Learned Templates for Speech Recognition

Mrs. Madhura Phadke
Lecturer, DMCE, Airoli, Navi Mumbai
DMCE, sector 3, Airoli, Navi Mumbai.
91 - 9833307733
phadkemadhura06@gmail.com

Ms. Shubhangi Vaikole
PG student, MGMCOE, Panvel
DMCE, sector 3, Airoli, Navi Mumbai.
91 - 9324033401
shubhangiv@rediffmail.com

Mrs. Nusrat Parveen
Lecturer, DMCE, Airoli, Navi Mumbai
DMCE, sector 3, Airoli, Navi Mumbai.
91 - 9322988724
nusrat_athar@yahoo.co.in

ABSTRACT

The main focus in AI when it comes to sound-processing is to make a computer that can recognize what a person says to it. The aim is making a computer capable of automated speech recognition(ASR) which would be a next step in man-machine interface(MMI).Finding order in chaos without being overwhelmed is another major sub-division of artificial intelligence. Most phenomena in the world are analog--in other words, their actions are occurring continuously over time. A great example of analog devices that man has built is the tape recorder--it records every bit of sound every second, millisecond, nanosecond, etc .Computers, on the other hand, are digital machines in that it records data in absolute 0's and 1's. Thus, getting a computer to filter the analog world into meaningful digital representations is a major challenge that requires intelligence. One of the most effective ways to find meaning in data is to find the patterns that represent it. In this way, the information can be understood by the computer as well as be easier to store with less memory requirements. We are implementing a system for speech and speaker recognition which can be further used for various application domains as Health care, Military (High-performance fighter aircraft) , Training air traffic controllers etc. We are working to develop a system which will be capable of recognizing spoken words by various authorized users of the system.

Categories and Subject Descriptors

[Artificial Intelligence] Speech recognition is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. The recognized words can be the final results, as for applications such as commands & control, data entry, and document preparation. They can also serve as the input to further linguistic processing in order to achieve speech understanding .The implementation is based upon Hidden Markov model (HMM) and may use Dynamic time warping (DTW) technique for speech recognition.

General Terms

Pattern Recognition Algorithms, MATLAB, analog to digital conversion, defining vocabulary, authorized users of the system, Hidden Markov model (HMM), and Dynamic time warping (DTW) .

Keywords

Artificial Intelligence, Speech Recognition And Synthesis, Pattern Recognition, ASR, Phonemes, Acoustic Modeling, Language Modeling.

1 INTRODUCTION

The goal of this work is to understand the characteristics of speech produced by a person . Automatic speaker recognition [1] is an exciting technology that uses computers to automatically recognize a person based on his/her voice. The signal from a speaker's voice has information that can lead to the identity of the speaker. The system will be designed to recognize speech i.e. spoken words which can be connected speech or isolated words .Speaker recognition technology is currently being designed for gaining access to privileged databases such as personnel records, proprietary data, banking transactions over a telephone network, telephone shopping, voice mail, security control for confidential information areas, and remote access to computers. Applications involving fund transfers, entry to restricted premises, and telephone network business transactions would require authenticity of personal identity. To accommodate these applications, speaker recognition technology is divided into the two areas: speaker identification and speaker verification. Speaker identification systems identify a speaker from a group of speakers in a database, so there are several choices (alternatives) that can be made. Speaker verification systems can only make two choices: ``accept" the speaker (correct match), or ``reject" a speaker (speaker is an impostor.) Speaker recognition concentrates on choosing the proper speaker, therefore the researchers seek to find unique characteristics that will make the differences between speakers as large as possible. Speaker recognition differs from ``speech" recognition which seeks the proper message (words.) SPEECH recognition systems are designed to make voices sound as similar as possible so that the words can be determined regardless of the speaker. SPEAKER recognition systems enhance the differences in voices so that individual speakers can be identified or verified.

2 SPEECH PROCESSING ALGORITHMS

2.1 Hidden Markov model (HMM)

Modern general-purpose speech recognition systems are generally based on HMMs. These are statistical models which output a sequence of symbols or quantities. One possible reason why HMMs are used in speech recognition is that a speech signal could be viewed as a piece-wise stationary signal or a short-time stationary signal. That is, one could assume in a short-time in the range of 10 milliseconds, speech could be approximated as a stationary process. Speech could thus be thought as a Markov model for many stochastic processes (known as states).[2]

