American Journal of Applied Sciences 10 (10): 1148-1153, 2013
ISSN: 1546-9239

doi:10.3844/ajassp.2013.1148.1153 Published Online 10 (10) 2013 (http://www.thescipub.com/ajas.toc)

A Hybrid Speech Recognition System with Hidden Markov Model and Radial Basis Function Neural Network

[1]Judith Justin and [2]Ila Vennila

[1]Department of Biomedical Instrumentation Engineering, Avinashilingam University, Coimbatore, India
[2]Department of Electrical and Electronics Engineering, PSG College of Technology, Coimbatore, India

Received 2013-07-17; Revised 2013-08-20; Accepted 2013-08-23

ABSTRACT

We analyze the performance of continuous speech recognition of a speaker independent system using Hidden Markov Model and Artificial Neural Network. Modern speech recognition systems use different combinations of the standard techniques over the basic approach to improve performance accuracy. One such combination which has gained more attention is the hybrid model. Our hybrid system for continuous speech recognition consists of a combination of Hidden Markov Model in the front end and the Neural Network with Radial basis function as the back end. The speech recognition process consists of the training phase and the recognition phase. The speech sentences are pre-processed and the features are extracted. The extracted feature vector is clustered into a model database by Hidden Markov Model and is trained by the Radial Basis Function Neural Network. During the recognition phase, the continuous sentence is pre-processed and its feature vector is modelled. This is compared with the database model which contains models stored during the training process. When a match occurs, the model is recognized and the recognition is made for the least error. From the recognized output the word error rate is computed, which is a measure of recognition performance of the hybrid model. The audio files of continuous sentences are taken from the TIMIT database. The performance of our hybrid HMM/RBFNN gives 65% recognition rate.

Keywords: Hidden Markov Model, Radial Basis Function Neural Network, TIMIT Database, Feature Extraction

1. INTRODUCTION

Automatic Speech Recognition has been the most investigated topic in Speech Processing. There are many aspects of speech recognition that are already well understood yet the human quality of speech recognition is still not achieved. There are many reasons why speech recognition is quite difficult. Natural speech is continuous; it does not have pauses between the words. Recognizing isolated speech (words demarcated by silence) is less difficult than recognizing continuous speech in which boundaries are not so apparent. Another difficulty faced is when the task is speaker independent. It is easy when the system is trained on one particular speaker and tested on the same speaker. But when the system is trained on many speakers and tested on a disjoint set of speakers, the performance degrades rapidly. In a large sized vocabulary system like the TIMIT speech corpus, recognizing words is much harder than recognizing 'digits'-zero to nine or isolated words, since there is a greater variability in the acoustics associated with each type of speech sound. The larger the task, the more it is confused with recognizing words. Recognition results widely report an accuracy of above 90% for isolated words and digits. Accuracy for speaker independent continuous speech recognition from large vocabulary sized speech corpus is poor. Hence a variety of techniques have been tried to improve the recognition accuracy of machine

Corresponding Author: Judith Justin, Department of Biomedical Instrumentation Engineering, Avinashilingam University, Coimbatore, India

recognition and newer areas of research explores speech recognition field in an attempt to conquer the practical problems faced.

General purpose speech recognition systems are based on Hidden Markov Models (HMM). The HMM gives a sequence of a real valued vector. The vector comprises of Cepstral coefficients obtained by taking the FFT of a short time window of speech and then taking the most significant coefficients. In each state it tends to have a statistical distribution that is a mixture of diagonal covariance Gaussians which will give likelihood for each observed vector. Each word or each phoneme will have an output distribution; a Hidden Markov Model for a sequence of words or phonemes is made by concatenating the individual trained HMM's for separate words or phonemes. A large vocabulary system gives a reduced accuracy -Word Error Rate.

Hybrid systems take advantage of both the systems (HMM and Neural Nets) thereby improving flexibility and recognition performance as reported by Tang (2009). Umarani *et al*. (2009) illustrated a hybrid system for continuous speech recognition, which consists of a combination of Hidden Markov Model and Radial basis function Neural Network. Our hybrid system uses speech features extracted, which takes inputs to HMM followed by training using Radial Basis Function Neural Networks (RBFNN).

