Available online at www.elixirpublishers.com (Elixir International Journal)

Advanced Engineering Informatics

Elixir Adv. Engg. Info. 60 (2013) 16250-16253

ISSN: 2229-712X

Speech Recognition for Large Vocabulary

Suma Swamy[1,*] and K.V Ramakrishnan[2]

[1]Department of Computer Science & Engineering, Sir M Visvesvaraya Institute of Technology, Bangalore.
[2]Department of Electronics & Communication Engineering, Anna University of Technology, Chennai.

ARTICLE INFO

Article history:
Received: 9 May 2013;
Received in revised form:
24 June 2013;
Accepted: 6 July 2013;

Keywords
ASR,
MFCC, DTW,
Spectral Subtraction,
Feature extraction,
Template matching.

ABSTRACT

This paper presents an approach to the development of a speaker independent, continuous word Speech Recognition System for a large vocabulary. The feature extraction is based on Mel-scaled Frequency Cepstral Coefficients (MFCC) and template matching employs Dynamic Time Warping (DTW). In general, efficiency of the speech recognition system in noise free environment is impressive. But, in the presence of environmental noise the efficiency of the speech recognition system deteriorates drastically. As an attempt to overcome this drawback, Spectral Subtraction (SS) is used for de-noising the speech signal before feature extraction and Convolutional Noise Removal is performed after feature extraction.

Introduction

With the development of Automatic Speech Recognition (ASR) technology, the robustness requirement of speech recognition system is becoming more and more important. One of the challenges of speech technology is to be able to provide robust and accurate recognition systems that can operate in a wide range of environments, including those having high levels of noise and vibration. Aircraft cockpit is an example of a demanding environment in which reliable Automatic Speech Recognition is becoming an important requirement [1].

Though the existing techniques are able to provide some compensation for adverse acoustic environments, they only provide a limited improvement and require excessive computational power. Most current speech recognition systems employ only one type of speech feature [1]. This may not result in a robust end product. So, the incorporation of different speech features into speech recognition potentially offers a significant degree of acoustic noise immunity at a reasonable computational cost.

BASIC Principles Of Speech Recognition

Speech is the vocalised form of human communication. It is based on the syntactic combination of lexical and names that are drawn from very large vocabularies. The smallest unit of spoken language is known as a Phoneme. The English language contains approximately 44 phonemes representing all the vowels and consonants that are used for speech Phonemes act as identifying markers since they remain at a constant value and can be broken down further. An algorithm has to be used to interpret the speech further.

A speech recognizer consists of a number of components. One of the important components is the Speech Corpus that consists of recordings of speech and their textual transcriptions. The Speech Recognizer learns to make correspondences between sounds and words.

This processes the signals recorded by the micro-phone into Feature Vectors that provide a snapshot of what is going on in the speech signal, emphasizing those features that are relevant to speech recognition. Typically, 100 feature vectors per second are produced [2].

Proposed Model

The Automatic Speech Recognition system involves two phases-(1) Training phase and (2) Testing Phase.

Training phase

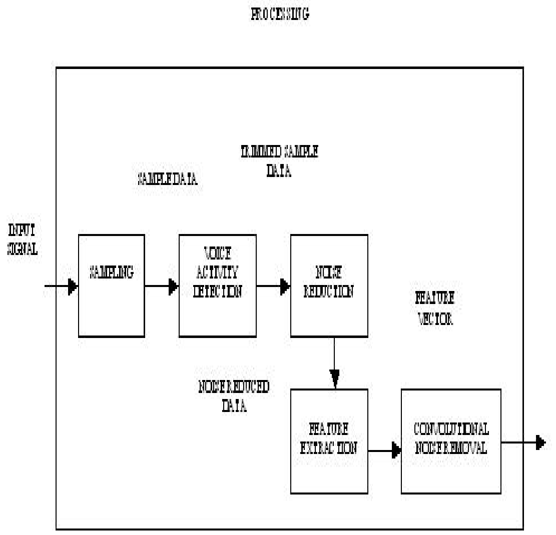


Fig 1: Flowchart for the processing phase, which is common to both training and testing phases.

