Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 5– No. 2, May 2016 – www.caeaccess.org

Isolated Word Recognition for Dysarthric Patients

Sheena Christabel Pravin
Assistant Professor
Rajalakshmi Engineerig College
Chennai-602105

Abhiroop Chellu
Student
Rajalakshmi Engineering College,
Chennai-602105

P. Kannan
Student
Rajalakshmi Engineering College
Chennai-602105

ABSTRACT

In this paper, a HMM based speech recognition system is proposed for the Dysarthric patients. The speech samples recorded from patients with Spastic Dysarthria with mid to high intelligibility are taken from the UA Research database. The speech samples are de-noised using Discrete Wavelet Transform (DWT), after which the MFCC, LPC features are extracted. A comparative study on speech recognition using MFCC and LPC are presented. The recognition efficiency is found to be 68.50% using MFCC features and 66.54% using LPC features.

General Terms

Spastic Dysarthria, UA Research Database, Recognition

Keywords

DWT, MFCC, LPC

1. INTRODUCTION

Automatic Speech Recognition (ASR) can be defined as the independent, computer-driven transcription of spoken language into readable text in real time In a nutshell, ASR is a technology that allows the computer to identify words that a person speaks into a microphone or telephone and converts it into written text.

Having a machine to understand fluently spoken speech has driven speech research for more than 50 years. Although ASR technology has not grown to the point where machines understand all speech, in any acoustic environment, or by any person, it is growing on a day by day basis and is continually implemented in a number of applications and services. The ultimate goal of ASR research is to allow a computer to recognize real time speech, with 100% accuracy; all the words spoken intelligibly by any person, independent of vocabulary size, noise, speaker characteristics or accent. Today, if the system is trained to learn an individual speaker's voice, then much larger vocabularies are possible and accuracy can be greater than 90%. These systems are called as speaker dependent systems. In speaker independent systems, the recognizer is trained with speech samples from random speakers. These systems tend to have a slightly lower accuracy due to the randomness of the training samples used.

Speech disorders in patients are an outcome of genital disorders of the speech system. Dysarthria is a condition in which the muscles used for speech are weak and are difficult to control. It is often characterized by slurred or slow speech that can be difficult to understand. The type and severity of dysarthria depend on the area of the nervous system that is affected. Mostly, it is caused by damage to the brain. This may occur at birth, as in cerebral palsy or muscular dystrophy, or may occur later in life due to few other conditions that involve the nervous system, like stroke, brain injury, tumors, Parkinson's disease, Amyotrophic Lateral Sclerosis (ALS), Huntington's disease and multiple sclerosis. A person with dysarthria may demonstrate the following speech characteristics: "Slurred," "choppy" or "mumbled" speech that may be difficult to understand, Slow rate of speech, Rapid rate of speech with a "mumbling" quality, Limited tongue, lip, and jaw movements, abnormal pitch and rhythm when speaking, Changes in voice quality, such as hoarse or breathy voice or speech that sounds "nasal" or "stuffy". The neurological damage that causes dysarthria also usually affects other physical activities which can drastically restrict mobility and computer interaction. Studies have shown that severely dysarthric people are 150-300 times slower than typical users in keyboard interaction. However, since dysarthric speech is only 10-17 times slower than that of typical speakers, speech is a desirable input modality for computer-assisted interactions[1]. Hence, we propose a speaker-independent isolated word recognition system for the dysarthric, built using the Hidden Markov Model Toolkit.

In [2], Rabiner, et.al. reviews the theory of discrete Markov chains and the theory of hidden states, where the observation is a probabilistic function of the state, the three fundamental problems of HMM and techniques to solve them. He demonstrates various types of HMMs including ergodic as well as left right models and model features including the form of observation density functions, the state duration density, and optimization criterion for choosing optimal HMM values. He also discusses issues for implementation like topics of scaling, initial parameter estimation, model size, model form, missing data and multiple observation sequences. In [3], GarimaVyas, et.al. examine and present an approach to the recognition of speech signal using frequency spectral information with Mel frequency. It is a dominant feature for speech recognition. Mel-frequency Cepstral Coefficients (MFCCs) are the coefficients that collectively represent the short term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a non-linear Mel scale of frequency.

