Speaker Independent Kannada Speech Recognition using Vector quantization

M.A.Anusuya
Asst.Prof. Dept. of CS&E SJCE,Mysore

S.K.Katti
Proffessor, Dept. of CS&E SJCE,Mysore

ABSTRACT
In this paper a statistical method is used to remove the silence from the speech signal. This method is applied on vector quantization technique to identify the minimum speech patterns that are required while creating the training set of the speech samples. This paper also discusses the importance and efficiency of the algorithms used in vector quantization for the clustering purpose. Also speech recognition accuracies for speaker dependent and speaker independent methods have been evaluated and tabulated in the tables given below. The paper shows the importance of the statistical method analysis of the signal than the normal analysis.

Keywords
Speech Recognition, Isolated word, Uncertainty, Vector quantization, Euclidean distance.

1. INTRODUCTION
Automatic speech recognition is an active research topic for the researchers. With the advent of digital computing and signal processing, the problem of speech recognition was clearly posed and thoroughly studied. These developments were complemented with an increased awareness of the advantages of conversational systems. The range of the possible applications is wide and includes: voice controlled applications, command control applications, dialog system, fully featured speech-to-text software, automation of operator-assisted services, and voice recognition aids for the handicapped.

Isolated word recognition is based on the premise that the signal in a prescribed recording interval consists of an isolated word, preceded and followed by silence or other background noise. Thus, when a word is actually spoken, it is assumed that the speech segments can be reliably separated from the nonspeech segments. (Clearly, in the case when there is no speech in the recording interval, a request to repeat the spoken word must be made.) The process of separating the speech segments of an utterance from the background, i.e., the nonspeech segments obtained during the recording process, is called endpoint detection. These background noise, unvoiced segment are called as uncertainties in speech. These parts increase the storage area of memory and also the computation time. In isolated word recognition systems, accurate detection of the endpoints of a spoken word is important for two reasons, namely:

1) reliable word recognition is critically dependent on accurate endpoint detection

2) the computation for processing the speech is minimum when the endpoints are accurately located.

According to Ladefoged [1], one of the main difficulties in the acoustic analysis of speech is the lack of possibility of analyzing the original sound. It occurs because when the sound is stored through analogical or digital devices what is analyzed is not the produced sound, but the captured one instead. Even if the sound is captured in a soundproof booth it brings with it a series of uncertainty because of the characteristics of the human speech natures and the intrinsic imprecision's of the circuits used in the process. Moreover, the conversion of analogical audio into digital audio for computational analysis involves two discretizations, sampling, in time domain, and quantization, in the amplitude domain [2]. As well as any substitution of a infinite model for a finite one, the sampling and the quantization produce errors [3] which reflect directly in the accuracy of the speech recognition results. For any type of the application these uncertainties exits and these has to be solved for the better recognition accuracies irrespective of isolated, phoneme, continuous speech , spontaneous speech recognition. Hence signal pre-processing and feature extraction phase plays a major in the recognition experiments. These uncertainities can be well solved with statistics and fuzzy concepts as discussed in [1,15]

2. RELATED WORK
Different approaches in speech recognition have been adopted. They can be mainly classified into Template approach, Stochastic approach, Dynamic Time warping approach, Vector quantization approach, Connectionist approach, Artificial intelligence approach, support vector machine approach. Depending the application and the requirement of the problem different approaches are used[7]. But the introducing of speech to HMM has made an impact and has enabled great progress during these last few years. However, there is a lot to be accomplished in this area in order to improve their quality, i.e. the reinforcing of the discrimination between different models, which seems to be very promising. Also creating of minimum speech reference pattern for the use of HMM and for any other application is very crucial. Hence this paper shows the minimum reference patterns that can be used to create speech patterns.

The speech recognition system relied on measuring spectral resonances during the vowel region of each digit. In the 1960s several fundamental ideas, such as filter bank spectrum analysis, zero crossing analysis, time-normalization methods, in speech recognition were published[6,8]. Dynamic programming method for time aligning of a pair of speech utterances was proposed by [9]. In the 1970s isolated word recognition became advanced technology due to fundamental studies by Velichko[10].Speech recognition systems were made truly speaker independent In the 1980s a focus of research was the problem of connected word recognition. Speech research was shifted from template based approaches to statistical modeling methods, hidden Markov model (HMM) approach and neural network methods. In the 1990s main focus of research was large vocabulary continuous speech recognition, robust speech recognition including syntax, semantics, pragmatics into speech recognition higher level [11,12,13,14]

3. FEATURE EXTRACTION

Feature extraction is a special form of dimensionality reduction. When the input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant (much data, but not much information) then the input data will be transformed into a reduced representation set of features (also named features vector). Transforming the input data into the set of features is called *feature extraction*. Once the features are extracted selection of good features also matters in reduction of computation and storage of the information, which are used in better representing the data. If the features extracted are carefully chosen it is expected that recognition accuracy will be increased by minimizing the classification error.