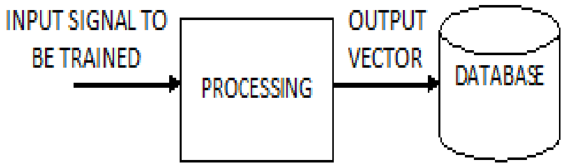


Fig 2: Flowchart for the training phase.

Testing phase

The input signal to be trained undergoes the following processes as shown in fig 1:

Tele:
E-mail addresses: suma_swamy@yahoo.com

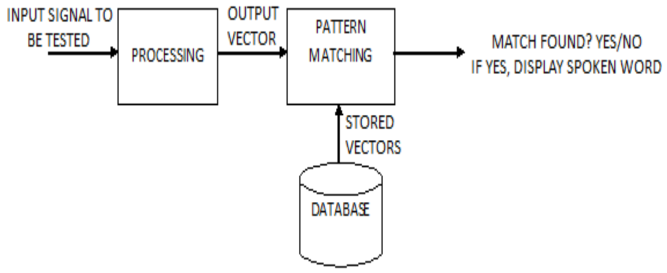


Fig 3: Flowchart for the testing phase.

1) Sampling: The sampling frequency of 8 KHz is sufficient for human speech. This frequency gives the window of 125μs between the two consecutive samples. Thus a sizable part of the processing can be done in real time. A higher frequency would shrink this window and would also demand greater memory and higher processing time. For a word of duration of 0.5 s, the number of samples at 8 KHz is 4000. With each sample stored as 16-bit value, it amounts to about 8 KB of storage space [3].

2) Voice Activity Detection

Voice Activity Detection (VAD) is done before spectral subtraction. It is also known as Speech Activity Detection or speech detection. It is a technique used to detect the presence or absence of human speech [2]. This is the first step towards noise reduction. This takes the audio sample as input and returns the sample with the non-speech sections trimmed off. This is done as follows:

Speech signal is segmented.

The zero-crossing rates for all segments in the speech signal are computed.

Frame energy for all segments in the speech signal is computed.

Frames with energy greater than ITU (Initial Upper threshold) and frames with energy lesser than ITL (Initial Lower threshold) are searched for.

The start and end indices for crossing rates higher than IZCT (Initial Zero Crossing Threshold) are searched for.

Thus the speech sample is trimmed using the start and end indices.

3) Noise reduction: Sample data is obtained as input for this process. The spectral subtraction algorithm carries out the noise reduction. The noise-reduced data is the output of this process.

Spectral subtraction is a traditional approach for reducing background noise in single channel systems [4]. And it is popular since it can suppress noise effectively, even in some real-life scenarios. Consider a speech signal s(k) corrupted by an additive background noise n(k). The observation signal y (k) can be expressed by

$$y(k) = s(k) + n(k) \qquad (1)$$

$$Y(\omega,r) = S(\omega,r) + N(\omega,r) \qquad (2)$$

where $Y(\omega,r)$, $S(\omega,r)$ and $N(\omega,r)$ denote the short-time Fourier transforms of y(k), s(k) and n(k) for frame r, respectively. Also, s(k) is assumed to be uncorrelated with n(k). If the noise spectrum $| N (\omega,r)|$ is estimated as $|N^1(\omega,r)|$, the estimation of the short-time speech spectrum $|S^1(\omega,r)|$ is represented by,

$$|S^1(\omega,r)|= H (\omega,r) \,|Y(\omega,r)| \qquad (3)$$

$$H (\omega,r)= \sqrt{(1-\alpha\, SNR_{post}(\omega,r)^2)} \quad \text{if } J>=0$$
$$H (\omega,r)= \sqrt{(\beta\, SNR_{post}(\omega,r)^2)} \quad \text{otherwise} \qquad (4)$$

$$J= (1/(\alpha+\beta))-SNR_{post}(\omega,r)^2 \qquad (5)$$

$$SNR_{post}(\omega,r)=| N^1(\omega,r)|/| Y(\omega,r)| \qquad (6)$$

where $H(\omega,r)$ is gain function, $\alpha(>=1)$ is the oversubtraction factor and β (>=0) is flooring level factor. When J>=0 spectral subtraction is carried out. On the other hand, spectral flooring is carried out when J<0.

