IAETSD International Journal of Multee-Disciplinary Research, Volume 3, Issue 2, Mar-April/2017
ISSN (Online): 2454-1524

Isolated Words Recognition in Speech Signal Processing

RAHUL KHANNA1, Dr.S.KALAIVANI2
[1,2]School of Electronics Engineering, VIT University, Vellore, India
[1]rahulkhanna9953@gmail.com, [2]kalaivani.s@vit.ac.in

ABSTRACT.

An automated algorithm for recognizing the isolated spoken words is presented in this paper. Hilbert-Huang Transform (HHT) and Adaptive Time-Frequency Masking are used to extract features of the speech signal by Mel-frequency cepstral coefficients (MFCC). The extracted features data obtained are used to train and formulate the neural. The training and testing of the algorithm is done on real data set.

Index Terms—**isolated words, speech signal, recognition, Hilbert-huang transform, Mel frequency cepstral coeffecients.**

I. INTRODUCTION

The Automatic Speech Recognition (ASR) system majorly depends on two important steps, feature extraction and decision making. The complexity of different ASR models differs mainly in these two steps only. Feature extraction is the step of converting the recorded speech data into some meaningful data which more clearly describes the information that it carries. .For feature extraction many of the techniques available in the literature. Some of commonly used feature extraction techniques are Perceptual Linear Prediction (PLP)[1], Relative Spectra Filtering of Log Domain Coefficients PLP (RASTA-PLP) [2] , Linear Predictive Coding[1], Linear Predictive Cepstral Coefficients (LPCC) [1], Mel-Frequency Cepstral Coefficients (MFCC) [1,3] and some other auditory features; such as first order derivative, energy normalization, formant [1,3].

Decision making is the step to recognize the recorded speech by using the data, collected from the training data set. This step uses the concepts of probability and statistics to formulate the algorithm for correct pattern recognition. Some widely use algorithms are Hidden Markov Model (HMM) [3, 4], Gaussian Mixture Model (GMM) [1], Dynamic Time Warping (DTW)[2], Decision Tress[5]. Nowadays, Artificial Neural Networks techniques [5] are more preferred and reliable than above techniques. Some ANN techniques are Multi-Layer Perception (MLP)[4], Recurrent Neural Network[5], Convolution Neural Network [5] and Support Vector Machines (SVMs)[5].

But the presence of unwanted noise in the recorded speech signal leads to reduction of efficiency of these two important steps. To mitigate this loss, adaptive time-frequency masking and Hilbert-huang transform techniques are used in this paper. MFCC technique is adopted for the feature extraction step. This paper is organized as follows: Section 2 discusses the related methods and mathematical operations used to implement the proposed method. Section 3 discusses the proposed method. Section 4 shows the experiment and results. Section 5 discusses the conclusion and future enhancements.

II. RELATED WORK

This section provides the brief information of the speech recognition system, HHT, MFCC, Linde-Buzo Gray Algorithm, and Adaptive Time Frequency Masking

A. Isolated Word Recognition System (General Method)

Speech recognition system implements the acoustic model to extarct the features of the speech signal. Acoustic model used in this paper is Mel-Frequency Cepstral Coefficients (MFCC) [1]. This system can be implemented for either known speakers or for unknown speaker. Known speaker system means that system can only recognize the words spoken by those users whose voice data is already stored in its memory. Unknown speaker system is the system which can be used or operated by any user. Here, unknown speaker means that voice data of that speaker is not stored in system’s memory and known speaker means that voice data of that speaker is stored in system’s memory.

To Cite This Article: RAHUL KHANNA and Dr.S.KALAIVANI ,. **Isolated Words Recognition in Speech Signal Processing. International Journal of Multee-Disciplinary Research; Pages: 264-270**

General steps of SRS are:

1. Speech signal is recorded at f_s sampling rate.
2. Recorded signal is preprosed using first order high pass filter,$S[n] \;=\; S[n] - 0.95 * S[n-1]$ [3,2].
3. Signal is windowed using hamming window of 25ms with overlapping of 25% [1].
4. Spectrum of windowed signal is used to obtain MFCC features, delta and delta delta features
5. Then these extracted features are mapped using vector quantization algorithm.
6. Mapped data is use to form train and test data sets, which are finally used for decision making step.