An important property of feature extraction is the suppression of information irrelevant for correct classification, such as information about speaker (e.g. fundamental frequency) and information about transmission channel (e.g. characteristic of a microphone). The feature measurements of speech signals are typically extracted using one of the following spectral analysis techniques: LPC,MFCC. Currently the most popular features are Mel frequency cepstral coefficients MFCC and their delta co-efficients provides good accuracy[3]. This paper also uses MFCC algorithm for feature extraction and the silence removl technique proposed by G.Shah et.al,. Comparision of the speech reference pattern requirement with different code book sizes are evaluated from the proposed paper with the paper [16].

4. SPEECH DATABASE CONSTRUCTION FOR TRAINING AND TESTING PHASE

4.1 Construction of speech Database for training:

An adult female, native speaker of Kannada was asked[4] to utter the Kannada words(1 through 10, see table-1), and her voice was sampled at 8KHz Using mono channel. The figure 1 shows the fundamental frequency determination of spoken digit Idu and its spectrum representation. Each speaker was asked to utter each word 10 times. Totally 100 signal were captured for the training purpose. The speech samples were collected from the ordinary environment where back ground noise of fan, outside noise is considere.The speech signal of each word was then isolated from silence by using statistical method that is discussed in the paper [5]. The samples were then stored in ascending order: first, the ten samples corresponding to word one("Ondu") were stored, then the ten samples of two and so on. To acquire the speech signal, PRAAT software is used.

TABLE I

Number	Kannada Word	Symbol used in the paper
1	"Ondu"	One
2	"Eradu"	Two
3	"Muru"	Three
4	"Nalaku"	Four
5	"Iydu"	Five
6	"Aaru"	Six
7	"Yelu"	Seven
8	"Entu"	Eight
9	"Ombatthu"	Nine
10	"Hatthu"	Ten

4.2 B. Construction of database of testing phase:

Among the ten utterances of each digit 3 utterances were used for the testing purpose. Totally we had 30 test signals . For speaker dependent speech recognition test utterances were taken with the training set and for speaker independent test signals were taken out of training set .

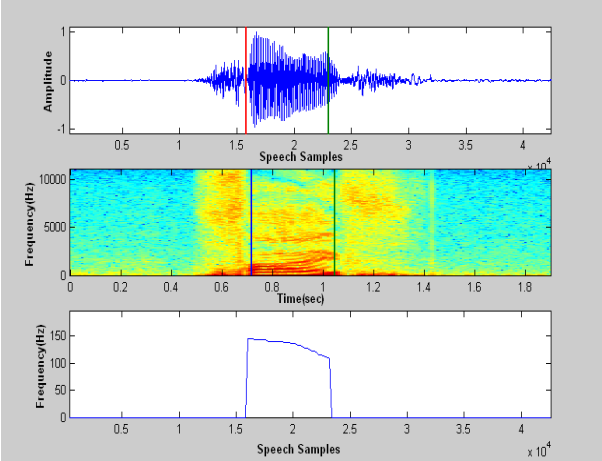


Figure 1. Fundamental frequency determination of a spoken digit Idu. The voiced segment is shown between the two lines. (a) time signal, (b) spectrum of the signal, (c) smoothed contour of the fundamental frequency

5. STEPS TO CALCULATE OF MFCC CO-EFFICIENTS:

- Silence Removal algorithm [5]
- The fist step is pre-emphasis which basically makes one sample in speech influence the next sample by a certain weight.
- S1(n) = s(n) - a*s(n-1)
- Signals are framed into frames of 160 samples in length
- Overlap between successive frames is kept at 80 samples
- Each frame is multiplied by a 160 point Hamming window. This step is primarily to have a smooth transition between samples of a frame.
- FFT is applied to obtain the magnitude frequency response of each frame.
- The next step is to pass the data through Mel filters. Mel filters are triangular equidistant filters in Mel scale, which is a logarithmic scale.
- Taking logarithm of this we obtain Mel spectrum coefficients.
- The final step in obtaining MFCC is performing discrete cosine transform (DCT) on the Mel cepstrum coefficients. The output of DCT is Mel-cepstral Ist order coefficients of 13values.
- The 2nd order coefficients called delta coefficients are calculated. In this 24 cepstral coefficients are calculated.

