Isolated Word Recognition for Marathi Language using VQ and HMM

Kayte Charansing Nathoosing

Department Of Computer Science, Indraraj College, Sillod. Dist. Aurangabad, 431112 (M.S.) India
charankayte@gmail.com

ABSTRACT

This paper describes the implementation of Marathi *Swar*, an experimental, speaker-dependent, real-time, isolated word recognizer for Marathi. The motivation and the advantages of choosing Marathi as the language for recognition are discussed here. The results obtained with Marathi Swar for tests conducted on a vocabulary of Marathi digits for 2 male and 2 Female speakers where presented in the end. The rest of the paper discussed the implementation of the system. The scheme proposed uses a standard implementation, with some modifications to the noise detection/elimination algorithm and the HMM training algorithm. The experimental results showed that the overall accuracy of the presented system was 94.63%.

INTRODUCTION

The Speech is the most prominent and natural form of communication between humans. Work on speech recognition is not new to our times. For many years people have been trying to make our machines hear, understand and also speak our natural language. This arduous task can be classified into three relatively smaller tasks:

1. Speech recognition to allow the machine to catch words phrases and sentences that we speak
2. Natural language processing to allow the machine to understand what we speak and
3. Speech synthesis to allow machines to speak.

The work described in this paper falls under the first category. Although a complete implementation of the first part would require that the computer be able to catch words and sentences from a continuous and natural speech over broad variations in accents, ages etc., to simplify the task, we focus on speaker dependent, isolated word recognition. The choice of Marathi language, on its part, has not been arbitrary. A major part of the motivation for choosing Marathi as the language for our recognition system comes from its local relevance. Whereas the English speaking community forms a very small percentage of India's population, Marathi being the national language of Maharashtra State is much more widely accepted. Marathi is an Indo-Aryan Language, spoken in western and central India. There are 90 million of fluent speakers all over world Marathi also offers several advantages as the language for speech recognition. Marathi language uses Devanagari, a character based script. A character represents one vowel and zero or more consonants. There are 12 vowels and 36 consonants present in Marathi languages (Bharti W, 2010). In other words, the Alphabet itself is the phoneme-set. Besides, the Marathi Alphabet is very well categorized on the basis of similarities in articulation methods of its letters. This property of Marathi makes it free of homonyms, thus obviating the complexity of handling them in design of speech recognition systems. The advantages are even more in the case of phoneme based recognizers, where a phonetic dictionary is not required for speech to text conversion.

Marathi Swar, is an experimental speaker-dependent, real-time, isolated word recognizer for Marathi. It uses a LPC (Linear Prediction Coding)VQ (Vector Quantization) front-end for processing speech signals along with HMMs (Hidden Markov Models) for recognition (Rabiner, 1983). The scheme proposed uses a standard implementation, with several modifications to the noise detection/elimination algorithm, and the HMM training algorithm, to achieve better results.

This paper is organized as follows. In Section II we give the implementation details for Marathi Swar. The various algorithms and techniques used, and the modifications suggested are discussed in this section. A discussion of the results obtained with Marathi Swar is presented in Section III. Section IV outlines the conclusions. Suggestions for improvement in the recognition accuracy of this experimental system are also presented in this section.

