y(k) (signal + noise) and a predicted measurement variable(assumed) $Y_p(k)$.

$$e(k) = Y(k) + Y_p(k) \qquad (4)$$

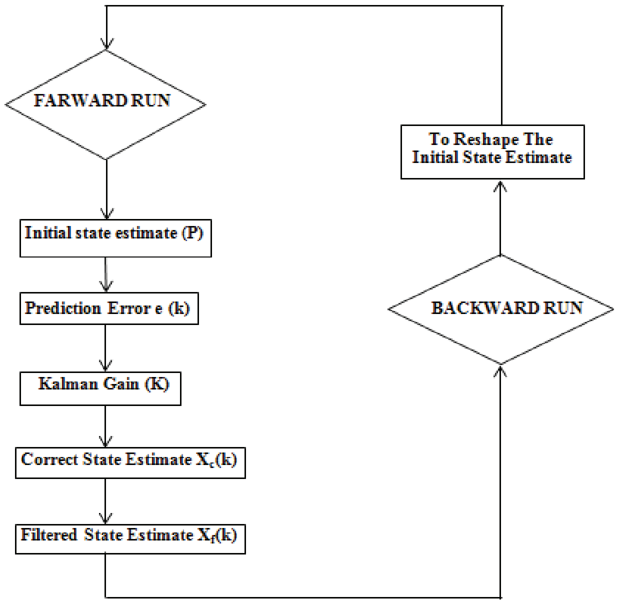

FARWARD RUN
To Reshape The Initial State Estimate
Initial state estimate (P)
Prediction Error e (k)
BACKWARD RUN
Kalman Gain (K)
Correct State Estimate $X_c(k)$
Filtered State Estimate $X_f(k)$

Figure 3: Bidirectional Nonstationary Kalman Filter Loop Operation

As we know the nature of noise is nonstationary and cannot be measured. In that case Kalman filter is applied to minimize the mean square error of the clean speech signal and the noiseous speech signal to enhance the speech quality.

Equations of the bidirectional Kalman filter algorithm are divided into two parts: *forward run* and *backward run*. In the forward run first we set the initial state estimate P to '0'. Then the second stage is the error prediction e(k) that can be calculated with the Y(k) (signal+noise) and a predicted measurement variable(assumed) $Y_p(k)$ [10].

$$e(k) = Y(k) + Y_p(k) \qquad (5)$$

Next step is to calculate the Kalman gain K. Its value can be calculated with the following parameters as [10]:

$$K(k) = \frac{P_p\ (k)C^T}{CP_p\ \ (k)C^T+R} \qquad (6)$$

'C' is the measurement gain matrix and '$P_p(k)$' is the Autocavriance matrix of the predicted state variable 'R' is the Autocovariance matrix of the measurement noise.

The correct state estimate is enumerated with Kalman gain and the predicted state estimate $X_p(k)$ as follows [10]:

$$x_c(n) = x_p(n) + Ke(n) \qquad (7)$$

The filtered state estimate $X_f(k)$ is equal to the correct state estimate. After this the filtered state estimate is applied to reshape the initial state estimate. Loop is executed again and to achieve the best estimate.

V. *RESULTS AND COMPARATIVE ANALYSIS*

Real time results were obtained in different environments. First we tested this system for some sentences accuracy in the presence of fan noise at home. Then we tested the system for same sentences in college laboratory with some background music and also fan noise. In the figure:4 a phrase 'throw down the glove' was uttered by the user through microphone and the 1st figure in this figure shows the utterance plot of all the four words of the sentence. After this the Kalman filter filters the noise word by word separately and provides the correct state estimate.

