ISSN 2395-1621

Content Based Retrieval of Text From Speech

#1Ninad Gokhale, #2Rohan Hulsure, #3Ashish Deshpande, #4Omkar Thombre

1gokhale.ruturaj.m@gmail.com
2rohanhulsure89@gmail.com
3ashishdeshpande2592@gmail.com
4thombare.omkar17@gmail.com

#1234Computer Engineering Department
PVPIT Pune Maharashtra India, Pune University

ABSTRACT

A text to speech conversion system is an application that converts text into speech (spoken words), by analyzing and processing the text using natural language processing (NLP) and then using Digital Signal Processing (DSP) to convert speech into text format. Here we developed a useful speech to text conversion system in simple form that convert inputted speech format into text which can then be saved in text file format. The development of text to speech conversion system will be of great help to people with visual formation.

Keywords: **Key term extraction; Automatic Text Retrieval; Speech to text conversion; Speech Processing; Database; Hidden Markov Model, Artificial Intelligence.**

ARTICLE INFO

Article History

Received: 31st May 2016

Received in revised form :

31st May 2016

Accepted: 2nd June 2016

Published online :

13th June 2016

I. INTRODUCTION

The Speech is one of the most important tool for communication and environment. The speech recognition made it feasible for machine to understand human language. As information technology has increased more aspects of our lives with every year, the problem of communication between human and devices become increasingly significant. Up to now communication has almost based on the use of keyboards and screens, but speech is most widely used, natural and fastest means of communication for people. In many systems the voice can be taken as input but it is not efficient due to noise. So there are many models that give perfection which reduces noise. Several models use the different approach that helps to speech retrieval, processing of speech and conversion of speech to text. In a speech to text system, many parameters affect the accuracy of the System. These parameters are: dependence or independence from speaker, discrete or continuous word recognition, vocabulary, environment, acoustic model, language model, etc. Problems such as noisy environment, different pronouncing of one word by one person in several times, dissimilar expressing of one word by two different speakers. Resolving each of these problems is a good step towards this aim.

II. LITERATURE SURVEY

Speech recognition came into existence during 1920. The first machine i.e. Radio Rex, a toy to recognize voice was manufactured. Bell Labs developed a speech synthesis machine at the World fair in New York. But later on they discarded efforts based on an incorrect conclusion that the AI is ultimately required for success. In order to develop systems for ASR, attempts were made in 1950s where researchers studied the fundamental concepts of phonetic-acoustic. Most of the systems in 1950[1] for recognizing speech examines the vowels spectral resonancews of each utterances. At Bell Labs Davis, Biddulph and Balashek (1952) premeditated a isolated digit recognition system for a single speaker[2] using formant frequencies estimated during vowel regions of each digit. At RCA Labs, Olson and Belar (1950) built 10 syllables recognizer of a single speaker [3] and Forgie and Forgie built a speaker-independent 10-vowel recognizer [4] at MIT Lincoln Lab, by measuring spectral resonances for vowels. Fry and Denes (1959) tried to build a phoneme recognizer to recognize four vowels and nine consonants [5] at University College in England by using a spectrum analyser and a pattern matcher to make the recognition decision. Japanese labs entering recognition field in 1960-70. As computers are not fast enough, they designed special purpose H/W as a part of their system. In Tokyo, Nagata et.al described a system of the Radio Research lab, was a H/W vowel recognizer. Another effort was the work of Sakai and Doshita in 1962, of Kyoto University who built a H/W phoneme recognizer. In 1963, Nagata and co-workers at

NEC Labs built a digit recognizer. This led to a long productive research program.

In 1970, the key focus of research was on isolated word recognition. IBM researchers studied in large vocabulary speech recognition. At AT&T Bell Labs, researchers began speaker independent speech recognition experiments. A large number of clustering algorithms were used to find the number of distinct patterns required to represent words to lexical model for non-native speech recognition. This research has been refined so that the techniques for speaker independent patterns are widely used. Carnegie Mellon University's Harphy system recognize speech with vocabulary size of 1011 words with reasonable accuracy. It was the first to make use of finite state network to reduce computation and determine the closest matching strings efficiently.

In 1980, the key focus of research was on connected words speech recognition. In the beginning of 1980, Moshey J. Lasry studied speech spectrogram of letters and digits and developed a feature based speech recognition. There is a change in technology in 1980 from template based approaches to statistical modelling approach specially HMM in speech research.