B. Hilbert-Huang Transform (HHT)

Usually, Fast Fourier Transform is use to convert signal from time domain to frequency domain. But, the drawback is, FFT assumes the data to be stationary and linear. But, speech signal are not stationary. So to construct more physically meaningful system HHT is implemented in two steps [3,4,6]:

B1.Empirical Mode Decomposition process

This step finds all the intrinsic mode functions (IMFs) in the signal $x(t)$as,

a. Find the local peaks $P_{max}(t), P_{min}(t)\ of\ x(t)$ signal
b. Form the cubic spline of these peaks
c. Compute the average, say avg, of splines
d. Perform $c(t) \;=\; x(t) - avg$
e. Calculate the number of zero-crossings, z, and peaks, p, in $c(t)$; to check the monotonicity of $c(t)$.
f. If $|z-p| \;>\; 1$ repeat steps: ‘a’ to ‘e’, otherwise calculate residue $r(t) \;=\; x(t) - c(t)$
g. Repeat steps ‘a’ to ‘f’ till r(t) have $|z-p| \;\leq\; 1$ such as $x(t) \;=\; \sum c(t) \;+\; r(t)$

Every summing vector in above equation is an IMF of $x(t)$

B2. Amplitude and frequency extraction

It is a step to find frequency domain data of the signal x(t)

a. Calculate the Hilbert transform of x(t), h(t)
b. Form the analytical value, y(t), by combining the value calculated in above step as imaginary value and x(t) as real value. $y(t) \;=\; x(t) \;+\; i * h(t)$ [6]
c. Now calculate the DCT of y(t) [7]

C. Mel Frequency Cepstral Coefficients (MFCC)

The acoustic model uses the MFCC model and 12 MFCC features are extracted from speech signal. MFCC uses the 20 overlapped warped triangular shaped filters. In mel scale these filters are equally spaced. Since the relation between mel and frequency scale is non- linear, these filters are termed as warped filters [3]. Relation between mel and frequency scale are given by Eq.1 [3]

$$m = 2595\ log10\ (1 + \frac{f}{700}) \qquad (1)$$

Steps for MFCC are as follows [1]:

a. Set the range of filter bank, R having lowest and highest filter bank frequency values
b. Set the number of desired coefficients from each filter, say k
c. Get 20 equally spaced values between range R in mel scale
d. From the above calculated values and k, form 20 overlapped triangular shaped filters in frequency scale, using Eq.2 [1]

$$H_m[k] = \begin{cases} \frac{2(\mathrm{k}-\mathrm{f}[\mathrm{m}-1])}{(\mathrm{f}[\mathrm{m}+1]-f[m-1])(f[m]-f[m-1])} & f[m-1] \leq k \leq f[m] \\ \frac{2(\mathrm{f}[\mathrm{m}-1]-\mathrm{k})}{(\mathrm{f}[\mathrm{m}+1]-f[m-1])(f[m+1]-f[m])} & f[m] \leq k \leq f[m+1] \\ 0 & else \end{cases} \qquad (2)$$

e. Pass the signal through these filters and obtain filter output.
f. For each filter perform sum of logarithm of each filter output, to obtain 20 different values
g. Now perform DCT on these values to obtain the time domain signal, set n as 12
h. Calculate the delta and delta delta of MFCC and their log energy features are calculated.

D. Linde-Buzo-Gray Algorithm

It is a vector quantization technique use to form the codebook for the given data. It is similar to K-mean algorithm.
Steps to form codebook are as follows [8,9]:

a. Decide the K value, it the number of centroids needed to form codebook, and a variable err, it is a small number to show deviation in values from centroid
b. Calculate the mean vector (centroid c) and mean square error, say D, of deviation of every data vector from the centroid.
c. Now set variable m as 2 and centroid with range c-err to c +err. Using these values assign centroid to each data vector.
d. From above data update codebook by calculating new centroid for each area
e. Now calculate the mean square error for deviation of every data vector from their respective centroids, say D'
f. Change the value of m to 2*m if value D' < D, otherwise perform steps 'c' to 'e' again
g. Repeat steps 'b' to 'f' till value of m is less than that of K.