A. *RESULTS OBTAINED WITH THE TIDIGIT DATABASEIN COMPARISON TO HMM BASED SPEECH RECOGNITION SYSTEM*

First of all we tested our system with TIDIGIT database which constitute the database of recording sampled digits from '0' to '9'. Results were obtained with TIDIGIT dataset with the computation of correlation figure. The results obtained were compared to HMM based speech recognition system's results given in [9]. Good results were obtained that are given in the Table 1 and the results were obtained in different real time scenarios. First we tested this system at home with fan noise. Secondly we tested this system in college lab with background music. After that we tested this system with different SNRs 20dB, 15dB, 10dB, 5dB and 0dB respectively. Correlation figure was obtained with the correlation of similar samples of TIDIGIT dataset and real time filtered speech input. It is observed from the tables that correlation of clean digits and filtered digits of the TIDIGIT dataset gives excellent results in different SNRs while the results obtained with real time speech in different real time noise environments are comparatively having low correlation figure.

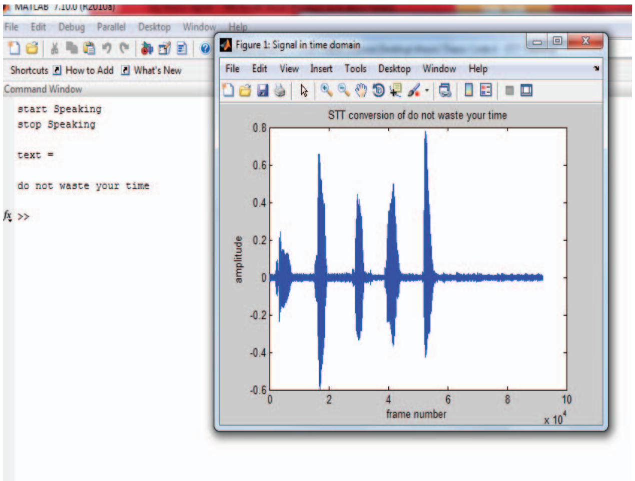
TABLE-I CORRELATION OF CLEAN SPEECH AND THE FILTERED SPEECH OF TIDIGIT DATASET IN DIFFERENT SNRS

Sr. no.	Number of repetitions	At 20db SNR	At 15db SNR	At 10db SNR	At 5db SNR	At 0dB SNR
1.	First	0.972	0.963	0.958	0.938	0.877
2.	Second	0.978	0.969	0.962	0.941	0.884
3.	Third	0.981	0.970	0.967	0.948	0.889
4.	Fourth	0.986	0.972	0.976	0.962	0.898
5.	Fifth	0.989	0.978	0.984	0.963	0.902
6.	Sixth	0.992	0.982	0.986	0.965	0.905
7.	Seventh	0.994	0.984	0.987	0.967	0.907
8.	Eighth	0.996	0.989	0.988	0.968	0.907
9.	Ninth	0.998	0.993	0.989	0.968	0.908
10.	Tenth	0.999	0.996	0.989	0.969	0.909

TABLE-II CORRELATION OF REAL TIME FILTERED SPEECH OUTPUTS WITH TIDIGIT DATASET

Sr. no.	Number of repetitions	At home with fan noise	At college laboratory with fan noise and background music	At 20dB SNR	At 15dB SNR	At 10dB SNR	At 5dB SNR
1.	First	0.892	0.822	0.884	0.838	0.828	0.818
2.	Second	0.897	0.834	0.889	0.840	0.837	0.820
3.	Third	0.899	0.841	0.891	0.842	0.838	0.824
4.	Fourth	0.908	0.844	0.896	0.845	0.841	0.827
5.	Fifth	0.916	0.853	0.898	0.859	0.844	0.829
6.	Sixth	0.927	0.892	0.899	0.862	0.846	0.835
7.	Seventh	0.939	0.896	0.902	0.870	0.850	0.837
8.	Eighth	0.944	0.916	0.906	0.877	0.857	0.842
9.	Ninth	0.953	0.931	0.908	0.881	0.866	0.844
10.	Tenth	0.965	0.944	0.909	0.899	0.870	0.845