1.1. Preprocessing and Feature Extraction

Speech signal carries information about the message to be conveyed, speaker identity and language information. For communication among human beings, there is no need for speech processing, since they are endowed with both speech production and perception mechanisms. But if a machine is placed in the communication chain, it needs speech processing because it does not have the knowledge of production and perception.

1.2. Pre-Processing

To extract the features from the speech signal, the signal must be pre-processed and divided into successive windows or analysis frames. So the following steps are performed before extracting the features (**Fig. 1**). The steps involved are pre-emphasis, frame blocking and windowing.

Higher frequencies of the speech signal are generally weak. As a result there may not be high frequency energy present to extract features at the upper end of the frequency range. Pre-emphasis is used to boost the energy of the high frequency signals. The output of the pre-emphasis is given by the following Equation 1 as follows:

$$\tilde{s}(n) = s(n) - \alpha s(n-1); \text{ where } \sigma = 0.95 \quad (1)$$

Speech is dynamic or time-varying. Speech analysis usually assumes that the signal properties change relatively slowly with time. This allows examination of a Short-time window of speech to extract parameters presumed to remain fixed for the duration of the window. The signal must be divided into successive windows or analysis frames. The next step in pre-processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and the end of the frame. The window must be selected to taper the signal to zero at the beginning and end of each frame. Hamming window is commonly used, which has the form Equation 2:

$$\omega(n) = 0.54 - 0.46\cos\frac{2\pi n}{N-1}, \quad 0 \le n \le N-1 \quad (2)$$

1.3. Feature Extraction

The features of a speech signal selected are the Short Time Average Zero Crossing Rate, Pitch Period Computation, Mel Frequency Cepstral Coefficients (MFCC), Formants and Modulation Index.

1.4. Short Time Average Zero Crossing Rate

A zero crossing occurs if successive samples have different algebraic signs. The rate at which zero crossing occurs is a measure of the frequency content of a signal. The average zero crossing rate gives a reliable means to estimate the frequency content of a speech signal. Rough estimates of spectral properties can be obtained using a representation based on short time average zero crossing rate. High and low frequencies imply high and low zero crossing rates respectively. The short time average zero crossing rate is computed using the Equation 3:

$$Zi = \frac{1}{N}\sum_{n=0}^{N-1}\frac{1}{2}\left|sgn\left(\tilde{x}i(n)\right) - sgn\left(\tilde{x}\left(n+1\right)\right)\right| \quad (3)$$

Where:

$$sgn\left(\tilde{x}i(n)\right) = 1; \text{ if } x(n) \ge 0$$
$$sgn\left(\tilde{x}i(n)\right) = 1; \text{ if } (n) < 0$$

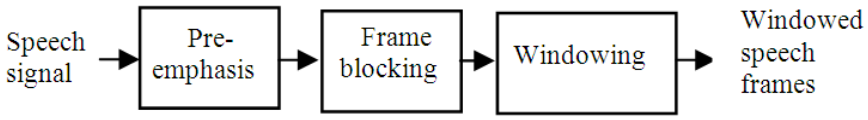

Fig. 1. Steps involved in pre-processing

1.5. Pitch Period Computation

The pitch is the fundamental frequency of the vocal cord vibration followed by 4-5 formants at higher frequencies. The typical values of pitch for male are 85- 155 Hz and for female 165-255 Hz. Pitch period is the reciprocal of the pitch. Average values for pitch period are around 8ms for male speakers and 4ms for female speakers. Pitch period can be calculated as follows: for each frame of the signal calculate the autocorrelation function and convert it to a binary signal. Set to logical 1 where the autocorrelation exceeds a pre-selected threshold and to logical 0 where the autocorrelation does not exceed the pre-selected threshold. We calculate autocorrelation function of the binary signal. Then detect peaks in the autocorrelation function of the binary signal. We use the distance between the peaks in the autocorrelation function of the binary signal as an estimate of the pitch.