E. Adaptive Time-Frequency Masking:

The adaptive mask is estimated for every frame. This method utilizes the concept of constant noise energy, to build an adaptive mask for every frame.This technique includes three steps [3]:

E1. SNR estimation

This step calculates the SNR of speech signal. SNR is the measure to calculate the amount of noise present in speech signal. But, the non-uniformity of speech signal creates problem to calculate SNR. So, SNR of speech signal is calculated by dividing the total speech and noise signals in speech. Also, present of short time energy in silent frame is called as noise and that noise signals have low STE than speech signals [3,10]. So, calculating the noise from low STE frames and speech from higher STE frames, SNR is estimated. Steps for calculating ESNR are:

a. Short time energy of every frame is calculated
b. SNR is estimated using Eq.3

$$ESNR = 20\log \sum_{i=1}^{k} \frac{STE_i - k \cdot \min(STE_i)}{k \cdot \min(STE_i)} \qquad (3)$$

E2. Time Frequency Masking

In this step the mask is created and is applied on the frame. In this step actual function of the technique happens, i.e. to concentrate the speech information to lower frequencies. Steps are:

a. Convolving frame with 11*11 mean filter [11].
b. Now normalize and dc shift the convolution result by subtracting it with minimum value and dividing it by the largest value
c. Set the threshold value using eq.4, where a=0.047 and b=0.8.

$$Threshold = a * b^{ESNR} \qquad (4)$$

d. If the value in mask is less than threshold than value of that point is 1, otherwise it is 0.1

E3. Mean Smoothing Filtering

Since the noise power of AWGN keeps on changing from frame to frame. So to remove this instant variation, mean smoothing filter is to adaptive time frequency masked map. Average of 3 frequency frame is calculated.

III. PROPOSED METHOD

The proposed method is implemented with the desire to reduce the noise and to construct more physically realistic isolated word recognition system. Assumption used is that only Additive White Gaussian Noise (AWGN) is present in this system. Proposed method is designed to recognize digits 0-9 separately. Fig. 1 shows the steps implemented in the proposed method.

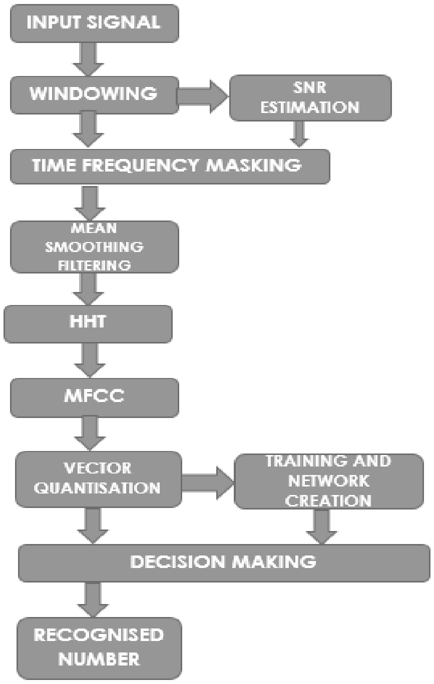

Fig. 1 Block diagram of r isolated word recognition system

In speech signals, most of the useful data is located in lower frequency regions only [1, 3]. So removing the highest frequencies from signal will not much affect the information of the speech. Also, the energy of this noise is almost constant. Fig 2 shows the spectrum of input signal obtained by general system. So to remove noise different approach is taken. For this new approach, firstly the pre-processing step is avoided, as this step boosts the noise amplitude and then FFT is replaced by HHT, to make system more realistic.

In proposed method the unwanted noise is reduced by the use of adaptive time frequency masking technique. Then the HHT of the masked signal is used for extraction of MFCC features. Then the extracted features are fed to decision making algorithm, which is based on Sum of Squared Error (SSE) method.

