Electronic Language Interpreter using Isolated word algorithm

Sujay Khole
sujaykhole@gmail.com

Rahul Gosavi
rahulgosavi9@gmail.com

Ashish Khenat
khenatashish@gmail.com

K.T.Talele
talelesir@gmail.com

***Abstract*— Speech recognition has found its application on various aspects of our daily lives from automatic phone answering service to dictating text and issuing voice commands to computers. In this paper, we present the steps involved in the design of a speaker-independent speech translation system. We focus mainly on the pre-processing stage that extracts salient features of a speech signal and a technique called Dynamic Time Warping commonly used to compare the feature vectors of speech signals. These techniques are applied for recognition of isolated words spoken. Finally, we design a simple Voice-to-Voice converter application using independent hardware system.**

***Keywords*: DTW, speech, windowing.**

I. Introduction

Language is mans most important means of communication and speech its primary medium. Speech provides an international forum for communication among researchers in the disciplines that contribute to our understanding of the production, perception, processing, learning and use. Spoken interaction both between human interlocutors and between humans and machines is inescapably embedded in the laws and conditions of Communication, which comprise the encoding and decoding of meaning as well as the mere transmission of messages over an acoustical channel. Here we deal with this interaction between the man and machine through synthesis and recognition applications. The paper dwells on the speech technology and conversion of speech into analog and digital waveforms which is understood by the machines. Speech recognition, or speech-to-text, involves capturing and digitizing the sound waves, converting them to basic language units. In this paper, we present how a speech translator system is designed. First, we describe the acoustic pre-processing step that aids in extracting the most valuable information contained in a speech signal. Then, we present an algorithm called Dynamic Time Warping used to recognize spoken words by comparing their feature vectors with a database of representative feature vectors[1]. We also build a simple Voice-To-Text converter application using Open source software Dhvani.

II. Overview of System Working

The system is implemented in 3 stages: speech to text, text to text and text to speech. First of all, input English speech is recognized by speech to text converter using automatic speech recognition system and English speech is converted into English text. Now, text to text conversion selects the hindi text for the corresponding English text from the database and atlast the Hindi text is given to text to speech synthesizer to convert Hindi text into Hindi speech. Thus, the English speech at the input is converted into Hindi speech at the output.

III. Speech to Text

The input analog speech is given to the various stages as shown in Fig. 1[2] and then at the ouput the recognized is converted into text form.

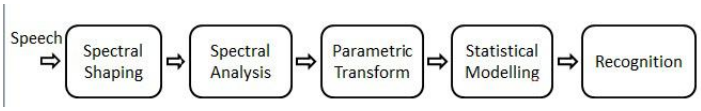

Fig. 1. Speech Recognition Process

A. Spectral Shaping

Spectral shaping involves two basic operations: A/D conversion conversion of the signal from a sound pressure wave to a digital signal; and digital filtering emphasizing important frequencycomponents in the signal. This conversion process is shown in Fig. 2.

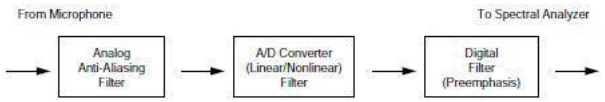

Fig. 2. Spectral Shaping Process

B. Spectral Analysis

The pure speech signal is then divided into no. of frames and windowing is performed using hamming window for each frame and then DFT of each frame is found out as shown in Fig.3.

On the Spectral graph, freq scale is divided into bark scale subbands and the speech vector is formed for each frame computed by summing up amplitude of all frequency components in each subband as shown in Fig.4.

All audio is cut up into frames of length 512 samples with 50 percent overlap and windowed using a 512 point Hamming window. A 1024 (512 non-redundant) point FFT of each of these frames is then taken laying the groundwork for the frame to be divided into sub-bands. In this case the subbands