MATERIALS AND METHODS

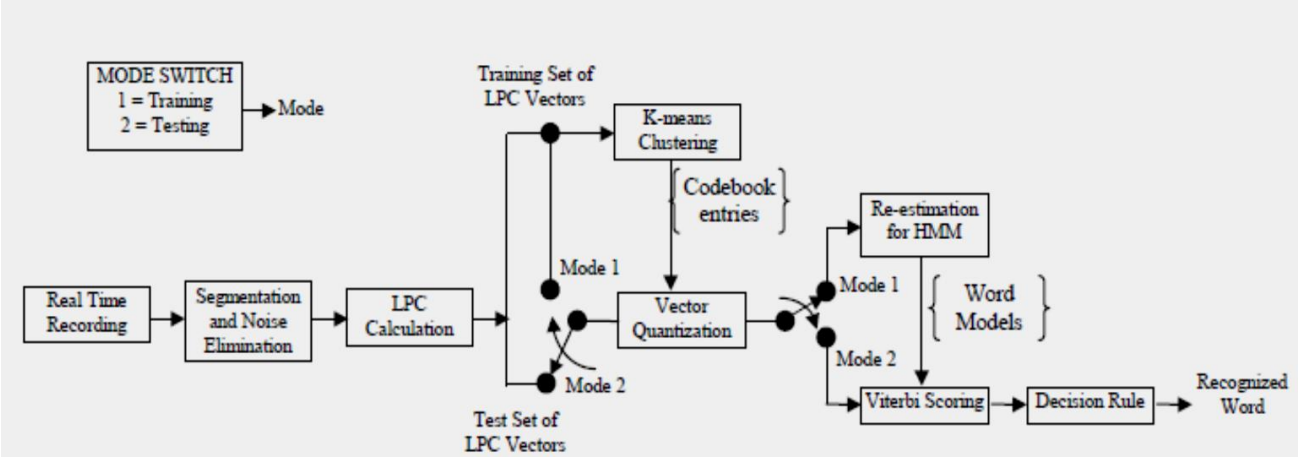

Fig. 1: Structural Design of Marathi Swar

A block diagram of our implementation is given in Figure 1. The whole system can be broadlycategorized into three main sections: Segmentation and Noise Elimination, Feature Extraction using LPC and VQ (Rabiner, 1993; Rabiner, 1983; Linde, 1980), and Recognition using HMM (Alan. Poritz, 1988). When VQ is running in mode 1, HMM part is non-existent. VQ training is implemented in offline mode. In this mode, the training set of LPC vectors is used by the K-means clustering algorithm (Rabiner, 1978) to iteratively update the vector quantizer codebook until the average distance falls below a preset threshold. The desired codebook size is obtained by using LBG algorithm (Rabiner, 1983). This codebook is used by the vector quantizer in the testing mode. In the HMM training mode, a set of LPC vectors (corresponding to an utterance of the word) is quantized by the vector quantizer to give a vector of codebook indices. A set of such vectors (corresponding to multiple utterances of the same word) is used to re-estimate the Hidden Markov Model for that word. This procedure is repeated for each word in the vocabulary. In the testing mode, the set of LPC vectors corresponding to the unknown word is quantized by the vector quantizer to give a vector of codebook indices. This is scored on each word HMM to give a probability score for each word model. The decision rule is used to choose the word whose model gives the highest probability.

A. Segmentation and Noise Elimination

Noise detection and elimination is critical to the design of a good real time dictation tool. In the absence of this, the recognizer is fed with several pseudo-words, which are actually segmented portions of the background noise. Noise detection can be done by defining certain parameters of speech, and a rule using these parameters to distinguish between noise and speech. Marathi Swar uses ZCR (Zero Crossing Rate) and short-time energy to distinguish between noise and speech. Speech is characterized by high ZCR values for unvoiced sections, or high short time energy for voiced sections. However, it was found that in noisy environments noise could have sufficiently high energy and ZCR values, making it virtually impossible to find thresholds such that all speech lies above the thresholds and all noise below it. After studying the noise characteristics in such noisy environments, we proposed a slight modification to the simple procedure. It is shown in Figure 2 (Mayukh Bhaowal, 2004). The thresholds here are determined dynamically. Glitches, another characteristic of noisy environments, have been handled by assuming a minimum sequence length for the utterances.

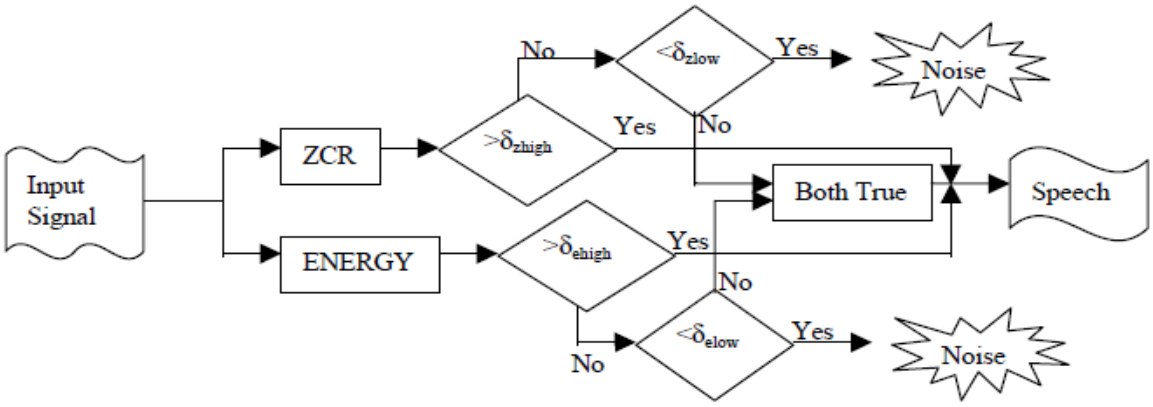

Figure 2 Noise/Speech detection algorithm

B. Feature Extraction

Linear Prediction Coding is used to obtain a feature vector from windowed input speech. The feature vector consists of cepstral coefficients up to order 10. The feature vectors are vector quantized. For this purpose, a codebook of size 128 is obtained using LBG algorithm. The splitting method uses standard deviation of the vector cluster, to determine the split vectors. The corresponding equations:

$X+ = X + \alpha*\sigma$

$X- = X - \alpha*\sigma$

Where:

X is the old centroid,

X^+ and X^- are the new centroids,

α is the coefficient of split, σ is the standard deviation for the vector cluster.