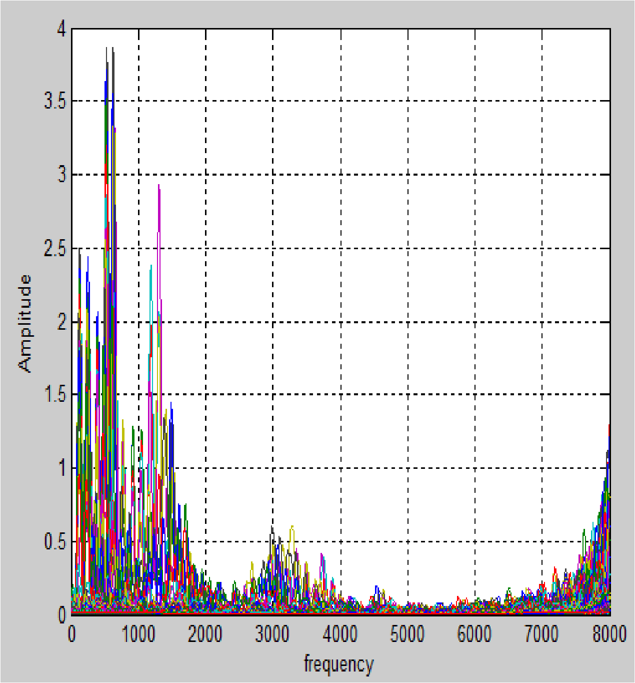

Fig 2. Spectrum of input signal, obtained by general system

IV. EXPERIMENT AND RESULTS

A. Experiment

The proposed method is implemented and compared with the state of the art systems like adaptive time frequency masking and with FFT. The system is trained on the speech data recorded by the three speakers; each of them three times pronounces the 0-9 English numbers. The numbers are recorded at 16 kHz.

In isolated word recognition system each number signal is framed by 25ms overlapped hamming window with the overlapping of 25%. The ESNR is calculated for entire signal. This ESNR will be utilized later in the adaptive time frequency making. Then, HHT of each frame is calculated and adaptive time frequency masking is applied on each HHT frame. Then, the MFCC and dynamic features are calculated. Extracted features are clustered by LBG algorithm.

The testing of the system is also implemented on the speech data recorded by the same users that had contributed for training purpose. Also, one more unknown user's speech data is recorded to find the stability of designed system to unknown user. Each speaker, three times pronounces the 0-9 English numbers. The numbers are recorded at 16kHz. All the training and testing is done in MATLAB [10].

For the general system, FFT and pre emphasizing are utilized with no HHT and no masking. Also, in every system, two sub-system are made. One which utilizes vector quantization technique and one which do not utilize it.

B. Results

In this section the experimental results are presented. The results are presented in three sub-section; improvement in estimated signal to noise ratio, unknown speaker's word recognition rate and improvement in known speaker's word recognition rate. The decision making is based on statistical results given by Sum of Squared Error (SSE) method [13]. The training set's digit with minimum SSE is the recognised ouput.

B1. Improvement in estimated signal to noise ratio

Improvement in the estimated signal to noise ratio is presented in Table 1. The table shows the ESNR of the input signal and that of maked signal. Table 1 shows that proposed method reduces noise in the input signal by atleast 25%. Fig 4 shows the spectrum of input signal otained by proposed system. By comparing Fig 3 and Fig 4, graphically, it can be said that ESNR is increased.

B2.Unknown speaker's word recognition rate

The word recognition rate of unknown speaker's data set is zero. This means that given system cannot be used by the unknown speakers.

B2. Improvement in known speaker's word recognition rate

Table 2 shows improvement in considered systems with respect to general system. The word recognition rate varies from sub-system to sub-system. The improvement by proposed system for the sub-system with LBG algorithm is 3% and for the sub-system without LBG algorithm is 23.33%.

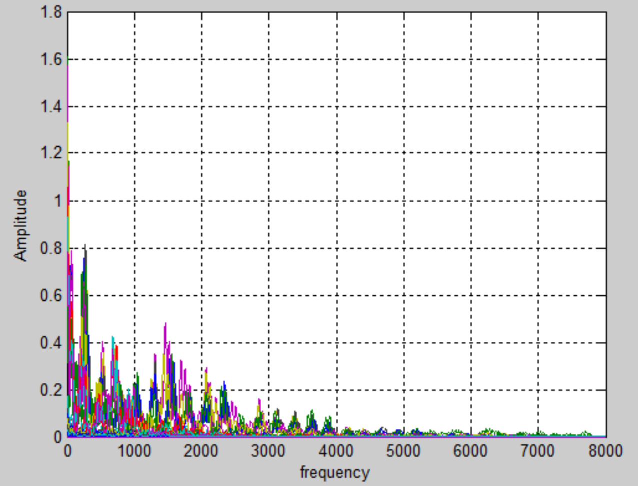

Fig 4: Spectrum of input signal, obtained by proposed system

V. CONCLUSION AND FUTURE WORK

In this paper a proposed method has been developed to increase the word recognition rate of the isolated word recognition system. It utilized HHT to make system physically more meaningful by considering non-stationary data also. It also utilized adaptive time-frequency masking technique to reduce noise from the speech data.

