International Journal of Information Technology and Knowledge Management
July-December 2010, Volume 2, No. 2, pp. 447-450

ISOLATED WORD RECOGNITION SYSTEM FOR ENGLISH LANGUAGE

Sukhminder Singh Grewal[1] & Dinesh Kumar[2]

This paper proposes an optimum speaker independent isolated word recognizer for English language. The recognizer consists of following phases:-speech acquisition, extraction, classification and recognition phase respectively. Firstly the analog speech signal is converted into digital speech signal. From the digital signal the physical acoustic features are extracted and stored in the database. These features are used for the computation of the threshold values. Finally in the recognition phase the isolated word uttered by an independent speaker is recognized or rejected. The system has been trained and tested on words spoken by different speakers. The system in its extended form can be used for more words of common use speech, speaker identification, speech recognition, security …etc.

Keywords: Speech Recognition, Isolated Word Recognition, Acoustics

1. Introduction

Humans interact with their environment in many ways and receive information through many modalities: sight, audio, smell and touch. To communicate with the environment humans send out signals or information visually, auditory and through gestures. With the development of technology the human dependency on the machines has increased manifold. Humans have to interact with the machines to get the data processed. Human –computer interaction often uses a mouse, keyboard …etc as machine input and screen, printer, speaker...etc as output.

The reasons for opting the project were the following: Keyboard a popular medium requires a certain amount of skill for effective usage. A mouse requires good hand- eye coordination. Physically challenged people find computer difficult to use. Partially blind people find reading from monitor difficult. Moreover current computer interface assumes a certain level of literacy from the user. It expects the user to have certain level of proficiency in English. Speech interface can help us tackle these problems. Speech synthesis and speech recognition together form a speech interface. Speech synthesizer converts text into speech. Speech recognition in a computer domain may be defined as the ability of computer systems to accept spoken words in an audio format such as wav, raw recognize the audio format and take appropriate action of operation.

Speech is a complicated biometric signal produces as a result of several transformations occurring at semantic, linguistic, acoustic and articulatory level. Speaker related differences are result of a combination of anatomical differences inherent in the vocal tract and the learned speaking habits of different individual. The task of recognition is not easy one the variation in dialect, speaking rate, vocal tract length that exist between speakers account for many of the difficulties encountered during recognition.[10]

A Model for Computer Perception of Speech:

The speech signal $x_1(t)$ is generated by a discrete and finite sequence of actions:

$$A = a_1(t_1)\, a_2(t_2)\, a_3(t_3)....a_k(t_k)... a_K(t_K)$$

Where $a_k(t_k)$ denotes an action ending at time t_k; $a_1(t_1)$ represents the silence preceding the beginning of a sentence. When a person reads a sentence S, a relation

$$R_1(S, A)$$

is applied which produces A. The relation R depends on the speaker, his/her mood, state of health. [11]

In a language, elementary sounds that distinguish meaning are called phonemes. For each Speech Recognition, a language is built upon a set of phonemes and several other types of sound (noise...etc).for example a lexicon, the pronunciation rules, acoustic library of language.[12]

The speech recognition refers to the ability to listen to spoken words and identify various sounds in it and recognize them as a word.

In the computer domain the speech recognition is the ability of the computer to accept an input speech data and recognize the word.

Speech recognition involves pre-processing of acoustic signal so that meaningful parameters can be extracted. These parameters are used for recognizing the word.

[1]Ludhiana College of Engineering and Technology, Ludhiana, Punjab

[2]DAV Institute of Engineering & Technology, Jalandhar, Punjab

Email: [1]Sukhminder123@sifymail.com, [2]dinesh_daviet@hotmail.com

Isolated Word Recognition

Let each spoken word be represented by a sequence of speech vectors O defined as:

$$O = o_1 o_2 o_3 \ldots o_t$$

Where o_t is the speech vector at time t. The isolated word recognition can be regarded as that of computing

$$\text{Arg max } \{p (w_i|O)\}$$

Where w_i is the i^{th} vocabulary word and $p (w_i)$ is a given set of prior probabilities for a given set of probabilities. The most probable spoken word depends only on the likelihood $p (O|w_i)$. [13]

Recognition involves mapping the given input in the form of features to one or the known sounds. This may involve use of various knowledge models for the precise identification and ambiguity removal. Knowledge models refer to the model such as phone acoustic model, language model...etc. which help in recognition system. To generate the knowledge model one needs to train the system. During the training period one needs to show the systems a set of inputs and what outputs they should map to.