B. RESULTS OBTAINED IN DIFFERENT NOISE SCENARIOS

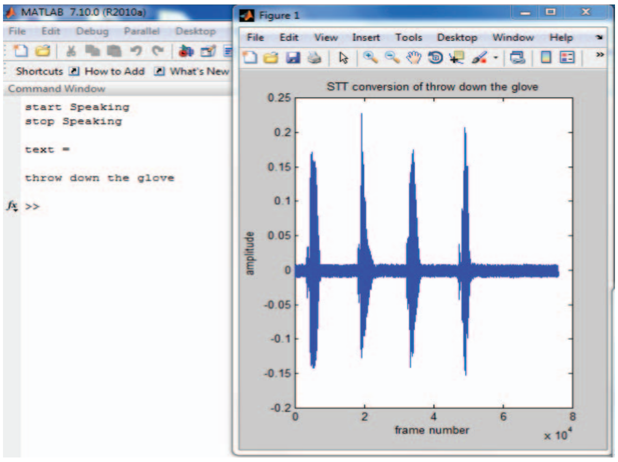

MATLAB 7.10.0 (R2010a)
File Edit Debug Parallel Desktop Window Help
Shortcuts How to Add What's New
Command Window
Figure 1: Signal in time domain
File Edit View Insert Tools Desktop Window Help
STT conversion of do not waste your time
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
amplitude
0
2
4
6
8
10
frame number
x 10^4

```
start Speaking
stop Speaking

text =

do not waste your time

>>
```

Figure 4: STT conversion of ‘do not waste your time’

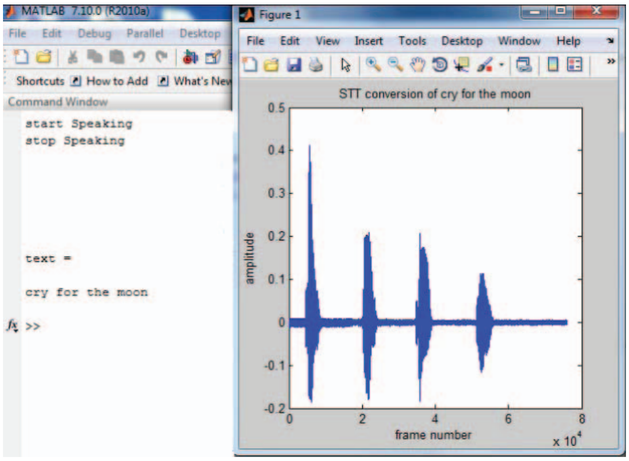
TABLE III CORRELATION FIGURE BETWEEN PRERECORDED CLEAN VOICE SAMPLES STORED AND REAL TIME FILTERED SPEECH OUTPUTS FOR 1ST SENTENCE

Sr. no.	Words	At home with fan noise	At college lab with background noise	At 20dB SNR	At 15dB SNR	At 10dB SNR	At 5dB SNR
1.	Do	0.989	0.888	0.978	0.932	0.886	0.772
2.	Not	0.979	0.878	0.962	0.933	0.922	0.777
3.	Waste	0.977	0.898	0.922	0.952	0.923	0.799
4.	Your	0.978	0.877	0.952	0.899	0.945	0.798
5.	Time	0.986	0.867	0.932	0.912	0.897	0.788


MATLAB 7.10.0 (R2010a)
File Edit Debug Parallel Desktop
Shortcuts How to Add What's New
Command Window
Figure 1
File Edit View Insert Tools Desktop Window Help
STT conversion of throw down the glove
0.25
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
amplitude
0
2
4
6
8
frame number
x 10^4

```
start Speaking
stop Speaking

text =

throw down the glove

>>
```

Figure 5: STT conversion of ‘Throw down the glove’


MATLAB 7.10.0 (R2010a)
File Edit Debug Parallel Desktop
Shortcuts How to Add What's New
Command Window
Figure 1
File Edit View Insert Tools Desktop Window Help
STT conversion of cry for the moon
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
amplitude
0
2
4
6
8
frame number
x 10^4

```
start Speaking
stop Speaking



text =

cry for the moon

>>
```