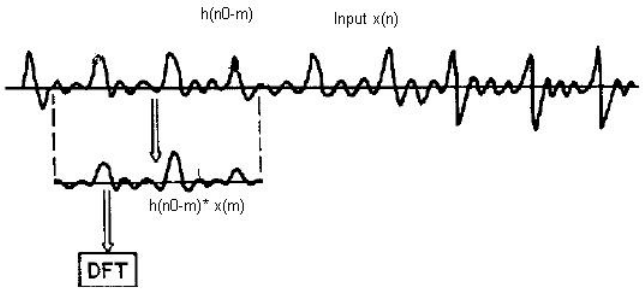

Fig. 3. Windowing using Hamming Window and forming DFT

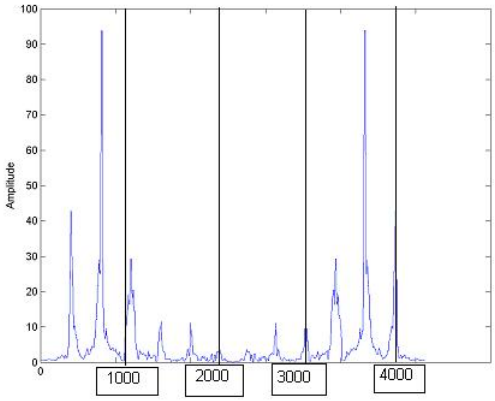

Fig. 4. Dividing Frequency Components into Bark Scales

correspond to the 17 critical bands of human hearing between 0 and 4000Hz. Sub-bands with such a division are often referred to as Bark bands[3].

C. Speech Vector Formation

The speech vector is formed by using the 13 parameters present in the speech depending upon the vocal tract of the speaker. A 2 D vector is formed with the no of rows indicating bark scales and the no of columns indicating no of frames as shown in Fig. 5. So ,a feature vector is obtained for a particular word[4].

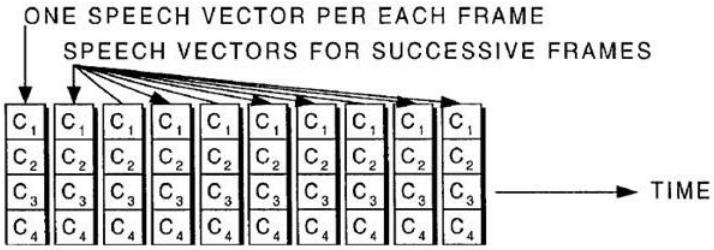

Fig. 5. Speech Vector Formation

D. Dynamic Time Warping Algorithm

A speech signal is represented by a series of feature vectors which are computed every 10ms. A whole word will comprise dozens of those vectors, and we know that the number of vectors (the duration) of a word will depend on how fast a person is speaking. In speech recognition, we have to classify not only single vectors, but sequences of vectors. Lets assume we would want to recognize a few command words or digits. For an utterance of a word w which is TX vectors long, we will get a sequence of vectors X= x0, x1, . . . , xTX-1 from the acoustic preprocessing stage. What we need here is a way to compute a distance between this unknown sequence of vectors X and known sequences of vectors W = w0,w1, . . . ,wTW which are prototypes for the words we want to recognize. The main problem is to find the optimal assignment between the individual vectors of unequal vector sequence X and W . In Fig. 6 we can see two sequences X and W which consist of six and eight vectors, respectively. The sequence W was rotated by 90 degrees, so the time index for this sequence runs from the bottom of the sequence to its top. The two sequences span a grid of possible assignments between the vectors. Each path through this grid (as the path shown in the figure) represents one possible assignment of the vector pairs. For example, the first vector of X is assigned the first vector of W, the second vector of X is assigned to the second vector of W, and so on[5].

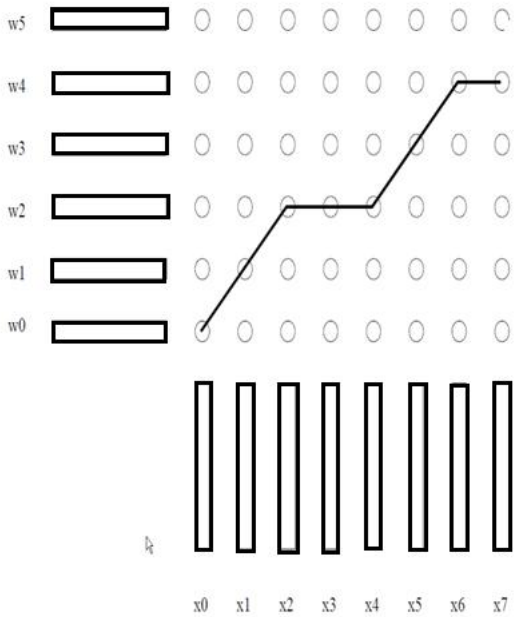

Fig. 6. Implementation of Dynamic Time Warping

IV. Text to Text Conversion

The text to text conversion performs the one to one mapping and finds the words from the database for the corresponding recognized english words in the speech and gives the

ouput in the text form to the speech synthesizer. Fig.7 shows the one to one mapping from English words into Hindi words.