The general approach to speech recognition consists of five steps: digital speech data acquisition, feature extraction, pattern matching, and decision making to accept or reject the request. The fiq1. Describes the flow diagram of the isolated word recognizer system

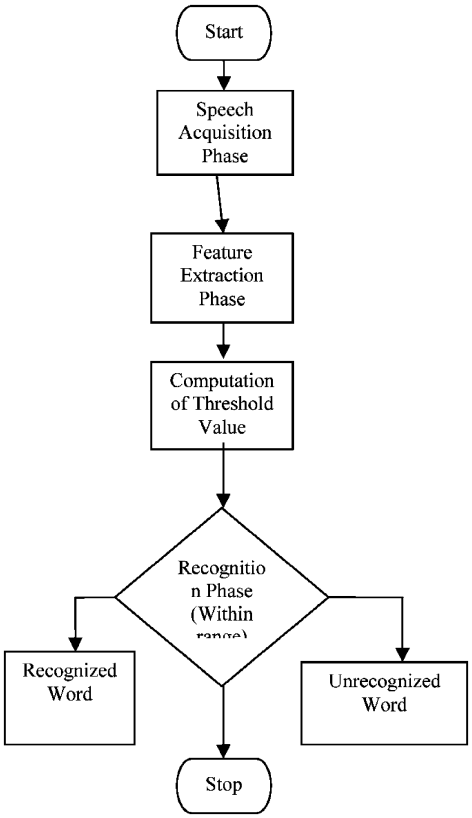


Figure 1

There are many classes of algorithm which can be used for making speech recognition engine.

Template matching algorithm: It recognizes the word by comparing the extracted features template of spoken words with stored template.

Hidden Markov Model (HMM): In HMM each word is represented by a set of state with probability of transition from one state to other state. These states are generated by training procedure.

Neural network: this is a kind of associative memory which can be used to associate to each speech signals with words.

Literature Review

In 2001[1] speech signal modeling techniques for recognition of isolated word were given. In These techniques features spectral and temporal were computed for recognition of alphabet based on ISOLET database and HMM toolkit was used to implement it. Voice dialing [2] simulator were developed by Ericsson for recognition of isolated word for use in mobile phones. The speech recognition algorithm based on discrete HMM adapting the spectral characteristics of the speech were used. The multi resolution models using HMM achieved higher computer recognition of speech mixed with other sounds. [3]. MVDR was used for robustly estimating envelop of the speech signal proved to be accurate and relatively less sensitive to additive noises. It removed the traditional Mel scaled filter bank, a perceptually motivated frequency partitioning. [5]. Feed forward multi layered perception by back propagation in speech recognition used neural network and was used for speaker independent isolated word recognition on a small vocabulary. [6]. A non linear AM-FM speech model was used toe extract robust features for the speech recognition. It measured the amount of amplitude and the frequency modulation that exist in speech recognition.[7]. Pattern recognition and audio processing were important aspects in visual and audio stimuli in order to reflect intelligent behavior. The use of neural network and linear predictive coding techniques together improved the performance of speech and text dependency recognition. [9] Machine vision and image processing was widely used to recognize different patterns and extract distinctive features of the images. These techniques were used for spectral image analysis.

Design and Implementation

A speech recognition system consists of two main parts: training unit and testing unit. Training speech data is input in the training unit which generates a model. This model is then used by testing unit. The testing speech data is fed to the testing unit which performs pattern matching using the model obtained from the training unit. The speech data is

pre-processed and set of features are extracted from the speech data. The proposed system is implemented using matlab 6.1.

Data Speech Acquisition: The analog speech signals for the isolated word forward are recorded using microphone, converted and stored into digital speech signal. The stored speech signal is in the form of wave files. The speech samples thus obtained are stored for further computation. The speech sample are:

Audio sampling rate	22 kHz
Bit rate	352 kbps
Audio sampling rate size	16 bit
Channel	1
Audio formats	PCM

Feature Extraction: Meaningful features can be extracted using the following methods:

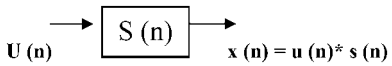
A Fast Fourier Transform (FFT) Algorithm: FFT extracts frequency amplitude for discrete intervals from a given segment of a signal.

