Recognition of vocal commands of a driver while driving in a noisy environment

Navid Samimi Behbahan , navid_samimi@yahoo.com

Sama Technical and Vocational Training College, Islamic Azad University, Behbahan Branch, Behbahan, Iran

Paper Reference Number: 5

Name of the Presenter: Navid Samimi Behbahan

Abstract

In this paper, we are going to investigate the topic of recognition of certain words in a noisy environment. This topic is very important due to loss of the driver's concentration while working with the audio system of a car. The words which was mentioned by the driver are: next (next track), were prior (prior track), increase (volume) and decrease (volume).

To exact the feature of each frame, we recruited Mel Frequency Cepstral Coefficient, First derivative, Second derivative with audio energy (in aggregate 39 features). The Hidden Morkov Model algorithm was recruits signify that the proposal algorithm set has a high level performance.

Key words: Hidden Markov Model, MFCC, Speech, noisy environment

1. Introduction

Speech recognition technology is a new method to identify vocal messages and commands and is one of the important parts of speech processing. In recent years, several researches have been done in this area on different languages and also Persian. Automatic speech recognition could be classified into three general categories of discrete words recognition, connected words recognition and continuous words recognition. Several techniques of speech recognition systems, including the numbers have been used, the most important of which are the Hidden Markov Model and Neural Networks. Substantial works have been done on discrete recognition of Persian words, some of which are stated hereafter. In the system implemented by Rostamzadeh et al. (1998), discrete word recognition was done based on speaker-independent recognition in which HMM was continuously applied. In this system, the model of words was established based on zero to nine and every word was modeled in six states. In training this system, 200 samples of each word expressed by equal number of men and women speakers were used. The applied coefficients were LPCC coefficients with 14 dimensions. Recognition error in this system with the change in number of Gaussian function in each state from 1 to 5 has varied from 2/15% to 0.25%. In another research done by Babaeizadeh et al. (1999), the combined model was implemented. The system was used on the numbers" zero" to "nine" and the words "Yes" and "No". 14 HMM models were generated and outputs of these models were given to

MLP network for further improvement. The system was tested on a database including 230 speakers, recognition rate of which was 98% in discrete HMM model and 97/9% for combined model. In another research done by Homayounpour et al. (1999) to recognize discrete digits through telephone, MLP system was applied in which the dimensional feature vector is estimated by neural network. Each word is divided into N parts and each part is estimated by a predictive network. Each word is divided into N syllables and each syllable is estimated by a predictive network. This system was tested on phone database including the numbers of zero to nine. Six MFCC coefficients were extracted from each frame. The results of recognition were 96% for trained data and 81% for experimental data. Another research was done by Sayadian et al. (2000), in which a single-state HMM model was applied. In this system, the number of Gaussian functions was 8. The system was compared with a continuous HMM system with five states in every model and 16 Gaussian functions. Considering MFCC coefficients, it's derivate and energy derivate, recognition rate of the implemented system was 100% and of continuous HMM was 94.17%.

In the aforementioned researches, numbers recognition was generally done on the discrete numbers.

In this research, the speaker-dependant discrete words recognition for the words expressed by a person (automobile diver) in the noisy and noiseless environments was emphasized. The voice recorded by a microphone was recognized and it could be transformed into commands recognizable for an electronic device or a computer. The usage area of this research is all the electrical and electronic devices and computers which receive commands from user in different ways.

2. Words and Data Range

The range of words is one of the important factors in determining good quality discrete speech recognition. The purpose of this research is to achieve recognition in the range of four words. The word range of this research includes the words "previous" (means previous track), "next" (means next track), "low" (means low voice) and "high" (means high voice). Every word was repeated twenty times, ten of which were in noiseless environment and ten others in noisy environment. Afterward the data were categorized into the trained and experimental data. They are used as the main data in this part of the research project of Persian speech processing.

3. Features Extraction

Proper feature extraction process is a basic and key step is solving any problem of pattern recognition. The feature applied in this part is Mel Frequency Capstral Coefficients (MFCC). The reason for selecting this feature is that filter bank is more resistant to noise. In order to extract MFCC parameters, speech signal is divided into 35ms frames with frame intervals of 10ms. After that, pre-stress process with α=0.975 is done on each frame signal, then Hamming window is applied. In filter bank analysis, 18 triangular filters distributed on Mel frequency spectrum are used and then Capstral coefficient is extracted (figure 1). Filter bank analysis process shows Mel scale for generating MFCC coefficients. The first and second derivatives of Capstral coefficients, also energy logarithm and the first and second derivatives of energy logarithm are added to Capstral coefficients. Capstral coefficients model static data of speech signal and are sensitive to its changes and dialect types, while Capstral coefficient derivatives include dynamic data and data on transfer between different dialects. The modulation of Capstral coefficients and its derivatives could express better features of speech signal. Finally, feature vectors with 39 coefficients were applied to model fames as it is explained in the next chapter.