C. Recognition Using HMM

Marathi Swar uses a feed-forward model (or Bakis model) of HMM for recognition. The model, with 6 states, allows a single skip. The method of obtaining the general HMM out of the models of individual utterances can be critical to the recognition level and the speed of learning of the Model (Tarun Pruthi, 1993). The usual method involves taking the mean of the values for all models. This however creates problems when some particular elements in some models have extremely low values (nearing zero), pushed up to some bare minimum (some epsilon) by post estimation constraints. The presence of these epsilons while taking average can result in grossly erroneous values for the general model. This may lead to oscillations, and hence, a very large number of training utterances is required to emphasize the peaks in the symbol probabilities. One way to tackle the problem with fewer training utterances is to ignore all the epsilons while taking the average. This would then mean that, we keep track of the number of values used to calculate the average for each parameter of the model. This clearly is an overhead and hence is undesirable. In Marathi Swar, we employed a simple method of peak picking, i.e., taking the maximum of all values as the general value representing them. The total probability is then normalized to 1. The idea here is to approximate the previous solution, which was inefficient. An acceptable approximation is achieved because with sufficient training, the more appropriate values will dominate, and so the more eccentric ones will be smoothened by repetitive normalization across different training iterations, utterances and training sessions.

RESULTS AND DISCUSSION

The training set for the vector quantizer was obtained by recording utterances of a set of Marathi words encompassing the Marathi alphabet. The recordings were done for two male speakers. The recognition vocabulary consisted of Marathi Digits (0, pronounced as "shoonya" to 9, pronounced as "nau"). The HMM for each of the words was trained with 20 utterances of the word for the 2 male speakers. The results obtained are shown in Table 1 below.

Table 1: Recognition results obtained with Swar for a vocabulary of Marathi Digits

Word	Speaker1	Speaker1
0 ("shoonya")	88.23 %	94.44 %
1 ("ek")	71.35 %	76.74 %
2 ("do")	88.09 %	86.53 %
3 ("teen")	97.14 %	95.23 %
4 ("char")	89.47 %	84.13 %
5 ("paanch")	73.90 %	74.46 %
6 ("saaha")	82.60 %	85.67 %
7 ("saat")	91.42 %	84.00 %
8 ("aath")	81.08 %	83.78 %
9 ("nau")	81.63 %	77.77 %
Average	84.49 %	84.27 %

Much of the error in recognition can be attributed to the presence of plosives at the beginning or end of some of the words, as in, "paach"(begins and ends in a plosive), and, "ek"(ends in a plosive),which are misinterpreted as more than one word. Further, the Marathi digit vocabulary is a confusing vocabulary. The digits 4 (pronounced as "char"), 7(pronounced as "saat") and 8 (pronounced as "aath"), have the same vowel part, and differ only in their unvoiced beginnings and endings. Similarly, the digits 2 (pronounced as "do"), and 9(pronounced as "nau") have a very similar vowel part, and differ in their beginnings. These recognition errors were found to be more prominent in the case of Speaker 2, whereas the problem with plosives was much more prominent for Speaker 1. The digits 0 (pronounced as "shoonya") and 3 (pronounced as "teen") which are very distinct from the rest of the digits are seen to have a very high recognition rate for both the speakers.

An experimental isolated word recognition system for Marathi language was implemented. The results were found to be satisfactory for a vocabulary of Marathi digits. The accuracy of the real time system can be increased significantly by using an improved speech detection/noise elimination algorithm. Further improvement can be obtained by a better VQ codebook design, with the training set including utterances from a large number of speakers with variation in ages and accents.

LITERATURE CITED

Alan B Poritz, 1988, Hidden Markov Models: A Guided Tour. *Electrical and Electronics Engineers,* **1**:7-13

Bharti W Gawali, Santosh Gaikwad, Pravin Yannawar, Suresh C Mehrotra, 2010. Marathi Isolated Word Recognition System using MFCC and DTW Features*. Journal of Computer Applications,* **10**:3.

Linde Y, Buzo A and Gray RM, 1980. An algorithm for vector quantizer design*,* Institute *of Electrical and Electronics Engineers Transaction on Communication,* **28**(1):84.

Mayukh Bhaowal and Kunal Chawla, **2004 .** Isolated word recognition for English language using LPC, VQ and HMM*,* International Federation for information Processing 18th World Computer Congress - Student Forum Student Forum.343-352

Rabiner LR and Schafer RW, 1978. Digital Processing of Speech Signals, *Prentice-Hall, Englewood Cliffs, New Jersey*.

Rabiner LR and Juang BH, 1993. Fundamentals of Speech Recognition, *Prentice Hall, Englewood Cliffs*, *New Jersey*.

Rabiner LR, SE Levinson and MM Sondhi, 1983. On the application of Vector Quantization and Hidden Markov Models to speaker independent, isolated word recognition*, The Bell System Technical Journal.* **62**:4.

Tarun Pruthi, Sameer Saksena, Pradip K Das, 1993, Swaranjali: Isolated Word Recognition for Hindi Language using VQ and HMM. *Journal of Computing and Business Research,* **2**(2):3.