Once the subtraction is calculated in the spectral domain with (3) and (4) the enhanced speech signal $S^1(k)$ is obtained as

$$S^1(k) = IFFT[|S^1(\omega,r)| \,.\, e^{j\, arg(Y(\omega,r))}] \qquad (7)$$

where the phase of the observation signal is used for the enhanced speech signal.

4) Feature Extraction: Features are extracted from the noise reduced sample data using the MFCC algorithm. Feature vector is obtained at the end of this phase.

The feature extraction involves identifying the formants in the speech, which represent the changes in the speaker's vocal tract. There are many approaches used viz. Linear Predictive Coding (LPC), Mel-scaled Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC), Reflection Coefficients (RCs). Among these, MFCC has been found to be more robust in the presence of background noise compared to other algorithms [5]. Also, it offers the best trade-offs between performance and size (memory) requirements.

The primary reason for effectiveness of MFCC is that, it models the non-linear auditory response of the human ear which resolves frequencies on a log scale [6]. The mapping from linear frequency to mel frequency is defined as,

$$Mel(f) =c.\, \log_{10}(1+f/700) \qquad (8)$$

To capture the auditory frequency content usefully, speech signal is best passed through a filter bank consisting of over-lapping triangular filters called the Mel Filter Bank. On the Mel scale, the centre frequencies of these filters are linearly spaced and the bandwidths are equal. The mel scale is often approximated by a linear scale for f<1 K*Hz* and logarithmic afterwards. The approximation of the Mel filter bank is as follows

$$U_m(k)= 1- (|k-c_m| / \Delta_m) \quad \text{if } |k-c_m| < \Delta_m$$
$$U_m(k)= 0 \quad \text{if } |k-c_m| >= \Delta_m$$
$$(9)$$

$$c_m= c_{m-1}+ \Delta_m \qquad (10)$$

$$\Delta_m= 4 \quad \text{if } f < 1 \text{ Khz}$$
$$\Delta_m= 1.2 * \Delta_{m-1} \quad \text{if } f>=1 \text{ Khz}$$
$$(11)$$

where k is the DFT domain index, $2\Delta_m$ is the bandwidth and c_m is the central frequency of the mth filter in the bank of size M. The mth energy co-efficient for nth frame of input signal X(k) is given by,

$$Y_n(m)= \sum_{k=cm-\Delta m \text{ to } cm+\Delta m} X_n(k) \,.\, U_m(k) \qquad (12)$$

The logarithm of the magnitude of each of these energy coefficients is taken to account for the logarithmic relation of intensity and loudness. These log energy coefficients so obtained are then orthogonalized by using inverse DCT (IDCT) [6]. The resulting parameters are called Mel-scaled Frequency Cepstral Coefficients (MFCC). Mathematically, this is as follows:

$$Y_n(j)= \sum_{m=1 \text{ to } M} \log |Y_n(m)| \cos((j(m-1/2)\pi)/L), \; j=0, 1, \ldots ,L \qquad (13)$$

16 filter banks (M = 16) are used. L = 15 which is the final number of co-efficients per frame of input signal.

5) Convolutional Noise Removal: Feature Vector is the input for this phase. It returns the Cepstral mean normalized version of the input matrix.

The output vector thus obtained is stored in the database as template shown in fig 2.

The input signal to be tested undergoes the following processes as shown in fig 3:

1) Processing: Processing involves phases such as sampling, noise reduction, feature extraction and convolutional noise removal as described earlier. After undergoing all these processes, an output vector is obtained from the given input vector.