Linear Predicting Coding (LPC): LPC input waveform is considered to be a convolution of various periodic functions. In LPC periodic signal of input waveform can be approximated from linear combination of past signal values.

$$X(n) = a_1 x(n-1) + e(n)$$

Group of few LPC coefficient and error signals can be used to predict entire waveform.

Cepstral Processing: It uses the source filter model of speech. When the spoken word is comprised of



Cepstrum is defined as

$$C = IDFT(\log |DFT(x(n))|)$$

Training the System

The different speakers speech signals stored in the database are read using wavread command. The physical features of the speech signal are extracted. The extracted features are used for calculating the threshold range for the respective feature category

Testing the System

For testing the accuracy of the system in the real application the following steps are performed on the system. A. The utterance of the unknown isolated word is analyzed. B. The features of the unknown are extracted. C. The extracted features are compared with the threshold values. D. If the extracted feature value falls within the range the isolated word is recognized or rejected.

The recognizer was trained using 10 No of utterances of the isolated word forward from different speakers. 16 No of utterances were used for the testing the system.

CONCLUSIONS

Speaker independent optimal isolated word recognition was implemented for an English word forward. The results were found to be satisfactory. The accuracy percentage for the trained set recognized was 95% and for the test /real set was 81.23%. The energy feature had the highest accuracy rate. It is concluded that physical features can also be used for speech recognition.

Future Scopes

Speech recognition is widely preferred area of research. It has brought technology to new level. The work of the project can be carried forward by an integrated approach of using different feature extracted methods. The integrated approach may increase the accuracy rate. The challenging task is of developing techniques for acquisition of quality speech signals. The accuracy rate may further be enhanced by eliminating the noise from the speech signal. Another great task is of dealing with misspoken words, extreme emotions, poor or inconsistent room acoustics, sickness...

REFERENCES

[1] Karnjanasecha Montri and Zahorian Srephen A., " Signal modeling for high performance robust isolated word recognition" , IEEE transactions on speech and audio processing , vol.9, no.6, September 2001

[2] Kovacevic Mauricio Aracena, Dwhlbom Anna, Ekeberg Jakob, Gariazzo Guillaume, Lasth Eric and Tronoso Vanessa," Ericsson T18s Avoice Dialing Simulator," Royal Institute of Technology, Sweden ,2001.

[3] Harding Sue and Meyer Georg. "Multi Resolution Auditory Scene Analysis: Robust Speech Recognition using Pattern Matching from a Noisy Signal," Euro speech 2003, Geneva.

[4] Oh Wook Kwon, Kwokleung Chan, Jiucang Hao and Te-Won Lee, "Emotion Recognition by Speech Signals ," Euro speech 2003-Geneva.

[5] Yapanel Umit H. and. Hansen John H.L, "A New Perspective on Features Extraction for Robust in –vehicle Speech Recognition." Euro Speech 2003 Geneva.

[6] Chin Luh Tan and Adznan Jantan, "Digit Recognition using Neural Networks", Malaysian Journal of Computer Science, 17, No.2, December 2004, pp40-54.

[7] Panagiotakis Costas and Tzititas George, "IEEE Transaction on Multimedia", 7, No. 1, February 2005.

[8] Dimitriadis Dimitrios, Maragos Petro and Potamianos Alexandros, "Robust AM-FM Features for Speech Recognition ," IEEE Signal Processing Letters, 12, No.9, September 2005.

[9] Francisca Natalia , Solano Gonzalez , " Pattern Recognition and Text Dependent Voice Recognition System for a Biped

Robot," University of Puerto, Rico, 2006.

[10] Jyoti, Singhai Rakesh,"Automatic Speaker Recognition: An Approach using DWT Based Feature Extraction and Vector Quantization", IETE Technical Review, 24, No. 5, Sept-Oct 2007, pp 395-402.

[11] Demori, Merlo, Palakal, Rouat, "Use of Procedural Knowledge for Automatic Speech Recognition", pp 840-842.

[12] Sharma Vandana, "Speak Indic Telisma is Listening", Information Technology, Dec 2007, pp 36 - 37.

[13] http:// 128.59.66.180/doc/HTKBook21/node5.html.