TABLE1: Improvement in ESNR

Digit spoken (0-9)	**ESNR**		**Percentage Improvement by masking**
	Before masking	**After masking**	
0	76.09	110.26	44.91
1	53.61	84.13	56.92
2	84.53	121.9	44.21
3	51.67	83.57	61.73
4	85.9	109.8	27.8
5	82.45	105.72	28.22
6	79.01	102.12	29.25
7	84.71	115.87	36.78
8	78.08	100.56	28.8
9	74.25	107.62	44.94

TABLE2: Improvement by systems with HHT+masking and FFT+masking

Sub-System	**FFT + MASKING**	**HHT + MASKING**
WITHOUT LBG	13.33%	23.33%
WITH LBG	3%	3%

The performance of the proposed system was examined on the data set recorded by four speakers. The experimental results show that the proposed method cannot be used by unknown speakers. Also, it shows the improvement in the ESNR by at least 25% and thus the improvement in isolated word recognition by 3% and 23.33% in sub-systems in which one utilizes vector quantization technique and other do not respectively.

Our future work is to improve the decision making step by utilizing the knowledge and concepts of artificial neural networks, to increase the word recognition rate.

VI. REFERENCES

[1] L. R. Rabiner and Biing-Hwang Juang, “Fundamentals of speech recognition”, Printice HallSignal Processing Series, Alan V. Oppenheim, Series, Editor, 1993

[2] S. Pannirselvam and G. Balakrishnan, “Comparative study on preprocessing techniques on automatic speech recognition for Tamil language”, IJCA, 0975-8887, NCRIIAMI, 2015

[3] A. M. Gouda, M. Tamazin and M. Khedr, “Robust Automatic speech recognition system based on using adaptive time frequency masking”, IEEE, pp.181-186, 2016

[4] Fenglei Ma, X. Zhou and Y. Zheng, “Voice signal noise reduction based on Hilbert-huang transform”, Applied Mechanics and Materials, ISSN: 1662-7482, Vols. 303-306, pp. 1039-1042, 2013

[5] W. Song and J. Cai, “End to end deep neural network for automatic speech recognition system”, Stanford NLP group, 2015

[6] Huang, N. E., and Z. Wu, “A review on hilbert-huang transform: Method and its applications to geophysical studies”, Rev. Geophysics., 46, RG2006, 2008

[7] Vani H.Y., “Hilbert-huang transform based speech recognition”, Second International Conference on CCIP, 2016

[8] MIT, “Vector quantization and clustering”, Lecture-6, Session 2003, 6.345 Automatic Speech Recognition

[9] Roma Bharti and Priyanka Bansal, “Real time speaker recognition system using MFCC and vector quantization technique”, IJCA, 0975-8887, Vol. 117 – No. 1, May 2015

[10] M. Saleem, Z. U. Rehman, U. Zahoor, A. Mazhar and M. R. Anjum, “Self learning speech recognition model using vector quantization”, IEEE, pp.199-2013,2016

[11] R. C. Gonzalez, R. E. Woods and S. L. Eddins,“Digital image processing using MATLAB”, 2nd edition, 2002

[12] Stephen J. Chapman, “MATLAB Programming with Applications for Engineers”, 1st edition, 2013

[13] A. Lipeika, J. Lipeikiene and L Telksnys, “Development ofisolated word recognition system”, INFORMATICA, Vol. 13, No. 1, 37–46, 2002

BIBOGRAPHY

Dr. S. Kalaivani : is currently working as Associate Professor in VIT University, Vellore. She received her B.E. degree in Electronics & Communication Engineering from Government College of Engineering, Salem, Tamil Nadu under Madras University, India in 2000 and M.E. degree in Communication Systems from Anna University, Chennai, India in 2006. She completed her Ph.D. in 2012 under Anna University, Chennai with the specialization of Image Processing. She is having 15 years of teaching experience in under graduate and post graduate level of Engineering courses in Anna University affiliated colleges. She has presented her research papers in 15 international, National Journals and 10 International conferences, more than 15 National conferences. Her research interest includes speckle reduction in SAR and ultrasound images, Multi-resolution processing, Laplace transforms and diffusion

RAHUL KHANNA is UG student in School of Electronics Engineering, VIT University, and Vellore. He is working in the area of speech signal processing. His research interest includes transform domain signal processing, multi scale analysis and automated algorithm design.