2) Template Pattern Matching-DTW: After feature extraction of a spoken word, a frame-wise sequence of feature vectors is obtained. In this step, the output vector that is obtained from the processing phase is compared with each of those stored in the database. This is done using the DTW algorithm [7]. It is a technique, which "warps" the time axis to detect the best match between given two sequences.

For the spoken word S, let Sik denote the kth co-efficient of the ith frame. The DTW comparison of S and a template word T starts with calculation of a Local Distance Matrix of size 20x20 where each entry LD_{ij} is given by,

$$LD_{ij}= \sum_{k=1 \text{ to } 15} (S_i^k – T_j^k)^2 \qquad (14)$$

Thus the local distance LD_{ij} is the vector distance between the corresponding coefficients of ith frame of the spoken word and the jth frame of the template word.

A minimal warping path is found through the local distance matrix. The corresponding warping cost gives the DTW distance between the two sequences [8]. The minimum warping cost can be found efficiently using dynamic programming having the following recurrence:

$$D_{ij}=LD_{ij}+ \min \{ D_{i-1,j-1}, D_{i-1,j}, D_{i,j-1}\} \qquad (15)$$

where D_{ij} is the cumulative distance and $D_{20,20}$ gives the DTW distance.

Thus, DTW distance is calculated between the spoken word and each of the template words. The template word having the least distance is taken as a correct match, if its distance is smaller than a predetermined threshold value.

If the obtained output vector matches any of the stored vectors, then the spoken word is given as the output. If no match is found, then a suitable error message is displayed as the output.

Implementation

The designed Automatic Speech Recognition System uses the above-discussed algorithms mainly for two applications. They are: 1) Car application and 2) PC (Personal Computer) application. Two databases are maintained, one for car and another for the PC application. Each database consists of ten words.

Fig 4 shows snapshot of the home screen of the automatic speech recognition system. The user can choose one of the options from the screen. If the user needs to train the database, the first option is chosen, or if testing has to be done, the second option is chosen.

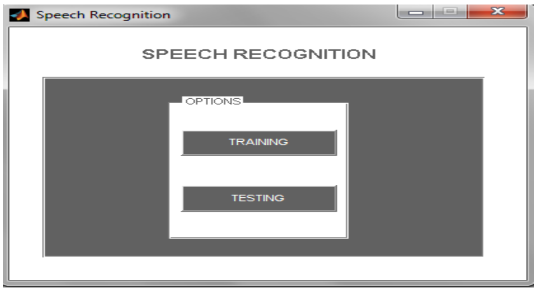
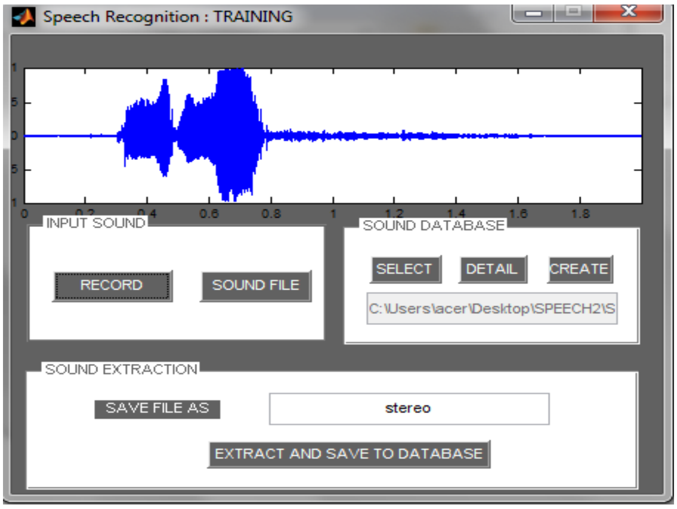


