Indian Journal of Science and Technology, Vol 8(32), DOI: 10.17485/ijst/2015/v8i32/92130, November 2015
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Automatic Speech Recognition of Pathological Voice

Algabri Mohammed[1], Alsulaiman Mansour[1], Muhammad Ghulam[1], Zakariah Mohammed[1], Tamer A. Mesallam[2], Khalid H. Malki[2], Farahat Mohamed[2], M. A. Mekhtiche[1] and Bencherif Mohamed[1]

[1]College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia; malgabri@ksu.edu.sa, msuliman@ksu.edu.sa, Ghulam@ksu.edu.sa, mzakariah@ksu.edu.sa, mmekhtiche@ksu.edu.sa, mbencherif1@yahoo.com

[2]College of Medicine, King Saud University, Riyadh, Saudi Arabia; tmesallam@ksu.edu.sa, kalmalki@ksu.edu.sa, mfarahat@ksu.edu.sa

Abstract:

Background/Objectives: Automatic speech recognition (ASR) benefits human beings in many useful applications. Various ASR systems exhibiting good performance have been developed for normal speakers. The speech produced by a voice disordered patient is not like a normal speaker due to irregular vibration and incomplete closure of vocal fold. Therefore, an investigation is required by exploring the different speech features to develop an ASR system which can perform well for both pathological and normal speakers. **Methods:** In this paper, we proposed an automatic speech recognition system using Hidden Markov Model Toolkit (HTK) for normal and pathology voice. Four techniques are applied for feature extraction; Mel Frequency Cepstral Coefficient (MFCC), Perceptual Linear Prediction (PLP), RelAtiveSpecTrA - Perceptual Linear Predictive (RASTA-PLP), and linear prediction coefficients (LPC). The database that used to evaluate the performance of the developed system; includes a total of 297 speakers 121 of them were normal speakers and the remaining containing five types of vocal fold disorders. **Findings:** Experimental results show that the developed system gives good accuracies for normal and pathology voice. The highest accuracy of 94.44 % with a word error rate 5.55% is achieved in case of normal voice, and 88.63 % with a word error rate 11.63 % in case of pathology voice. Fuzzy logic controller is proposed to automatically segmentation the normal and disorders voice.

Keywords: Automatic Speech Recognition, Fuzzy Logic Control, HTK, Voice Pathology

1. Introduction

Speech is the fundamental medium by which the humans communicate with each other. This has attracted many researchers for the last 5 decades to do research on Automatic Speech Recognition (ASR). Speech is also an easy means to communicate with the computers compared to other communicating devices like mouse and keyboard. The communication with the computers could be accomplished by developing ASR system to identify the words uttered by the speaker on microphone or telephone and convert them into readable text. The basic definition of automatic speech recognition is an independent computer driven transcription of spoken language into readable text in real time[1]. ASR techniques has many goals to accomplish among them are it should be able to comfortably convert the oral speech into some readable written text which should be totally independent of the speaker, the environment where the speech is recorded like environment, and also it should not get effected by the device used to record the speech and finally size of the record and should be able to tolerate the noise and accent of the speaker. The recent development in ASR by the researchers has led the technology to recognize more than 90% of the speech accurately. Although lot of research has been done in the past but still we could not conclude that machine the convert 100% of the speech into readable text accurately in any acoustic environment or by any person. The first requirement of ASR was the disabled peoples who have the disability to communicate

*Author for correspondence

with the community around them. For instance people musculoskeletal disability can get assistance from this technology which is usually an outcome of multiple sclerosis, and cerebral palsy. The actual processing of the system starts when the speaker starts saying something some sentence. The developed system collects all the words and sentences in addition to the extra sounds from the environment and breaks in the speech and also the exact input spoken. The actual technical aspect of the software is to decode the whole speech into a full sentence. Initially all the speech signals are converted into sequence of vectors, and this process continuous till the whole sentence is delivered by the speech and vectors are created. The follows the syntactic decoder which actually converts and develops a valid sequence[2]. The main contribution of the paper is to evaluate the performance of the ASR system when using four feature extraction techniques on normal voice and pathology voice recognition. We also propose a method to classify the speech silence, voiced, and unvoiced speech based on fuzzy logic of short-time energy and zero-crossing. This paper will be structured as follows: Section 2 presents the literature review. Section 3 gives the details of speech recognition. Sections 4 give the experimental results. Section 5 show automatic speech segmentation using fuzzy logic. Section 6 concludes the paper and suggests some future works.