In [4], Montrti Karnjanadecha, et. al. describe speech signal modeling techniques which are well suited to high performance and robust isolated word recognition. The authors present new techniques for incorporating spectral/temporal information as a function of temporal position within each word. In particular, spectral/temporal parameters are computed using both variable length blocks with a variable spacing between blocks. Although these simple HMMs may be adequate for small vocabulary and similar limited complexity tasks, they do not perform well when used for more complex, and larger vocabulary tasks such as broadcast news transcription and dictation [5].HMM creates stochastic models from known utterances and compares the probability that the unknown utterance generated by each model. HMMs are a broad class of doubly stochastic models for non-stationary signals that can be inserted in to other stochastic models to incorporate information from several hierarchical knowledge sources [6].An HMM-based recognizer using variable duration of Hamming window is proposed in [7] that raises the recognition rate of dysarthric speech up to 80 %. Discrete articulatory feature recognition has been applied to identify values for concurrent features in [8]. Here, articulatory features are collected into different categories, each with a number of possible values. For example, a segment of speech can be concurrently voiced, nasal, and static, which represent values for three distinct features. Knowledge of common articulatory features (e.g.,nasality in /m/ and /n/) allows states in HMM models for different phones to be trained on shared data.

2. SPEECH CORPUS

One thousand seven hundred and eighty (1780) speech samples from thirteen Dysarthric Speakers chosen from the UA Research Speech database were used for recognition. The Universal Access (UA) Research database is the audio visual database of dysarthric speech for research promoting universal access to information technology. The Speech materials in this database consist of 765 isolated words per speaker: 300 distinct uncommon words and 3 repetitions of digits, computer commands, radio alphabet and common words. Data is recorded through an 8-microphone array and one digital video camera.

3. ISOLATED WORD RECOGNITION

Isolated Word Recognition was done using speaker dependent Word-Based HMMs. Using the HTK toolkit, speaker-dependent HMM speech recognizers were trained and tested. The optimum number of Gaussians in the mixture Gaussian was 3. A Hidden Markov Model (HMM) is a finite state machine which generates a sequence of discrete time observations. At each time unit (i.e., frame), the HMM changes states according to state transition probability distribution, and then generates an observation at time ‘t’ according to the output probability distribution of the current state. Hence, the HMM is a doubly stochastic random process model.

An N-state HMM is defined by state transition probability distribution A = {aij} Ni,j=1, output probability distribution B = {bj (o)}N j=1, and initial state probability distribution Π = {πi}N i=1. For convenience, the compact notation λ = (A, B, Π) is used to indicate the parameter set of the model. Every state of the HMM model could be reached from every other state of the model in a single step, and generally, the left-to-right HMMs are used to model speech parameter sequences since they can appropriately model signals whose properties change successively [9].

In the first experiment, the HMM models were trained using a 10- word vocabulary that included the 10 digits. Test data included two utterances of each digit.

A full probabilistic description of the above system would, in general, require specification of the current state (at time ‘t’), as well as all the predecessor states. For the special case of a discrete, first order, Markov chain, this probabilistic description is truncated to just the current and the predecessor state, (i.e.),

$$P\ [q_t] = Sj|q_{t-1} = Si,\ q_{t-2} = S_k\ldots]$$

$$= P\ [q_t = Sj|q_{t-1} = S_i] \quad \ldots \quad (3.1)$$

Furthermore, we only consider those processes in which the right-hand side of eqn. (3.1) is independent of time, thereby leading to the set of state transition probabilities aij of the form

$$a_{ij} = P\ [q_t] = Sj\ [q_{t-1} = Si].\ 1 \leq i, j \leq N \quad \ldots (3.2)$$

with the state transition coefficients having the properties

$$aij \geq 0 \quad \ldots (3.3)$$

$$\sum_{j=1}^{N} a_{ij} = 1 \quad \ldots (3.4)$$

The above stochastic process could be called an observable Markov model since the output of the process is the set of states at each instant of time, where each state corresponds to a physical (observable) event.

1. For each word v in the vocabulary, we must build HMM λ_n, i.e., we must estimate the model parameters ***(A, B,*** Π) that optimize the likelihood of the set observation vectors for the nth word.
2. For each unknown word which is to be recognized, the processing shown below must be carried out, namely measurement of the observation sequence 0 = {O_1, O_2……O_T}, via a feature analysis of the speech corresponding to the word; followed by calculation of model likelihoods for all possible models, P (O| λ_n), 1 ≤ n ≤ V; followed by selection of the word whose model likelihood is highest.