Fig 4: Snapshot of the home screen of the speech recognition system

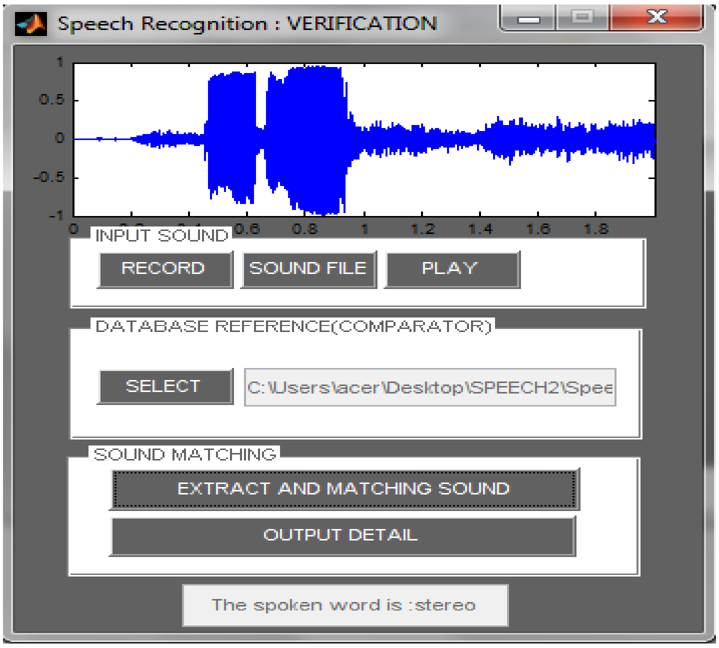


C:\Users\acer\Desktop\SPEECH2\S

stereo

Fig 5: Snapshot of the training phase

Fig 5 shows the snapshot of the training phase. In the frame called input sound, the user has two options to choose from Record button is used to record a sound dynamically. Sound file button is used to load an existing file. In the frame called Sound Database, there are three options to choose from. The select button is used to choose the database into which recorded sound or the loaded file has to be stored. The detail button is used to display the details of the chosen database. The create button is used to create a new database. In the frame called sound extraction, the save file as button is used to assign a name to the file that has the recorded sound. The extract and save to database button is used to extract the feature vectors of the recorded sound and store the file into the chosen database. The rectangular window on top shows the graph of the recorded sound.



C:\Users\acer\Desktop\SPEECH2\Spee

The spoken word is :stereo

Fig 6: Snapshot of the testing phase.

Fig 6 shows the snapshot of the testing phase. In the frame called input sound, the user has three options to choose from. Record button is used to record a sound dynamically. Sound file button is used to load an existing file. The play button is used to play the recorded sound or the chosen file. In the frame called database reference, the select option is used to select the database which has the set of words with which the recorded sound has to be compared with. In the frame called sound matching, the extract and matching sound button is used to extract the features of the recorded sound and compare it with the features of the words stored in the database and display the spoken word. The output detail button shows the details of comparison of the words in the database and the recorded sound. The text box at the bottom displays the spoken word if found, else, informs the user that the word could not be found.

Experimental Results

In the training phase, two male and two female speakers were asked to train the system by uttering each word from four different phonetically balanced wordlist and the files are stored as .wav file. During the testing phase they were asked to utter the same wordlist. The recognition result is summarised in Table 1. Both training and testing was done in noisy environment such as 56dB in a classroom.

Table I

Recognition Results for 50 phonetically balanced wordlist

Wordlist	Male1	Male2	Female1	Female2
List1	95%	97%	90%	92%
List2	98%	99%	95%	97%
List3	98%	98%	96%	96%
List4	100%	100%	98%	97%
Overall Efficiency	98%	99%	95%	96%

Thus the overall efficiency of the Speech Recognition System for large vocabulary is 97%.

Conclusion

In this paper MFCC and DTW techniques are used for developing a speaker independent Speech Recognition System for large vocabulary. The coding has been done using MATLAB. Spectral Subtraction is used to remove background noise to some extent. More efficient techniques can be used to further reduce the noise. Using an additional algorithm or set of algorithms along with the above-described algorithms in a speech recognition system may yield a better efficiency. Further work will be done using DHMM and Neural Networks for improving speech recognition efficiency.