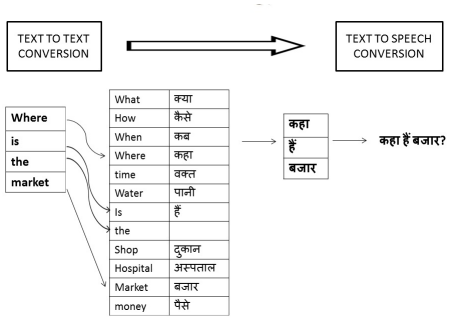

Fig. 7. Text Mapping of Words from Database

V. Text to Speech

Dhvani is a Text to Speech System for Indian Languages. It includes language specific synthesizer modules and an Indian voice database as shown in Fig.8[10]. Dhvani is written in C and has a command line interface as well as a C/C++ API Interface. It includes a speech database and total size of the application is less than 2 MB.

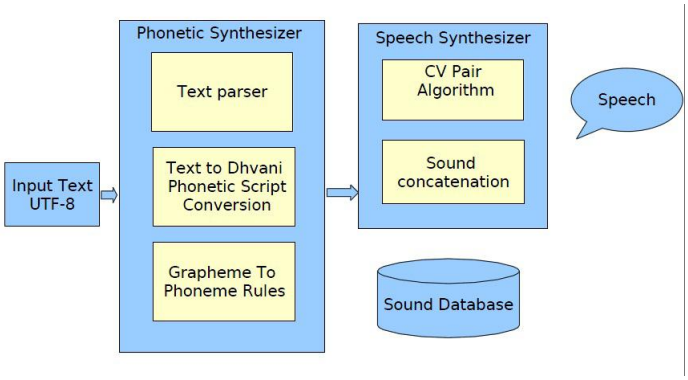

Fig. 8. Text to Speech Conversion Process

VI. Conclusion

Thus, we have developed a system to perform the translation of speech in English language into Indian languages using the three stages:

- Speech to text
- Text to text
- Text to speech

It is possible to develop a speech translation system which is independent of the speaker, with continuous speech input, increasing the database of words and using the speech recognition system for any language by building the acoustic and language models and dictionary. The problems with the systems to be dealt with in future is increasing the accuracy and make it robust system and making the speech synthesizer more clearer.

VII. Future Scope

There is huge scope for continuing future work in the project. Firstly, it is possible to increase the size of the limited vocabulary which currently has only 150 words. A dictionary with a vocabulary of around 600 words would be large enough and will include most of the commonly used words. Also, it would be advisable to add more languages to the supported languages. Focus should be on adding Indian languages such as Malyalam, Kannada, Sanskrit, Assamese and Punjabi. Scope of the project can be increased widely by integrating language support for major international languages such as French, German, Russian, Italian and Spanish. Online translation services, android-applications and devices for these languages are available com- mercially. These can be then integrated with our project. Mandarin is another international language option that should be considered.

References

[1] Speech Recognition on DSP: Algorithm Optimization and Performance Analysis YUAN Meng, The Chinese University of Hong Kong, July 2004.
[2] Lawrence Rabiner, Biing Hwang Juang, Fundamental of Speech Recognition , Copyright 1993 by ATT.
[3] L.R.Rabiner and B.H.Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliff, New Jersy, 1993.
[4] Signal Modeling Techniques In Speech Recognition by, Joseph Picone Texas Instruments, Systems and Information Sciences Laboratory, Tsukuba Research and Development Center, Tsukuba, Japan.
[5] Project report on Speech Recognition with Dynamic Time Warping using MATLAB by Palden Lama and Mounika Namburu.
[6] M.A.Anusuya and S.K.Katti,"Speech Recognition by Machine: A Review",(IJCSIS) In- ternational Journal of Computer Science and Information Security, Vol. 6, No. 3, 2009.
[7] Tool for the development of speech synthesis system Kalika Bali, Partha Pratim Talukdar, N. Sridhar Krishna HP Labs India; A.G. Ramakrishnan, Indian Institute of Science, Bangalore.
[8] Design of Multilingual Speech Synthesis System S. Saraswathi, R. Vishalakshi.
[9] Vani-An Indian Language Text To Speech Synthesizer by Harsh Jain, Varun Kanade, Kartik Desikan, Department of Computer Science and Engineering, Indian Institute of Technology Mumbai,April 2004.
[10] http://dhvani.sourceforge.net/doc/index.html