- The 3rd order called delta-delta coefficients are calculated obtaining 39 coefficients. Energy coefficients are also considered.

At the end of the feature extraction 39 coefficients for each signal is obtained. Then these features are fed into vector quantization algorithm to form the group of clusters for each words. Totally we have ten groups for the ten elements of the words one to ten. Vector quantization algorithm is applied for both speaker dependent and speaker independent signals. To decrease the problem of feature vectors for speaker dependent/independent recognition task, two clustering algorithms in vector quantization approach is used namely VQ1 and VQ2. VQ1 was developed by Juang *et al.* (1982) based on binary splitting algorithm i.e. splitting every cluster into two clusters, and is VQ2 is developed by Lipeika *et al.* (1995) based on splitting a cluster with largest average distortions into two clusters. Both the algorithms were evaluated and its results of speech reference pattern required and total error percentage has been tabulated in table 3 is presented. The standard Euclidean distance measure is used to find the distance between the test signal and reference templates of the speech signal. The codebook size is varied from 32 to 128 reference patterns and the results have been evaluated for both speaker dependent and independent. The comparison is drawn between table 2 and table 3.Table 2 shows the comparative results of speaker dependent and independent error percentage with the type of code book size suited to created the reference pattern for both algorithms i.e. VQ1 and VQ2[5].

6. RESULTS AND DISCUSSIONS:

By Table it is observed that for speaker independent application the recognition error is varied from 1.9% to 2.5% The table shows that the better results were obtained for vector quantization based on splitting a cluster with largest average distortions into two cluster. In this case only 32 feature vectors are needed to create a reference pattern template. But this method decreased the amount of reference pattern but increased speech recognition error rte.

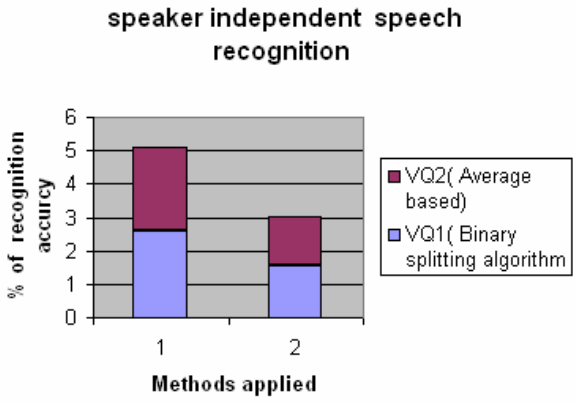
Table 2: Results for speaker independent speech recognition using Vector quantization

Speakers	Speaker depend. recogn.	Speaker independ recogn.	VQ1 32 ref. patt.	VQ1 64 ref. patt.	VQ1 128 ref. patt.	VQ2 32 ref. patt.	VQ2 64 ref. patt.	VQ2 128 ref. patt.
D0	0	0	2.8	1.8	20.3	2.8	4.6	5.5
D1	0	0.9	0.9	0.9	2.8	0.9	0.9	0.9
D2	0	3.7	2.8	0.9	1.8	1.8	1.8	1.8
D3	2.8	3.7	2.8	3.7	10.2	2.8	2.8	2.8
D4	0	3.7	5.5	8.3	20.3	6.5	4.6	5.5
D5	4.6	3.7	4.6	3.7	21.3	4.6	6.5	4.6
D6	0	0	2.8	1.8	4.6	0	2.8	1.8
D7	0	2.8	2.8	1.8	6.5	2.8	0.9	1.8
D8	0	0	0	0.9	2.8	0.9	0.9	0
D9	0.9	0.9	1.8	1.8	3.7	1.8	0.9	0.9
Total error %	0.83	1.94	2.68	2.59	9.44	2.5	2.68	2.59

From the proposed method the error recognition accuracy has been decreased from 0.83 to 0.14 for the speaker dependent application. And for speaker independent method. the error recognition accuracy has been decreased from 1.94 to 0.97 . Regarding the speech reference pattern vector code book of size 64 has be identified as the best reference pattern for the VQ1 clustering algorithm and for the second VQ2 clustering algorithm vector code book of size 128 has been identified as the best reference patterns for speech recognition application. Since we are interested in increasing the recognition accuracy more reference patterns are required as it is said in the literature. More the training set more the accuracy. Figure 2 shows the recognition accuracy plotted by the existing and proposed methods for speaker independent speech recognition application.

Table 3: Results for speaker dependent and independent speech recognition by the proposed method.