3.1 MFCC Features

Mel Frequency Cepstral Coefficients (MFCC) is one of the most commonly used feature extraction method in speech recognition. The technique is FFT based which means that the feature vectors are extracted from the frequency spectra of the windowed speech frames. The Mel frequency filter bank is a series of triangular bandpass filters.

Human hearing is not equally sensitive to all frequency bands. It is less sensitive at higher frequencies; roughly > 1000 Hz i.e. human perception of frequency is non-linear. Mel (melody) is a unit of pitch. Mel-frequency scale is approximately linear up to the frequency of 1KHz and then becomes close to logarithmic for higher frequencies. Human ear acts as filters that concentrate only on certain frequency components.

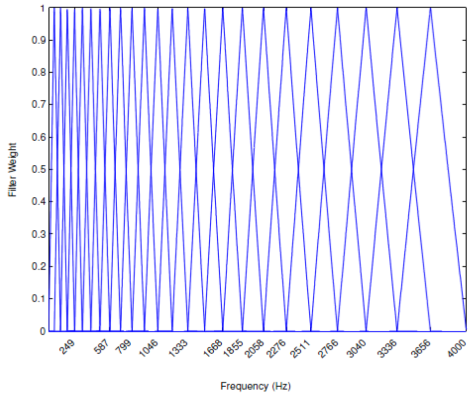


Fig. 3.1.1 Plot of pitch (Mel) versus frequency

Band-pass filters are non-uniformly spaced on the frequency scale, with more filters in the low frequency regions and less filters in the high frequency regions. Applying the bank of filters in Mel scale to the spectrum, each filter output is the sum of its filtered spectral components.

$$F(Mel) = [2595 * \log10[1+f] 700] \quad \ldots(3.1.1)$$

where 'f' denotes the real frequency, and Mel(f) denotes the perceived frequency. A new method for statistical estimation of Mel-frequency cepstral coefficients (MFCCs) in noisy speech signals is proposed in [9]. The cepstral domain face high complexity of distortion models caused by the nonlinear interaction of speech and noise in this domain [10, 11].

3.2 LPC Features

LPC (Linear Predictive coding) analyzes the speech signal by estimating the formants. In LPC system, each sample of the signal is expressed as a linear combination of the previous samples. It is a model based on human speech production. It utilizes a conventional source-filter model, in which the glottal, vocal tract, and lip radiation transfer functions are integrated into one all-pole filter that simulates acoustics of the vocal tract[12]. LPC is a frame based analysis of the speech signal which is performed to provide observation vectors of speech[13]. The de-noised signal is blocked into frames of N samples, with adjacent frames being separated by M samples. After frame blocking, the next step is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame; typical window being the Hamming window. Auto-correlation is performed on each frame of windowed signal

$$r_l(m) = \sum_{n=0}^{N-1-m} x_l(n) x_l(n+m) \quad \ldots(3.2.1)$$

$$m = 0,1, \ldots p$$

where the highest autocorrelation value, p, is the order of the LPC analysis. The next processing step being the LPC analysis converts each frame of (p +1) autocorrelations into LPC parameter set by using Durbin's method [14].

4. HTK TOOLKIT

The Hidden Markov Model Toolkit (HTK) is a portable toolkit for building and manipulating hidden Markov models. HTK is primarily used for speech recognition research although it has been used for numerous other applications.

4.1 Feature Extraction

The very first step in building a speech recognition system is "feature extraction". This process is achieved in HTK by using the HCopy function. The configuration file is a text file which contains information about the parameters of the extracted features. It must be created before using the HCopy function. After the prototype HMM is created, the initialization is carried out using the HInit function. After running this function, an updated HMM file gets saved in the output folder mentioned in the function. This updated HMM file will have the values of means and variances corresponding to the data in the extracted feature files. After running this function, an updated HMM file gets saved in the output folder mentioned in the function. This updated HMM file will have the values of means and variances corresponding to the data in the extracted feature files

4.2 Training Phase

The third step in the process is training the initialized HMM. In the training process, the initialized HMM will be trained to acclimatize more to the data in the feature files. This increases the overall accuracy of the system. For optimum accuracy to be achieved, the training has to be performed 6-10 times. The dictionary is a text file which contains the words that are to be trained. It can be created as a normal text file with the words listed in it. One more important step to be performed before training the data is the creation of label files. Each feature file will have a corresponding label file which contains the word that the feature denotes.