Acknowledgement

The author acknowledges Visvesvaraya Institute of Technology, Belguam and Anna University, Chennai for the encouragement and permission to publish this paper. The author would like to thank the Principal of Sir M. Visvesvaraya Institute of Technology, Dr.M.S.Indira for her constant encouragement and would like to thank Prof.Dilip.K.Sen, Head of the Department, CSE/ISE for his invaluable guidance and suggestions from time to time

References

[1] Xiaoqiang Xiao, Jasha Droppo and Alex Acero, “Information Retrieval methods for automatic speech recognition”, in *Proc. of ICASSP, 2010.*

[2] www.wikipedia.org

[3] Sujay Phadke, Rhishikesh Limaye, Siddharth Verma,Kavitha Subramanian, “On design and implementation of an Embedded Automatic Speech Recognition Sytem”, Dept. of electrical engineering Indian Institute of Technology, Bombay.

[4] V.K.Gupta, Anirban Bhowmick, Mahesh Chandra, S.N.Sharan, “Speech Enhancement using MMSE estimation and Spectral Subtraction Methods”, Electronics and communication department, BIT, Mesra, Ranchi, India.

[5] H.Combrinck and E.Botha, “On the mel-scaled cepstrum”, department of electrical and electronic engineering, University of Pretoria.

[6] W.H.Press, B.P.Flannery, S.A.Teukolsky and W.T.Vetterling, *Numerical Recipes in C- The Art of Scientific Computing*, 2nd ed. Cambridge University Press, Feb 1993.

[7] M.Brown, L.Rabiner, “Dynamic time warping for isolated word recognition based on ordered graph searching techniques,” in *Intl. Conf. on Acoust., Speech,Signal Processing, ICASSP’82*, vol. 7, May 1982, pp. 1255-1258

[8] E. J. Keogh and M. J. Pazzani, “Derivative dynamic time warping,” department of Information and Computer Science, University of California, Irvine.

[9] Ahmad.A.M. Abushariah, Teddy.S.Gunawan, Othman.O. Khalifa,” English Digits Speech recognition system based on Hidden Markov Models”, Electrical and Computer Engineering Department, International Islamic University, University of Malaya, Kaula Lumpur, Malaysia.

[10] Punit Kumar Sharma, Dr.B.R.Lakshmikantha and K.Shanmukha Sundar, “Real time control of DC motor Drive using Speech Recognition”.

[11] Christophe Lévy, Georges Linarès, Pascal Nocera, “Comparison of Several Acoustic Modeling Techniques and Decoding Algorithms for Embedded Speech Recognition Systems”, France.

[12] Soon Suck, “HMM Voice Recognition Algorithm Coding”, University Dept. of Control & Instrumentation, Robotics Eng. Chosun University, South Korea, IEEE 2011.

[13] Zhongming Pan, “A Novel Speech Recognition Method for Student Management System”, School of Computer, Beijing University of Posts and Telecommunications, IEEE Proceedings of IC-NIDC2010

[14] Ibrahim Patel , Dr.Y.Srinivasa Rao, “Speech Recognition using Hidden Markov Model with MFCC-Subband Technique”, A.P India, 2010 International Conference on Recent Trends in Information, Telecommunication and Computing.

[15] Z.Hachkari, A. Farchi, Bmounir, J. EL Abbadi, “A Comparison of DHMM and DTW for Isolated Digits Recognition System of Arabic Language”, International Journal on Computer Science and Engineering (IJCSE) 2011, ISSN: 0975-3397.

[16] Raji Sukumar.A, Firoz Shah.A , Babu Anto.P, “ Isolated Question Words Recognition from Speech Queries by using Artificial Neural Networks”, Kannur University, Kerala, India, 2010 Second International Conference on Computing, Communication and Networking Technologies Computer Science and Engineering (IJCSE) 2011, ISSN: 0975-3397.

[17] www.mathworks.com