2. Literature Review

Authors presented in[3] an English speech recognition, with the ultimate goal to design and implement a system which would help recognizing English speech in digital format; the system is developed using MATLAB (GUI). The system was developed using Hidden Markov Model (HMM) which lead to the development of highly reliable system to recognize the speech. Mel Frequency Cepstral Coefficients (MFCC) technique was used to extract the features. The paper focused on English digits from Zero through Nine. Kuldeep kumar[1] presented an ASR for Hindi language. The system is developed using Hidden Markov Model Toolkit (HTK). Acoustic word model is used to recognize the isolated words. The system is trained for 30 Hindi words. Training data was collected from eight speakers. The overall accuracy of the presented system is 94%.

K. Kumar et al.[2] worked for connected-words speech recognition system for Hindi language. Hidden Markov model toolkit (HTK) was used to develop the system and the system was trained to recognize any sequence of 102 words. The experimental results showed that the presented system gave an accuracy of 87.01 %. Have used HTK technique to develop an automatic Arabic speech recognition system. The system has a unique feature as it can recognize both continuous speech and isolated words. The dataset used to develop this system was an Arabic dictionary which was built manually with as many as 13 speakers with a total of about 33 words in vocabulary. Tarun Pruthi et al.[4] in 2000, a new system was developed for Hindi language called as Hindi speech recognizer, it basically was developed for male speakers, the authors here have used LPC and HMM for both feature extraction and speaker recognition system respectively. The outcome of the experiment was satisfactory but it was limited to only speaker specific, since more performance was expected so the system needed further enhancement in the performance. For this reason Gupta gave a special Hindi word SR to the Hindi language in 2006. In the experiment also HMM was used in the continuous form as recognizer and the text used here was also in Hindi language but with ten Hindi digits. In[5], the authors have used Shannon Algorithm to do a comparative study for several different speech features depending on two classes one for relevant class and the other for irrelevant class. The following are the features compared by the authors wavelet feature, Mel scale FFT, Cepstrum, and mel-cepstrum. The whole experiment was done on TIMIT corpus and among all the features compared to identify the speech, high performance was achieved by Mel-Scaled FFT feature. M.A. Anusuya et al.[6] A unique speech recognition system was developed for the kannada language. The system works as follows; initially it calculates the discrete wavelet transform for the whole speech and then calculates the MFCC coefficients then follows the principal component analysis to recognize the speech. The authors also did a comparative study for several wavelet techniques with the combination of PCA inducted for the recognition purpose. Among many wavelet techniques the authors decided to apply the Daubechies 4, and 5-level decomposition and then Discrete Mayerwavel et al. with the expectation that these techniques could be compared to get more comparable results to check the performance of the system. In order to enhance the performance of the accuracy other techniques were introduced by replacing PCS with HMM in the system. By looking at the development of SR this system it can be estimated that in the future this system could be applied to other languages also. The system could be enhanced more one step ahead

by taking the following things into consideration like the no. of speakers and different speakers with different age groups and accent of speech.

3. Automatic Speech Recognition Using HTK

Figure 1 shows a diagram of the ASR system. In the following, we will describe the different steps in detail. The input of the system is a waveform speech file and the output is the recognized words.

3.1 Preparation Phase (Speech Corpus)

Here is the description about the collection of the dataset to conduct the experiments, the dataset is a collection of speeches recorded in different sessions like one from Communication and Swallowing Disorder Unit, and some of them collected from a very experienced phoneticians with excellent sound proof rooms with standardized recording protocols at King Abul Aziz University Hospital, Riyadh, Saudi Arabia. The whole experiments were conducted under the sponsorship of National Plans for Science and Technology (NPST), Saudi Arabia and the funds were issued to do the work was for two years which included many sub tasks among them was collection of dataset which was considered as a major task in the whole project. A comprehensive dataset was developed and designed with the help of a team to remove the entire shortcoming encountered by MEET database[7]. The main objective was to detect the person with vocal fold disorders to do this task this project collected all the records of speeches with normal person without defect and the people with some defect and the system would verify this if the speaker is defected or he/she is normal. The recording is a collection of different types of text like, three vowels with onset and offset information, different words isolating with each other like Arabic digits and common words

Figure 1. Automatic speech recognition system.

with continuous speech. To do the experiments the text with which covers all the Arabic phonemes were selected. The speaker has to record three utterances of each vowel /a/, /u/ and /i/ but continuous speech and words which are isolated are recorded only once to make it easy for the patient and prevent the patient from any hardship to record his/her speech. All the recording is done with the frequency of 50 KHz and the recorder used to record the speech is Kay Pentax computerized speech lab (CSL Model 4300). In this research, we used the Arabic digits and some common words to build ASR system. Table 1 show the statistics of AVPD database. It contains 121 normal speakers and 176 pathological speakers have five different types of vocal fold disorders. These voice disorders are: Sulcus, Cyst, Polyp, GERD, and Paralysis.