Figure 6: STT conversion of ‘Cry for the moon’

Table IV Correlation figure between prerecorded clean voice samples stored and real time filtered speech outputs for 2nd sentence

Sr. no.	Words	At home with fan noise	At college lab with background noise	At 20dB SNR	At 15dB SNR	At 10dB SNR	At 5dB SNR
1.	Throw	0.889	0.888	0.978	0.932	0.886	0.772
2.	Down	0.899	0.878	0.962	0.933	0.822	0.777
3.	The	0.877	0.898	0.922	0.952	0.823	0.799
4.	Glove	0.879	0.877	0.952	0.899	0.845	0.798

TABLE V ACCURACY IN CASE OF SENTENCES

Sr. no.	Sentences	Recognition at college lab with background noise	Recognition at home in fan noise and background music
1.	Do not waste your time	Yes	Yes
2.	Earth revolves around the Sun	Yes	Yes
3.	Throw down the glove	Yes	Yes
4.	Break the ice	Yes	Yes
5.	Cry for the moon	No	Yes

$$\text{Overall sentence accuracy} = \frac{9}{10} \times 100 = 90\%$$

TABLE VI ACCURACY IN CASE OF DIFFERENT SNR CONDITIONS

Sr. No.	Sentences	Different SNRs					
		20dB	15dB	10dB	5dB	2dB	1dB
1.	Earth revolves around the sun	100%	100%	100%	100%	100%	80%
2.	Break the ice	100%	100%	100%	100%	80%	60.6 %
3.	Throw down the glove	100%	100%	100%	100%	100%	75%
4.	Cry for the moon	100%	100%	100%	100%	100%	75%
5.	Do not waste your time	100%	100%	100%	100%	100%	80%

VI.CONCLUSION

We have developed speech recognition system and used a bidirectional Kalman filter for its enhancement. We created our own database in Matlab and also used TIDIGIT dataset for its comparison analysis. We concluded that in comparison with the TIDIGIT dataset this system yields excellent results in real time scenarios while with our own created database the results are comparatively poor. Only five sentences and five similar words are shown in this paper. We tested this system in a real time environment for different noise conditions and compared the output with the prerecorded speech samples with the correlation figure. First we tested this system at home with fan noise. System was able to recognize each word and gave us 100% accuracy in that case. After that we tested this system at college laboratory with fan noise and background music. It was not able to recognize a single word ‘moon’ in the sentence named ‘cry for the moon’. At this scenario system was 80% accurate according to the sentence recognition. Based on this analysis our system is overall 90% accurate.

This system’s accuracy can be increased in more different noise scenarios. From the table VI it is concluded that the accuracy of the system is 100% up to 5dB SNR. After 5dB SNR the accuracy drops.

By using this code the system can be trained for more words and paragraphs. Kalman filter removes background noise very efficiently. But this filter takes large time to filter out noise. In case or continuous acquisition of speech it can take more time to display text because of its word by word filtration process.

ACKNOWLEDGEMENT

I would like to thank all the people who contributed in some way to the work described in this paper. With deep sense of gratitude I express my sincere thanks to my esteemed and worthy supervisor, Er. Shipra Sardana **,** Assistant Professor**,** ECE Department Chandigarh University, Mohali, Chandigarh for her valuable guidance in carrying out this work under her effective supervision, encouragement, enlightenment and cooperation. My special thanks also go to Dr. Amit Gupta, Prof. ECE Department, Chandigarh University for his motivational guidance throughout this Master’s Program to focus on the right problem and also to write a good research paper.

REFERENCES

[1] J. D. Tardelli, C. M. Walter, “Speech waveform analysis and recognition process based on non-Euclidean error minimization and matrix array processing techniques”. IEEE ICASSP, pp. 1237-1240, 1986.

[2] Takao Suzuki, Yasuo Shoji, “A new speech processing scheme for ATM switching systems”. IEEE,Digital Communications Laboratories, Oki Electric Industry Co. Ltd., Japan, pp. 1515-1519, 1989.