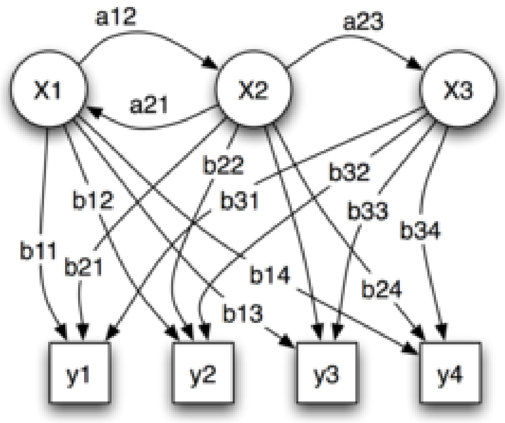

Figure 1. Sample Hidden Markov Model

Another reason why HMMs are popular is because they can be trained automatically and are simple and computationally feasible to use. In speech recognition, to give the very simplest set up possible, the hidden Markov model would output a sequence of n-dimensional real-valued vectors with n around, say, 13, outputting one of these every 10 milliseconds. The vectors, again in the very simplest case, would consist of cepstral coefficients, which are obtained by taking a Fourier transform of a short-time window of speech and decorrelating the spectrum using a cosine transform, then taking the first (most significant) coefficients. The hidden Markov model will tend to have, in each state, a statistical distribution called a mixture of diagonal covariance Gaussians which will give a likelihood for each observed vector. Each word, or (for more general speech recognition systems), each phoneme, will have a different output distribution; a hidden Markov model for a sequence of words or phonemes is made by concatenating the individual trained hidden Markov models for the separate words and phonemes[4].

Described above are the core elements of the most common, HMM-based approach to speech recognition. Modern speech recognition systems use various combinations of a number of standard techniques in order to improve results over the basic approach described above. A typical large-vocabulary system would need context dependency for the phones (so phones with different left and right context have different realizations as HMM states); it would use cepstral normalization to normalize for different speaker and recording conditions; for further speaker normalization it might use vocal tract length normalization (VTLN) for male-female normalization and maximum likelihood linear regression (MLLR) for more general speaker adaptation. The features would have so-called delta and delta-delta coefficients to capture speech dynamics and in addition might use heteroscedastic linear discriminant analysis (HLDA)[5]; or might skip the delta and delta-delta coefficients and use splicing and an LDA-based projection followed perhaps by heteroscedastic linear discriminant analysis or a global semitied covariance transform (also known as maximum likelihood linear transform, or MLLT). Many systems use so-called discriminative training techniques which dispense with a purely statistical approach to HMM parameter estimation and instead optimize some classification-related measure of the training data. Examples are maximum mutual information (MMI), minimum classification error (MCE) and minimum phone error (MPE)[3].

Decoding of the speech (the term for what happens when the system is presented with a new utterance and must compute the most likely source sentence) would probably use the Viterbi algorithm to find the best path, and here there is a choice between dynamically creating a combination hidden Markov model which includes both the acoustic and language model information, or combining it statically beforehand (the finite state transducer, or FST, approach).

2.2 Dynamic time warping (DTW)

Dynamic time warping is an approach that was historically used for speech recognition but has now largely been displaced by the more successful HMM-based approach. Dynamic time warping is an algorithm for measuring similarity between two sequences which may vary in time or speed. For instance, similarities in walking patterns would be detected, even if in one video the person was walking slowly and if in another they were walking more quickly, or even if there were accelerations and decelerations during the course of one observation. DTW has been applied to video, audio, and graphics. Indeed, any data which can be turned into a linear representation can be analyzed with DTW[6].

A well known application has been automatic speech recognition, to cope with different speaking speeds. In general, it is a method that allows a computer to find an optimal match between two given sequences (e.g. time series) with certain restrictions, i.e. the sequences are "warped" non-linearly to match each other. This sequence alignment method is often used in the context of hidden Markov models.

3 SPEECH PROCESSING

Speech processing extracts the desired information from a speech signal. To process a signal by a digital computer, the signal must be represented in digital form so that it can be used by a digital computer.

3.1 Speech Signal Acquisition

Initially, the acoustic sound pressure wave is transformed into a digital signal suitable for voice processing. A microphone or telephone handset can be used to convert the acoustic wave into an analog signal. This analog signal is conditioned with antialiasing filtering (and possibly additional filtering to compensate for any channel impairments). The antialiasing filter limits the bandwidth of the signal to approximately the Nyquist rate (half the sampling rate) before sampling. The conditioned analog signal is then sampled to form a digital signal by an analog-to digital (A/D) converter[7].

Speech recognition is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. The recognized words can be the final results, as for applications such as commands & control, data entry, and document preparation. They can also serve as the input to further linguistic processing in order to achieve speech understanding

The diagram below represents the sequence of operations required for speech recognition For speaker it can be speaker identification or speaker verification. This can be achieved by preparing templates for authorized users of the system. The speech (voice) captured by microphone is further converted into digital signals. This data consists of speech information along with some unwanted data known as noise. The required data is extracted and is further used with the pattern matching techniques to find out similar patterns and hence identify the speech with the help of stored templates. Speaker model templates are created and stored for speaker identification and verification.