3.2 Training Phase

The training phase is used to estimate the parameters of a set of HMM models using training utterances. The speech files are labeled with silence before and after each digit. We built the system using word models; hence, in our system the grammar is defined in the grammar.txt file as shown in Figure 2. The words are the digits zero to nine plus three words.

3.2.1 Acoustic Model of Training Data

The goal of this phase is to extract useful information from voice for solving speech recognition problems. MFCC, LPC, PLP, and RASTA-PLP features were extracted from speech.

Mel Frequency Cepstral Coefficients (MFCCs)

It developed by Davis and Mermelstein, 1980. The aim here is to apply hamming window for each frame of signal then generate a cepstral feature vector for each frame. The next step is to apply Discrete Fourier Transform (DFT) of each frame. We then keep the log of amplitude spectrum. Then smooth the spectrum and using the Discrete Cosine Transform (DCT) to obtain cepstral features for each frame. The MFCC is the most widely

Table 1. AVPD database

	Number of speakers
Normal	121
Pathology	176
Total	297

```
/*
* Task grammar
*/
$word = sefr | wahid | ethnan | thalathah | arbaah | khamsah |
sitah | sabah | thamaniah | tessah| ashraa| gamal | zarf | gazal;
(sil <$word> sil)
```

Figure 2. Grammer file.

used for acoustic signal[8-10]. In this study MFCC with a dimension of 39 are used (12 MFCC and log energy with their corresponding delta (Δ) and acceleration (ΔΔ) coefficients).

Linear Prediction Filter Coefficients (LPC)

LPC is the one of the most speech analysis techniques. The Linear Prediction (LP) model is used to estimate the speech sample at current time from past speech sample using some weights. It is useful method for encoding good quality speech at a low bit rate[11].

Perceptual Linear Prediction (PLP)

PLP is a technique of analysis of speech presented and examined by Hermansky 1989. PLP is the alternative to the MFCC but it used equal loudness curve, critical bands, and intensity-loudness power law[12]. The PLP feature extraction technique is to describe the psychophysics of human hearing to an estimate of the auditory spectrum more closely[13].

Relative Spectral Transform Perceptual Linear Prediction (RASTA_PLP)

Is developed by Hermansky 1991. A band pass filter was added to the energy in each frequency in PLP in order to smooth noise variations[14,15].

3.2.2 HMM Definition and Learning

In this paper, Arabic digits (0-10) and three words have been modelled as HMM prototypes. Each initial prototype has the structure shown in. the tool HCompV in HTK is used to initialize, the mean and variance of each Gaussian component in the HMM definition, to the average and the overall variance of the training data of speech. After initializing the HMM models by HCompv, we re-estimate the models using HERest, these models are then re-estimated globally with the Baum-Welch algorithm.

3.3 Evaluation

Once the acoustic model of test data has been generated, the next step is to analyze the results. The *HResult* tool is provided for this purpose. *HResult* compares transcripts containing the resulting file with the original reference transcriptions. The percentage of accuracy (Acc.) and word error rate (WER) defined as:

$$Acc. = \frac{N - D - S - I}{N} \times 100 \quad (1)$$

$$WER = \frac{S + I + D}{N} \times 100 \quad (2)$$

Where N is a total number of labels, D is a deletion error, S is a substitution error, and I is insertion errors.

4. Experimental Results

In this study, we used a 5-fold cross validation. Table 2 shows the accuracy of four feature extraction techniques for speech recognition system with voice normal samples. The best accuracy with the proposed features can be found using MFCC and PLP approximate 94.44 %.

In case of voice pathology samples, table three show the average accuracy using four feature extraction techniques. The best accuracy 88.63 is for applying PLP feature with 39 coefficients. The accuracy of MFCC with 39 coefficients and RASTA-PLP with 13 coefficients almost the same. In addition, when applying LPC with 39 coefficients, the accuracy is 76.16 %, which is the worse results.