[3] J. S. Lim "Evaluation of a correlated subtraction method for enhancing speech degraded by additive white noise", IEEE Trans. Acoustics, Speech and Signal Processing, vol. ASSP-26, no. 5, pp.471 -472 1978.

[4] R. E. Kalman and R. S. Bucy, “New results in linear filtering and prediction theory,” Trans. ASME Series D, J. Basic Engineering, pp. 95-108, 1961.

[5] Gabrea, M.: ‘Adaptive Kalman filtering-based speech enhancement algorithm’. IEEE Canadian Conf. on Electrical and Computer Engineering, 2001, vol. 1, pp. 521–526.

[6] Jeong, S., Hahn, M.: ‘Speech quality and recognition rate improvement in car noise environments’, Electron. Lett., 2001, 37, (12), pp. 800–802.

[7] Ma, J., Deng, L.: ‘Efficient decoding strategies for conversational speech recognition using a constrained nonlinear state-space model’, IEEE Trans. Speech Audio Process., 2003, 11, (6), pp. 590–602.

[8] Mathe, M., Nandyala, S.P., Kishore Kumar, T.: ‘Speech enhancement using Kalman filter for white, random and color noise’. IEEE Int. Conf. on Devices, Circuits and Systems (ICDCS), 2012, pp. 195–198.

[9] Yeh Huann Goh, Paramesran Raveendran, Yann Ling Goh, “Robust speech recognition system using bidirectional Kalman filter”, IET Trans. Pp. 1751-9675, 2015.

[10] Tony Lacey, “Likelihood interpretation of Kalman filter”, tutorial lecture, April 2006.

[11] Siva Prasad Nandyala, Dr.T.Kishore Kumar,’’ International Journal of Computer Applications’’ Volume- 12, pp.0975 – 8887, November 2010.

[12] Paliwal, K.K, Atal, B.S. Efficient vector quantization of LPC Parameters at 24 bits / frame. IEEE Trans. Speech Audio Process, pp.3- 14, 1993.

[13] W.B. Kleijn and K.K.Paliwal, “An introduction to Speech coding,” Speech coding and synthesis, Elsevier science, 1995, pp.1-47.

[14] J. Makhoul. S. Roucos. and H. Gish, "Vector quantization in speech coding," Proc. IEEE. vol 73, pp. 1551-1588, Nov.1985.

[15] Sharon Gannot, David Burshtein , Ehud Weinstein, “Iterative and Sequential Kalman Filter-Based Speech Enhancement Algorithms” IEEE Transactions on speech and audio processing, vol.6, pp. 373 – 385, july 1998.

[16] Zenton Goh , Kah-Chye Tan , Tan, B.T.G., “Kalman filtering speech enhancement method based on a voiced and unvoiced speech model”, IEEE Transactions on speech and audio processing, vol.7, pp. 510 – 524, August 1999.

[17] Ki Yong Lee and Souhwan Jung, “Time–Domain Approach Using Multiple Kalman Filters and EM Algorithm to Speech Enhancement with Nonstationary Noise”, IEEE Transactions on speech and audio processing, vol.8, pp. 282 -291, 2000.

[18] Lee, Ki Yong, Lee, Ki Yong, “Recognition of noisy speech by a nonstationary AR HMM with gain adaption under unknown noise”, IEEE Transactions on speech and audio processing, vol.9, pp. 741 – 746, 2002.

[19] Chi-Chou Kao, “Design of Echo Cancellation and Noise Elimination for Speech Enhancement”, IEEE Transactions on Consumer Electronics, Vol. 49, No. 4, NOVEMBER 2003

[20] Ma, J.Z., Li Deng, “Efficient decoding strategies for conversational speech recognition using a constrained nonlinear state space model”, IEEE Transactions on speech and audio processing, vol.11, pp. 590 – 602, January 2004.