1.6. Mel Frequency Cepstral Coefficients

Mel-Frequency Cestrum (MFC) is a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. It is a representation of the short-term power spectrum of a sound.

Mel Frequency Cepstral Coefficients (MFCC) can be extracted as in **Fig. 2**. The mapping of frequency in Mel scale is linear below 1000 Hz and logarithmic above 1000 Hz. So the band edges and center frequencies of the filters are linear for low frequency and logarithmically increase with increasing frequency. We call these filters as Mel scale filters and collectively a Mel scale filter bank. Here, the filters used are triangular and they are equally spaced along the Mel scale. Each short term Fourier transform magnitude co-efficient is multiplied by the corresponding filter gain and the results are accumulated. Then DCT is applied to the log of the Mel spectral coefficients to obtain the Mel frequency Cepstral coefficients. A common model for the relation between the frequencies in Mel and linear scales is as follows Equation 4:

$$\text{Mel frequency} = 2595 * \log(1 + \text{linear frequency} / 700) \quad (4)$$

The constants that define the filter-bank are the number of filters, the minimum and the maximum frequencies. These frequencies determine the frequency range of the filter-bank. Minimum frequency in speech is higher than 100 Hz and there is no speech information above 6800 Hz.

1.7. Formant Frequency

Formant features are adaptive, non-uniform samples of the signal spectrum that are located at the resonance frequencies of the vocal tract. They have higher signal-to-noise ratios than the other parts. The number and the position of these frequencies along the frequency axis differ depending on the phonemes and the position of the window along the phoneme. Along with the formants we could get the magnitude of the spectrum of that frequency to encode the properties of the speech and use it for speech recognition. The resonance frequencies of the vocal tract tube are called formant frequencies or simply formants. The formant frequencies depend upon the shape and dimensions of the vocal tract. Each shape is characterized by a set of formants. Different sounds are formed by varying the shape of the vocal tract. Formants in the human voice are essential components in the intelligibility of speech. The spectral peaks are known as formants and are numbered consecutively from low to high frequency.

1.8. Modulation Index

This ratio of frequency deviation to frequency of the modulating signal is useful because it also describes the ratio of amplitude to tone for the audio signal.

Modulation index = $\Delta f / f$m; where Δf = frequency deviation and fm = modulating frequency.

1.9. Training and Recognition Process

A speech sentence is represented by a series of feature vectors. A word will comprise of dozens of these vectors and in a sentence, it is required to classify a sequence of vectors. So, a model like HMM is required which is capable of dealing with variabilities.