In order to test the performance of the system, we apply the four feature extractions for speech recognition system with normal and pathology samples as shown in table 4. The MFCC and PLP outperforms the other techniques in terms of Accuracy and Word Error Rate (WER).

The results of all speech recognition experiments are summarized in the figure 3. The three techniques MFCC, PLP, and RASTA-PLP achieved good accuracy comparing

Table 2. Experiments with voice normal samples

Normal		
Feature Type	**Accuracy %**	**WER %**
MFCC	94.44	5.55
PLP	94.44	5.55
RASTA-PLP	89.62	10.38
LPC	77.25	22.75

Table 3. Experiments with voice pathological samples

Pathology		
Feature Type	**Accuracy %**	**WER %**
MFCC	87.64	12.63
PLP	88.63	11.36
RASTA-PLP	87.14	12.85
LPC	76.16	23.84

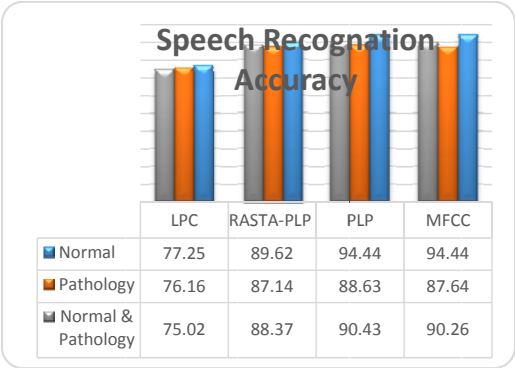

Figure 3. Summarize of all results.

with LPC. We can clearly see that, in case of pathology samples the recognition accuracy decreased comparing with normal samples.

5. Automatic Speech Segmentation

ASR techniques have many shortcomings and one among them is the segmentation which has to be done manually and the other shortcoming is to identify the where is the start and end of the voice sequence and also the speech is very sensitive with regards to the speed in which it is delivered. There is a dire need for a system which could do the segmentation automatically without the interference of the human skills. Partially the segmentation is done automatically with the combination of data from audio and visual format in the past works and with this temporal speech segmentation like AVSR systems were developed based on audio signals[16,17]. Automatic voice segmentation using fuzzy logic control was proposed. The voice signal was segmented into frames with duration 10 ms, then, hamming window was applied to prevent discontinuity. The mean of zero-crossing and short-term energy was computed for each frame and set as inputs of fuzzy logic

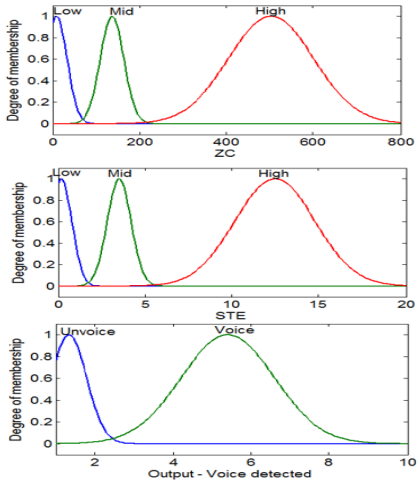

Figure 4. Fuzzy Logic controller for voice detection.

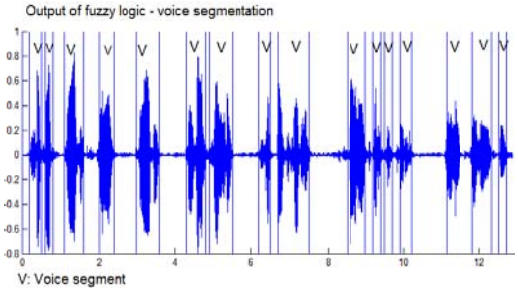

Figure 5. Voice pathology segmentation.

control. Voice and unvoiced detection is an output of fuzzy logic control. Figure 4 shows the membership functions of inputs and output of fuzzy logic controller.

After apply this fuzzy control on the voice pathological file contains the Arabic digits from zero to ten and three Arabic words. The output of the proposed method to detect voiced and unvoiced speech as shown in figure 5.