4.3 Testing Phase

After the training phase, speech samples are tested against the trained HMMs which are placed in a single folder. The function used for this process is HVite. The output of this step will be stored in a Master Label File (MLF) which will have details about each speech sample that has been used in the testing phase. The phonenet and DigitDict are essential files that are required during the testing phase. The DigitDict is a text file which contains the input and output as specified by the user. A phonegram is also a text file which contains all the words in the dictionary. Phonenet is a file that contains the system-friendly version of the phonegram. It is created using the function HParse.

5. RESULTS AND DISCUSSION

The overall and per-word accuracy can be checked using the HResults function. The output will be stored in a text file on the desktop window. In this function, the Master Label File that was created during the testing phase will be given as the input and compared to all the label files in the test data set. In this manner, the accuracy of the recognizer is easily calculated. The file that contains the recognition accuracy (in percentage) and results is stored as a text document.

Table 1. Comparison of HMM performance with MFCC and LPC for Dysarthric Speech recognition

Parameter type	HMM Recognition Accuracy
MFCC	68.50%
LPC	66.54%

MFCC features give slightly better recognition efficiency compared to LPC features. This is illustrated in the above tabulation.

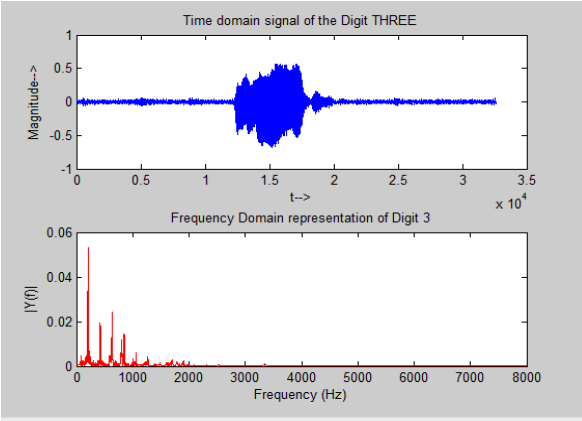


Fig 1: Time domain and Frequency domain representations of the digit "seven" spoken by a female with medium intelligibility

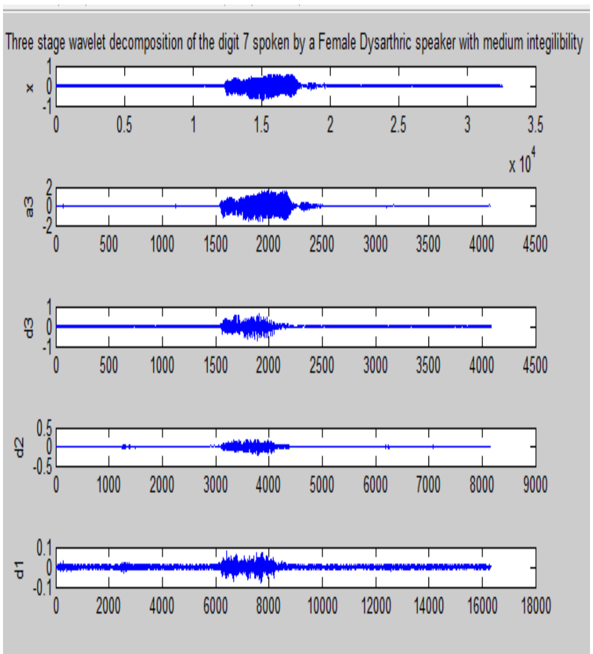
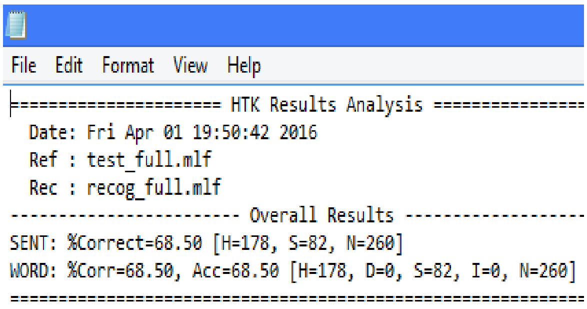
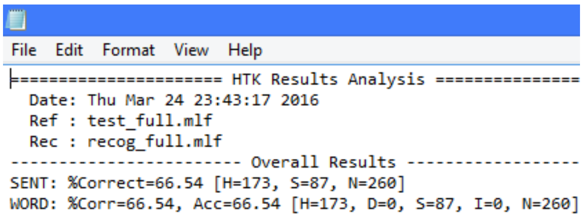