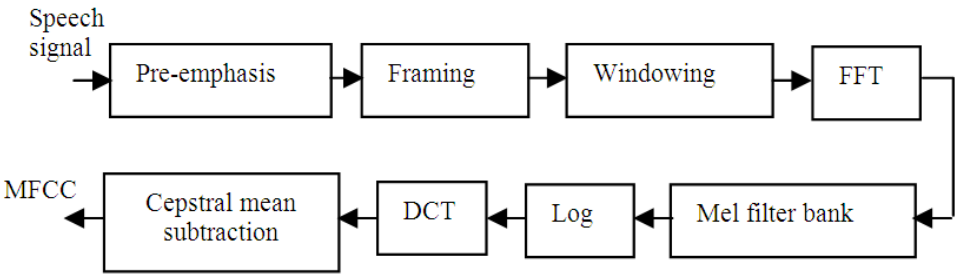

Fig. 2. Principle of extracting MFCC from a speech signal

1.10. Hidden Markov Model as Front End for Recognition Process

A Hidden Markov Model (HMM) is a statistical model in which the system being modelled is assumed to be a Markov process with unobserved state. The sole technique that gains the acceptance of the researchers to be the state of the art is a Hidden Markov Model (HMM) technique. It might be used successfully with other techniques to improve the performance, such as hybridizing the HMM with Artificial Neural Networks (ANN) algorithms. HMMs are popular because they can be trained automatically. They are simple and computationally feasible to use. In speech recognition, the hidden Markov model gives a sequence of n-dimensional real-valued vectors. These vectors consist of cepstral coefficients obtained by taking a Fourier transform of a short time window of speech and de-correlating the spectrum using a cosine transform. Then we consider the most significant coefficient. Rabiner (1989) has stated that, the hidden Markov model has a statistical distribution in each state which is a mixture of diagonal covariance Gaussians. This gives likelihood for each observed vector. Each word, or each phoneme, will have a different output distribution. A Hidden Markov Model for a sequence of words is made by concatenating the individual trained hidden Markov models for the separate words.

A large-vocabulary system requires context dependency for the phonemes (so phonemes with different left and right context have different realizations as HMM states); it uses Cepstral normalization to normalize for different speaker and recording conditions and for further speaker normalization Vocal Tract Length Normalization (VTLN) for male-female normalization and Maximum Likelihood Linear Regression (MLLR) for more general speaker adaptation is appropriate.

1.11. Radial Basis Function Neural Network (RBFNN) for Training

A RBFNN is an artificial neural network that uses radial basis functions as activation functions. It is a linear combination of radial basis functions. They are used in function approximation, time series prediction and control. It is mainly applied to problems of supervised learning. The main advantages of RBFNN over traditional Multilayer Perceptron (MLP) Neural Networks are its faster convergence, smaller extrapolation errors and higher reliability. Venkateswarlu *et al*. (2011a) have reported that performance of RBFNN is superior than MLP's. They have only one hidden layer and trains faster. The radial basis function is so named because the radius distance is the argument to the function. The hidden unit function used here is a Gaussian. Feature-response decreases monotonically with distance from a central point. Euclidean distance is computed from the test point being evaluated to the mean center of each neuron and a radial basis function is applied to the distance to compute the weight for each neuron. It gives better smoothing and interpolation properties. The feature vectors are clustered using the Kohonen Clustering algorithm. K-means clustering organizes the feature vectors into K number of groups. Grouping is done by minimizing the Euclidean distance between feature vector and corresponding cluster centroids. It takes a high-dimensional input and clusters it, but retaining some topological ordering of the output as in Cutajar *et al*. (2013). After training, an input will cause the output units in some area to become active. Such clustering (and dimensionality reduction) is very useful as a pre-processing stage. A typical RBF is given as Equation 5:

$$h(x) = \varphi\sigma(\|x - c\|) \quad (5)$$

Where:

φ = Gaussian activation function
σ = spread
x = input
c = mean centre

For a Gaussian function, φ (x) = e (-x)
Activation function of the i^{th} hidden unit for an input vector xj is characterized by the mean vectors and covariance matrices Equation 6:

$$gi(xj) = \exp\left[-\|xj - \mu i\|^2 / 2\sigma i^2\right] \quad (6)$$

where, xj = input vector, μi = mean vector (centers) and σi^2=variance.

Activation function is chosen such that it is neither too peaked nor too flat, $\sigma^2 = \eta x d^{2/2}$, where d = maximum distance between chosen cluster centers, η = empirical scale factor to control smoothness of the mapping function.

Hidden units are fully connected to output units through weights W_{ik}. The output units are linear and the response of K^{th} output unit for an input Xj given by, $Yk(Xj) = \sum_{i=0}^{Nu} Wik.gi.Xj$, where k =1, 2....Nc and g0(xj) = 1.