6. Conclusion

An automatic Arabic speech recognition system using HTK has been proposed. MFCC's, PLP, RASTA-PLP, and LPC are used for the feature extraction. Arabic digits (0 to 10) and three words (gamal, gazal, zarf) are used in the experiments for 297 speakers (121 normal and 179 pathology). The highest accuracy 94.44% is obtained of Arabic speech recognition by HTK using MFCC feature and PLP in case of normal voice. In case of pathological voice, we achieved the highest accuracy 88.63 % and word error rate 11.36 % using PLP features. In case of

all samples normal and pathology, we obtained 90.43 % highest accuracy. Moreover, process of automatic segmentation using fuzzy logic control is proposed. The proposed system is successfully tested for Arabic pathology speech. The future direction is seeking to apply this method to segment all database and compare it with manually segmentation. Genetic Algorithm and other optimization techniques can be used to improve the performance of fuzzy logic controller as a future work.

7. Acknowledgments

This project was funded by the National Plan for Science, Technology and Innovation (MAARIFAH), King Abdulaziz City for Science and Technology, Kingdom of Saudi Arabia, Award Number (12-MED2474-02).

8. References

1. Kumar K, Aggarwal R. Department of Computer Enginerring, National Institute of Technology, Kurukshetra: Hindi speech recognition system using HTK. Int J Comput Bus Res. 2011; 2. ISSN (Online).
2. Kumar K, Aggarwal R, Jain A. A Hindi speech recognition system for connected words using HTK. International Journal of Computational Systems Engineering. 2012; 1(1):25-32.
3. Al-Qatab BA, Ainon RN, editors. Arabic speech recognition using hidden Markov model toolkit (HTK). IEEE, 2010 International Symposium in Information Technology (ITSim). 2010.
4. Tripathy S, Baranwal N, Nandi GC, editors. A MFCC based Hindi speech recognition technique using HTK Toolkit. 2013 IEEE Second International Conference on Image Information Processing (ICIIP). IEEE, 2013.
5. Lee Y, Hwang K-W. Selecting good speech features for recognition. ETRI Journal. 1996; 18(1):29-40.
6. Anusuya M, Katti S. Comparison of different speech feature extraction techniques with and without wavelet transform to Kannada speech recognition. International Journal of Computer Applications. 2011; 26(4):19-24.
7. Saenz-Lechon N, Godino-Llorente JI, Osma-Ruiz V, Gomez-Vilda P. Methodological issues in the development of automatic systems for voice pathology detection. Biomedical Signal Processing and Control. 2006; 1(2):120-8.
8. Han W, Chan C-F, Choy C-S, Pun K-P, editors. An efficient MFCC extraction method in speech recognition. 2006 ISCAS 2006 Proceedings 2006 IEEE International Symposium on Circuits and Systems. IEEE, 2006.
9. Borde P, Varpe A, Manza R, Yannawar P. Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition. International Journal of Speech Technology. 2015; 18(2):167-75.
10. Hai NT, Van Thuyen N, Mai TT, Van Toi V, editors. MFCC-DTW Algorithm for Speech Recognition in an Intelligent Wheelchair. Springer: 5th International Conference on Biomedical Engineering in Vietnam. 2015.
11. Shrawankar U, Thakare VM. Techniques for feature extraction in speech recognition system: A comparative study. arXiv preprint arXiv:13051145. 2013.
12. Psutka JV. Comparison of MFCC and PLP parameterizations in the speaker independent continuous speech recognition task. In: 2001 E, editor. Eurospeech 2001; Scandinavia 2001. p. 1813-6.
13. Ramljak M, Stella M, Saric M, editors. Front-End Signal Processing for Speech Recognition. 1st International Conference on Wireless and Mobile Communication Systems (WMCS'13). 2013.
14. Saudi ASM, Youssif AA, Ghalwash AZ. Computer aided recognition of vocal folds disorders by means of RASTA-PLP. Computer and Information Science. 2012; 5(2):39.
15. Alsulaiman M, Muhammad G, Ali Z, editors. Classification of Vocal Fold Diseases Using RASTA-PLP. Proceeding of the 2013 International Conference on Bioinformatics and Computational Biology, (BIOCOMP'13). 2013.
16. Ma J, Cole R, Pellom B, Ward W, Wise B. Accurate automatic visible speech synthesis of arbitrary 3D models based on concatenation of diviseme motion capture data. Computer Animation and Virtual Worlds. 2004;15(5):485-500.
17. Musti U, Toutios A, Ouni S, Colotte V, Wrobel-Dautcourt B, Berger M-O, editors. Hmm-based automatic visual speech segmentation using facial data. Interspeech 2010. 2010.