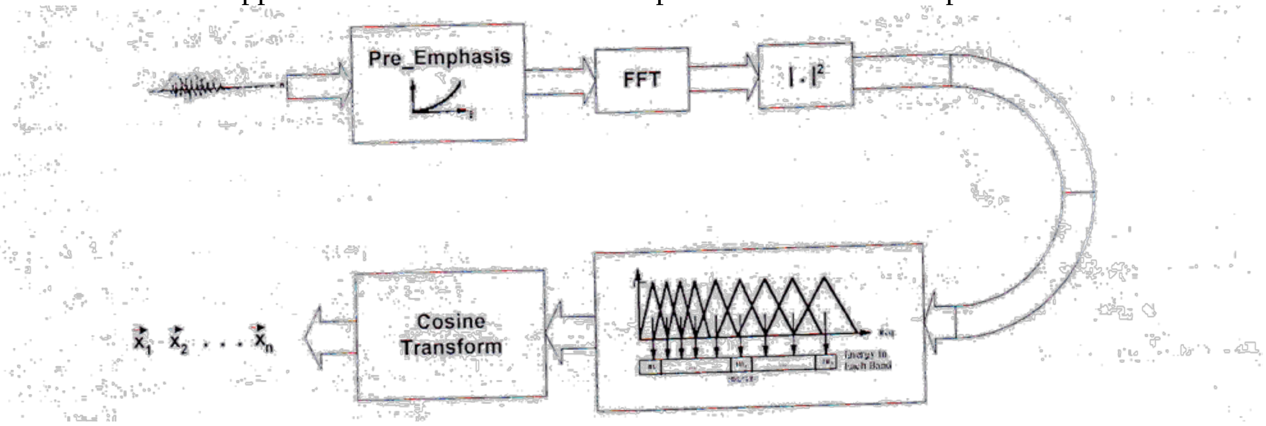

Fig 1: process of producing the MFCC coefficient

4. Hidden Markov Model

Today, Hidden Markov Model, as one of the successful methods, has so many applications in discrete and continuous speech recognition. Considering their high capability in modeling speech features, specially its dynamic features, these models have been greatly used and studied. Hidden Markov Model is a finite-state machine which encompasses two simultaneous random processes. A random process is sequence of the state the model takes and this sequence is hidden. Another random process is observations generation in each state. Depending on the manner of observation generation, there are two types of model: Discrete Hidden Markov Model in which observations are limited to a specific alphabet, therefore, feature vectors should be partitioned which would consequently lead to accuracy reduction. Another type is Continuous Hidden Markov Model in which observation vectors are generated using continuous probability density function. Therefore, there is no use of partitioning. In this research, Continuous Hidden Markov Model was applied. In the presented system, four models have been considered for four intended words in two noisy and noiseless environments and each word has been modeled with 6 modes. These HMMs have been considered from left-to-right without skip transition (Ahadi et al. 2000). Figure 2 shows the model topology.

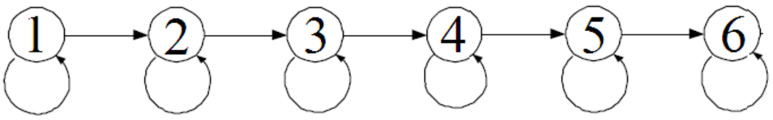

Fig 2: 6-state left-right HMM without skip transition

4.1. Training Models

To achieve an appropriate speech recognition system, it is essential for the models to be tested in an optimal way. One of the most applicable models in training HMM models is a method based on the maximum likelihood technique (ML), known as Baum-Welch. This algorithm is an iterative method and it is proved that it reaches a local maximum for likelihood. By using this algorithm and in the condition of multiple observation sequences, the reestimation relations for the main parameters in a CDHMM are as follows:

$$\hat{C}_{jk} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T}\gamma_t^{(l)}(j,k)}{\sum_{l=1}^{L}\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_t^{(l)}(j,m)} \tag{1}$$

$$\hat{\mu}_{jk} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T}\gamma_t^{(l)}(j,k).O_t^{(l)}}{\sum_{l=1}^{L}\sum_{t=1}^{T}\gamma_t^{(l)}(j,k)} \tag{2}$$

$$\hat{U}_{jk} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T}\gamma_t^{(l)}(j,k).(O_t^{(l)} - \mu_{jk})(O_t^{(l)} - \mu_{jk})^T}{\sum_{l=1}^{L}\sum_{t=1}^{T}\gamma_t^{(l)}(j,k)} \tag{3}$$