Utterances	Speaker dependent Recognition	Speaker independent Recognition	VQ1 32 ref. Pattern	VQ1 64 ref. Pattern	VQ1 128 ref. Pattern	VQ2 32 ref. Pattern	VQ2 64 ref. Pattern	VQ2 128 ref. Pattern
Ondu	0	0	2.1	1.3	17.2	1.7	3.2	0.9
Eradu	0	0	0.6	0.7	2.0	0.5	0.6	0.7
Muru	0	2.5	1.9	0.7	1.0	0.9	0.9	1.4
Nalakku	1.4	2.4	2.1	2.2	5.3	2.0	1.7	1.2
Idu	0	1.3	3.2	6.3	12.6	4.1	3.2	3.9
Aaru	0	2.4	3.1	1.8	18.5	2.4	3.1	2.4
Yelu	0	1.1	1.0	0.5	2.8	0.5	1.0	0.6
Enttu	0	0	1.9	0.7	2.1	1.3	0.6	0.9
Ombathu	0	0	1.7	0.5	2.3	1.0	0.7	1.6
Hathhu	0	0	1.1	0.6	2.1	1.0	0.7	0.9
Total error %	0.14	0.97	1.87	1.56	6.59	3.34	1.67	1.45



7. CONCLUSION

It is clear from the results, when new silence removal algorithm[5] is used the recognition error rate has been decreased from 2.59 to 1.56 in the case of VQ1 clustering algorithm and 2.5 to 1.45 for VQ2 algorithm. Also the speaker dependent error recognition rate has also been decreased from 2.5 to 1.45. This is because of the statistical measure used in the recognition system. This also reveals that statistical measures gives the finest distribution in analyzing the speech signal which is nonlinear and non stationary in nature. Using vector quantization for speaker independent mode best results was obtained when using VQ-2 for the codebook size chosen with 128 and VQ-1 for the codebook size of 64 patterns.

8. ACKNOWLEDGMENTS

The author would like to thank all the members who have supported in understanding this and implementing this paper.

9. REFERENCES:

[1] [1] Ladefoged, Peter. "Elements of acoustic phonetics". 2rd ed. The University of Chicago Press. Chicago. 1996.

[2] [2] Stranneby, Dag. "Digital Signal Processing: DSP and Applications". Oxford. 2001

[3] [3] Claudio, D.; Marins, "Calculo numerico computacional: teoriae pratica". 2rd ed. Editora Atlas. Sao Paulo. 1994.

[4] [4] M.A.Anusuya and S.K.Katti, "Discrete Wavelet transform for Noisy kannada speech recognition", International journal of computational Intelligence Research, Vol6, No.4, Published by Research India publication, India,2010.

[5] [5]. G. Saha, Sandipan et.al., " A New Silence Removal and Endpoint Detection Algorithm for Speech and Speaker Recognition Applications", Department of Electronics and Electrical Communication Engineering Indian Institute of Technology, Khragpur, Kharagpur-721 302, India.

[6] [6] E- Hocine Bourouba et.al. " Isolated Words Recognition System Based on Hybrid Approach DTW/GHMM" Department of electronic Faculty of Engineering, University of Annaba, Algeria Automatic and Signals Laboratory

[7] [7]. M.A.Anusuya and S.K.Katti, " Speech Recognition by Machine : A Review", International Journal of Computer Science and Information Security, Vol. 6, No. 3, 2009.

[8] [8]Rabiner L., B.-H. Juang, " Fundamentals of Speech Recognition". Prentice Hall,1993.

[9] [9]Vintsyuk T.K., "Speech Discrimination by Dynamic Programming". Kibernetika, 4(2), 81–88,1968.

[10] [10] Velichko V.M., N.G. Zagoruyko, "Automatic Recognition of 200 Words", Int. J. Man-Machine Studies, 2, 223, 1970.

[11] [11] Rabiner L.R. , "A Tutorial on hidden markov models and selected applications in speech recognition",Proc. IEEE, 77(2), 257–289,1989.

[12] [12]Rabiner L., B..H. Juang ,"Fundamentals of Speech Recognition. Prentice Hall, 1993.

[13] [13] Lipeika A., J. Lipeikien˙e, "Speaker identification using vector quantization". Informatica, 6(2), 167–180, 1995.

[14] [14]Lipeika A., J. Lipeikien˙e, "Speaker identification methods based on pseudostationary segments of voiced sounds". Informatica, 7(4), 469–484, 1996..

[15] [15] Jelinek F., "Statistical Methods to Speech Recognition", MIT Press, 1999.

[16] [16] Antanas Lipeik, et.al, " Development of isolted word speech recognition system", Informatica, Vol13, NO.1, pp.37-46, 2002.