2. MATERIALS AND METHODS

The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. It also includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16 KHz speech waveform file for each utterance. These sentences are spoken by different speakers. Irrespective of the gender, speaker independent datasets were taken for the study. Speaker independent task of recognition is more difficult than the speaker dependent one. The speech sentence is pre-processed in order to extract features. The signal is pre-emphasized and is divided into successive windows or analysis frames as in Maheswari *et al*. (2010), so that the parameters can be calculated often enough to follow the relevant changes. Frame blocking splits the speech signal into 32 ms frames with each frame having 50% overlap with the adjacent frames. To minimize the signal discontinuities Hamming window is used. The features extracted are the Short Time Average Zero Crossing Rate, Pitch Period Computation, Mel Frequency Cepstral Coefficients (MFCC), Formants and Modulation Index. The feature vector is obtained for every 32 ms frame. HMM is mainly required to classify a sequence of vectors. The HMM model database is computed using HTK toolkit. The sequence of features for a sentence is obtained and a model database is created which comprises of feature vectors of 800 sentences. This is used for the training phase using Radial Basis Function Neural Network as it trains faster as reported by Venkateswarlu *et al*. (2011b). The performance is measured with Word Error Rate (E) which is the common metric to assess the speech recognition system. This gives the number of correctly recognized words. E = (S+I+D/N) X 100; where N = total number of words in the test set, S, I and D are the total number of substitutions, insertions and deletions.

3. RESULTS AND DISCUSSION

The performance of the hybrid model HMM/RBFNN was analyzed with continuous speech sentences taken from the TIMIT database. Two test data sets were chosen 400 sentences were taken as the test set for trial 1 and 300 were taken for trial 2. Features were extracted from all the sentences taken for the training, the same way as before for both the test sets and feature vectors were obtained. The model database contained feature vectors, which were 800 models. These feature vectors are matched with the model database of the test set. Minimum error identifies and recognizes each of the tested sentences. A computer with large data handling capacity was used.

The overall performance of the hybrid recognition system for a continuous speech speaker independent system was found to be 65%. But when it comes to recognition of speaker independent sentences, the recognition accuracy reported by researchers, so far, is less than 65% this work could be extended to a speaker dependent comparison which will show higher recognition accuracy for RBFNN with data taken from the same TIMIT speech corpus.

4. CONCLUSION

This study is done with a continuous sentence, speaker independent, large vocabulary speech recognition system using a hybrid model using HMM at the front end and the radial basis function neural networks as a classifier. Radial basis function neural networks with Gaussian activation functions are successfully implemented for pattern recognition tasks. The advantages of using a hybrid model is that a large vocabulary system can be handled comfortably and these are typically suited for continuous speech recognition systems. A standard speech corpus like the TIMIT can be

used for further research on other hybrid models and a comparison could be made.

5. REFERENCES

Cutajar, M., E. Gatt, I. Grech, O. Casha and J. Micallef, 2013. Comparative study of automatic speech recognition techniques. IET Signal Process. DOI: 10.1049/iet-spr.2012.0151

Maheswari, U.N., A.P. Kabilan and R. Venkatesh, 2010. A hybrid model of neural network approach for speaker independent word recognition. Int. J. Comput. Theory Eng., 2: 912-915.

Rabiner, L., 1989. A tutorial on hidden markov models and selected applications in speech recognition. Proc. IEEE, 77: 257-285. DOI:10.1109/5.18626

Tang, X., 2009. Hybrid hidden markov model and artificial neural network for automatic speech recognition. Proceedings of the Pacific-Asia Conference on Circuits, Communications and Systems, May 16-17, IEEE Xplore Press, Chengdu, pp: 682-685. DOI: 10.1109/PACCS.2009.138

Umarani, S.D., P. Raviram and R.S.D. Wahidabanu, 2009. Implementation of HMM and radial basis function for speech recognition. Proceedings of the International Conference on Intelligent Agent and Multi-Agent Systems, Jul. 22-24, IEEE Xplore Press, Chennai, pp: 1-4. DOI: 10.1109/IAMA.2009.5228022

Venkateswarlu, R.L.K., R.V. Kumari and G.V. Jayasri, 2011a. Speech recognition using radial basis function neural network. Proceedings of the 3rd International Conference on Electronics Computer Technology, Apr. 8-10, IEEE Xplore Press, Kanyakumari, pp: 441-445. DOI: 10.1109/ICECTECH.2011.5941788.

Venkateswarlu, R.L.K., R.V. Kumari and G.V. Jayasri, 2011b. Novel approach to speech recognition by using radial basis function neural networks. Int. J. Comput., 1: 181-187.