In these relations, $\hat{C}_{jk}$ is the mixture weight *k*th from *j*th state, $\hat{\mu}_{jk}$ is its average vector and $\hat{U}_{jk}$ is its covariance matrix and the relations for $\hat{\mu}_{jk}$ and $\hat{U}_{jk}$ are written in vector (or matrix) form. Also, the $\gamma_t^{(l)}(j,k)$ is the probability of being in *k*th mixture from *j*th state in *t* moment and observation O_t.

Before CDHMM models could be trained, it is necessary to initialize original models with proper values, so that achieving a local maximum would be treated as synonymous with reaching the general maximum. Therefore, the initialization of parameters is of significant importance. For this purpose, it is performed in two phases as follows:

A) All values of training observation sequences for each model are uniformly distributed among the numbers of states of that model and the values of variance and mean vectors are obtained for all the vectors resulted from each state, and then they would be used as the initial values for the parameters of that state.

B) Using the mentioned initial values, all observation sequences would be simulated with the state of the respective model using Vitrebi algorithm. Then the simulated values of above clause “A” would be used to obtain parameters of each state. This action continues by reaching a convergence criterion or the intended number of repetition.

Then, parameters obtained from the above procedures are applied as the initial parameters in Baum-Welch algorithm.

With regard to training models possessing Gaussian mixed observation density, although applying the mentioned model is allowed, it is preferable to increase the number of mixed elements in the following procedures. So, the initial stages of training are generally implemented with single-Gaussian observation density and then the algorithm such as mixture splitter model would be used.

Another important point in this regard, is the application of likelihood display in the logarithm form which is essential for the implementation of algorithms such as Baum-Welch, because the likelihood values, due to being multiplied by very small probable values, tend to zero after a few moments, which would make the training process impossible. So it is essential to use the likelihood and probabilities logarithm display during the implementation. In this case the multiply will change to sum and the problem of being tend to zero will be solved.

4.2. Discrete Word Recognition

The process of discrete word recognition by HMM is generally implemented by Vitrebi algorithm. This algorithm is designed based on the dynamic programming method and has several advantages:

- Finding optimal path for states based on Bellman optimality principle
- Seriously reducing calculations due to the application of dynamic programming principles
- Simultaneously finding observation sequence provided that ($P(O|\lambda)$) model is applied which is used in speech recognition

- Not-using sum (unlike Forward Algorithm) and instead using maximization which facilitate its implementation in logarithm form.

With regard to the aforementioned cases, Vitrebi logarithm is one of the proper and suitable methods for speech recognition, and so, it has been used in this part of the present research project.

5. Conclusion

A dataset including four words enounced by speaker has sampled in noisy and silent environment (each word, 20 times). MFCC coefficient of speech energy with first and second derivatives has considered as feature vectors of each frame. Two Hidden Markov models (noisy and silent) have created for each word. 60 percent of samples have applied to training set and remained 40 percent of these have applied to testing set. The set of proposed algorithm has implemented by MATLAB software, and correct recognition rate has obtained 88.4 and 96 percent for testing and training set, respectively.

References

Ahadi. S.M., Shieikhzadeh, H., & Homayonpour M.M., (2001), *National Research of Farsi Speech Processing, Progressing Report(No. 5),* Iran.

Babaeizadeh, S., Gholampour, E., & Nayebi, K., (1999), Improvement System Efficiency Recognition of Discrete Farsi Speech Using Combine of Neural Networks and Hidden Markov Models, *7th Power Engineering Conference*, Iran, pp. 183-190.

Homayonpour M.M., & Najjari, A., (1999), Recognition of Speaker-Independent Farsi Digits Using Neural Predicting Model, *7th Power Engineering Conference,* Iran, pp. 75-81.

Rostamzadeh, Sh., Ahadi. S.M., & Shieikhzadeh, H.,(1998), Recognition of Speaker-Independent Discrete Farsi Speech Using Hidden Markov Models with Continuous Density, *6th Power Engineering Conference*, Iran, pp. 93-97.

Sayadian, A., Badi’e, K., Hakak, M., & Beykzadeh, M.R., (2000), Presenting the Statistical Method FPG-GMM for Speech Recognition, *8th Power Engineering Conference,* Iran, pp. 398-406.