Fig 2: DWT coefficients of the digit "seven" spoken by a female with medium intelligibility



```
====================== HTK Results Analysis ===============
  Date: Fri Apr 01 19:50:42 2016
  Ref : test_full.mlf
  Rec : recog_full.mlf
------------------------ Overall Results ------------------
SENT: %Correct=68.50 [H=178, S=82, N=260]
WORD: %Corr=68.50, Acc=68.50 [H=178, D=0, S=82, I=0, N=260]
===========================================================
```

Fig 3: HMM recognition using MFCC



```
====================== HTK Results Analysis ===============
  Date: Thu Mar 24 23:43:17 2016
  Ref : test_full.mlf
  Rec : recog_full.mlf
------------------------ Overall Results ------------------
SENT: %Correct=66.54 [H=173, S=87, N=260]
WORD: %Corr=66.54, Acc=66.54 [H=173, D=0, S=87, I=0, N=260]
```

Fig 4: HMM recognition using LPC

6. ACKNOWLEDGMENT

We gratefully acknowledge the sincere efforts gone into the making of the UA-Speech database which contains Audiovisual isolated-word recordings of talkers with spastic dysarthria to promote the development of user interface for talkers with gross neuromotor disorders and spastic dysarthria [15].

7. REFERENCES

[1] Frank Rudzicz,” Adjusting dysarthric speech signals to be more intelligible”, Computer, Speech and Language, Vol.27, pp.1163-1177, Dec. 2012.

[2] Lawrence R. Rabiner, “A Tutorial On Hidden Markov Models and Selected applications in Speech recognition”, PROCEDINGS OF THE IEEE, VOL.77,NO.2,February, 1989.

[3] Garima Vyas, Barkha Kumari, “Speaker Recognition System based on MFCC and DCT”, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-5, June 2013.

[4] Montri Karnjanadecha and Stephen A. Zahorian,”Signal Modeling for High Performance Robust Isolated word”, IEEE Transactions on Speech and Audio Processing ,Vol.9 , Issue: 6, pp.647 – 654, Sept. 2001.

[5] Mark Gales and Steve Young ,”The Application of Hidden Markov Models in Speech Recognition”, , Foundations and TrendsR in Signal Processing Vol. 1, No. 3, pp. 195–304, 2007.

[6] D.B. Paul,”Speech Recognition Using Hidden Markov Models”, The Lincoln Laboratory Journal, Vol. 3, Number 1, 1990.

[7] Mohammed SidiYakcoub, Sid-Ahmed Selouani, Douglas O’Shaughnessy,” Speech Assistive Technology to Improve the Interaction of Dysarthric Speakers with Machines”, Published in:Communications, Control and Signal Processing, pp. 1150 – 1154, March, 2008.

[8] Frank Rudzicz, “Articulatory Knowledge in the Recognition of Dysarthric Speech”, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, vol. 19, no. 4, May 2011.

[9] Nhan Nguyen-Duc-Thanh, et. al. “Two stage Hidden Markov Model in Gesture Recognition for Human Robot interaction”, International Journal of Advanced Robotic Systems, July 2012.

[10] Kevin M. Indrebo, et.al., “Minimum Mean-Squared Error Estimation of Mel-Frequency Cepstral Coefficients Using a Novel Distortion Model”, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, Vol. 16, No. 8, Nov. 2008.

[11] Bharti W. Gawali, Santosh K. Gaikwad, Pravin Yannawar, and Suresh C. Mehrotra, "Marathi Isolated Word Recognition System using MFCC and DTW Features", ACEEE International Journal on Information Technology, Vol. 01, No. 01, pp.21-24, Mar 2011.

[12] Urmila Shrawankar, and Dr. Vilas Thakare, "Techniques for Feature Extraction in Speech Recognition System: A Comparative Study", International Journal of Computer Applications in Engineering, Technology and Sciences (IJCAETS), ISSN 0974-3596, pp 412-418, 2010.

[13] Pratik K. Kurzekar, et. al., “A Comparative Study of Feature Extraction Techniques for Speech Recognition System”, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 3, Issue 12, Dec.2014.

[14] Thanh Le, et. al., “The Effectiveness of the Durbin Selection Method For Variance Reduction”, Proceedings of: The Annual Meeting of the American Statistical Association, August, 2001.

[15] http://www.isle.illinois.edu/sst/data/UASpeech/