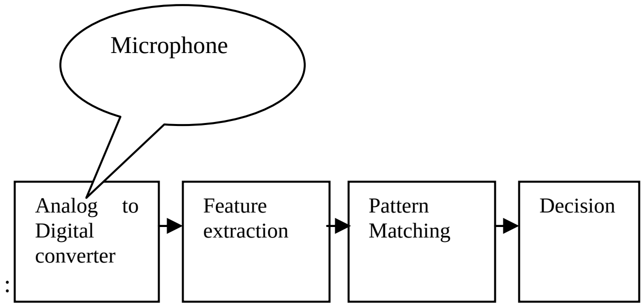

Figure 2. Speech processing

3.2 Feature Extraction

Feature extraction involves information retrieval from the audio signal. Extracted features are 13 MFCC (Mel-Frequency Cepstral Coefficients),[3] as well as a logarithmic energy measure. The speech signal is sampled at 8, 11 or 16 kHz and passed through first order offset compensation and pre-emphasis filters. The resulting signal is segmented into overlapping frames

3.3 Pattern Matching

Once the feature has been extracted , the task is to match the right pattern. Template matching is being used widely in recognition systems. Template Matching is one of the simplest methods to measure similarity. Initial samples are taken as reference (training sets). The test sample is compared with each of the training sets and the one with the best match is the one with the least Euclidean distance [8].

3.4 Decision

Either accept or reject the data based upon the result of pattern matching technique which compares the actual data with stored templates.

4 PROPOSED SYSTEM

We are planning to develop a system which will be useful for application areas with restricted vocabulary. We may use our system for operating a CNC machine. We will be using features of both the HMM and DTW techniques. The system will be armed with different learned templates for frequently required operations such as +,-, X, Y, Z etc. The application area is mainly targeted as mechanical production organizations where versatile types of CNC machines are working together for mass production.

5 EXPERIMENT DESIGN

Our work is based upon use of various advanced controls and functions for speech processing. We intent to design few learned templates for various authorized users of the system. Figure bellow represents speaker for the system with restricted vocabulary.

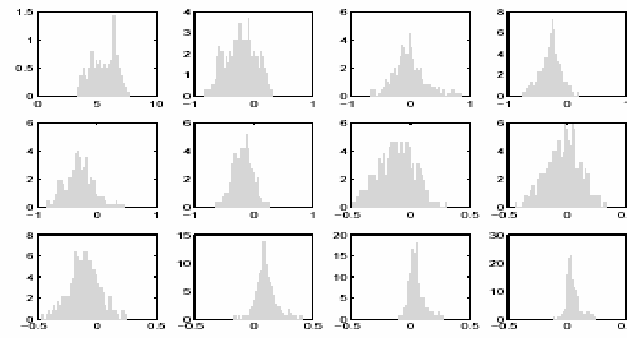

Figure 3. Features of speaker of the system

For each user of the system we prepare learned templates by taking into consideration the characteristics of speech. Accordingly we prepare models for users which are further used for speaker verification and identification.

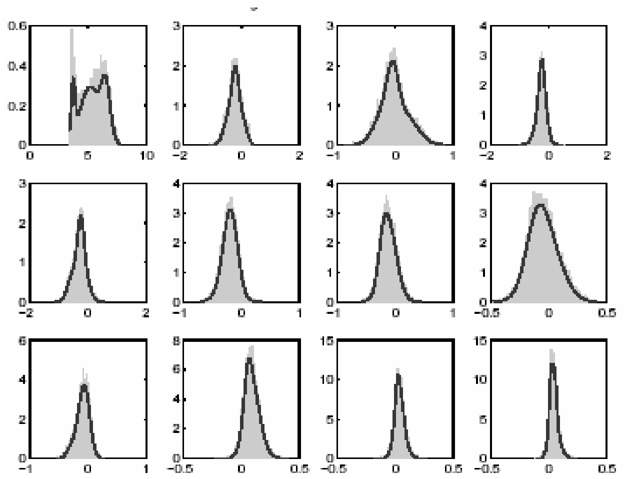

Figure 4. Learned templates for features of the speaker

With the pattern matching techniques we discover the nearly matching features and compute the likelihood. The next is testing of all newly discovered patterns that can be treated as templates for further processing. After the speaker identification phase the next task is to identify the speech. This will be tested by issuing various words as commands to a system.

6 CONCLUSIONS

In this work we have analyzed a method for speech and speaker identification. The developed system can be further used with variety of applications with ease of providing input by simply talking to a system.

7 ACKNOWLEDGMENTS

This work was partly supported by Datta Meghe Engineering College.

8 REFERENCES

[1] R. C. Rose and S. Partharathy, “A tutorial on ASR for wireless mobile devices”

[2] Handbook of Stochastic Analysis and Applications D. Kannan, V. Lakshmikantham

[3] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B.Dunn. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, 10(103), 2000.

[4] Che, C. and Lin, Q., "Speaker Recognition Using HMM with Experiments on the YOHO Database", Proceedings of Eurospeech, 1995.

[5] "Survey of the State of the Art in Human Language Technology (1997) by Ron Cole et all"

[6] Dennis van der Heijden. "Computer Chips to Enhance Speech Recognition", Axistive.com

[7] Jonathan Foote, An Overview of Audio Information Retrieval, ACM Multimedia Systems, Vol.7, 1999

[8] Rabiner and B.H. Juang, Fundamentals of speech recognition, Prentice Hall, Upper Saddle River, New Jersey 07458, 1993.