III. PROPOSED SYSTEM

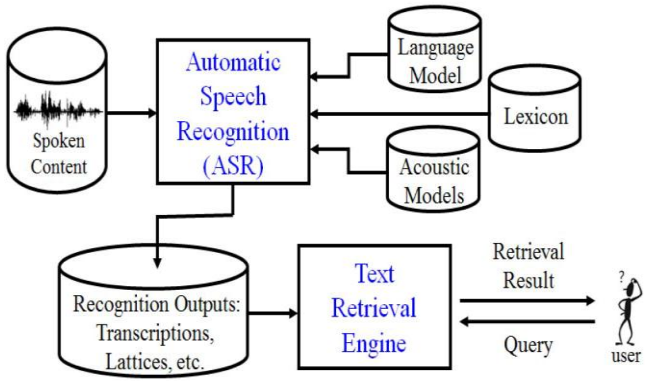

Fig 1: System Architecture

A waveform modifier takes an input speech signal and produces a modified signal. The modification might be a clipping of large values of the signal; a frequency spectrum filtering that alters the shape of signal or enhances the speech and de-emphasises noise that is present. A symbol transducer can take in one discrete symbol sequence and yields a modified sequence on its output. If the input were a sequence of words in one language and the output were an equivalent word sequence in another language, this transducer would be a language translation device. Parameter extractor takes an input speech signal and yielding parameters of speech wave. In recognizers, it is often called the pre-processor. A feature extractor can receive parameters and produce a more abstract set of important information carrying features such as determining what portions of speech are voiced, whether the sound is loud and resonant like a vowel etc. A segmenter and labeller can receive the set of features and produce a linear string of phonemes or other identified segments. The unit identifier takes input symbol sequence which may be compared to the expected reference sequences for various units to determine what linguistic units appear to be in the input. The most common unit identifier is a 'word matcher' which finds the closest matching word, based on which word's stored pronunciation string is most like the input string. With these building blocks, we have the essential prerequisites for discussing the main knowledge sources needed for machine understanding of speech. A speech recognition system consists of five blocks:

Feature extraction, Acoustic modelling, Pronunciation modelling, Decoder. The process of speech recognition begins with a speaker creating an utterance which consists of the sound waves. These sound waves are then captured by a microphone and converted into electrical signals. These electrical signals are then converted into digital form to make them understandable by the speech system. Speech signal is then converted into discrete sequence of feature vectors, which is assumed to contain only the relevant information about given utterance that is important for its correct recognition. An important property of feature extraction is the suppression of information irrelevant for correct classification such as information about speaker (e.g. fundamental frequency) and information about transmission channel (e.g. characteristic of a microphone).Finally recognition component finds the best match in the knowledge base, for the incoming feature vectors. Sometimes, however the information conveyed by these feature vectors may be correlated and less discriminative which may slow down the further processing. Feature extraction methods like Mel frequency cepstral coefficient (MFCC) provides some way to get uncorrelated vectors by means of discrete cosine transforms (DCT).

IV. RESULT ANALYSIS

Hidden Markov Model: The Hidden Markov Model Template comparison Methods of Speech recognition (e.g., dynamic Time warping directly compares the unknown utterance to known examples. Instead HMM Created stochastic Models from known Utterances and compares the probability that the unknown utterance was generated by each model. HMMs are a broad class of doubly stochastic

Models for non stationary signals that can be inserted into other stochastic models to incorporate information from several hierarchical knowledge sources. Since we do not know how to choose the form of this model automatically but, once given a form, have efficient automatic methods of estimating its parameters, we must instead choose the form according to our knowledge of the application domain and train the parameters from known data. In standard HMMs, the sequential data in a Markov state are assumed to be conditionally independent, and are represented by state-dependent Gaussian mixture models (GMMs). The corresponding state sequence is determined according to the accumulated likelihood function and the state transition probabilities. HMM parameters are prone to be over trained when using maximum-likelihood (ML) estimation. We question whether GMMs are a good model for representing training data as well as for generalizing to unknown test data. Predictive HMMs were accordingly proposed to address the over fitting problem by incorporating the uncertainties of HMM parameters, expressed by prior distributions, in a Bayesian model comparison. Another approach was given by buried Markov models, which relaxed the conditional independence assumption and were used to represent speech features with dependencies between observations. A set of state-dependent basis vectors was trained to express the conditionally dependent feature vectors. In yet another approach, subspace Gaussian mixture models were constructed to represent speech features by using state-dependent weights and a common large-scale GMM structure. The feature representation was seen as sensing based on different subspaces of a global GMM. Recently, canonical state models, which consist of state-dependent transforms and a set of canonical state parameters that act as the bases for feature representation or likelihood calculation, were proposed.

MFCC

In sound processing the mel-frequency cepstrum (MFC) is a representation of the short-term power cepstrum of a sound, based on a linear cosine transform of a log power spectrum on a mel scale of frequency.

Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound, for example, in audio compression.

MFCCs are commonly derived as follows:

1. Take the Fourier transform of (a windowed excerpt of) a signal.
2. Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
3. Take the logs of the powers at each of the mel frequencies.
4. Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
5. The MFCCs are the amplitudes of the resulting spectrum.

There can be variations on this process, for example: differences in the shape or spacing of the windows used to map the scale, or addition of dynamics features such as "delta" and "delta-delta" (first- and second-order frame-to-frame difference) coefficients. fast enough, they designed special purpose H/W as a part of their system. In Tokyo, Nagata et.al described a system of the Radio Research lab, was a H/W vowel recognizer. Another effort was the work of Sakai and Doshita in 1962, of Kyoto University who built a H/W phoneme recognizer. In 1963, Nagata and co-workers at NEC Labs built a digit recognizer. This led to a long productive research program.

In 1970, the key focus of research was on isolated word recognition. IBM researchers studied in large vocabulary speech recognition. At AT&T Bell Labs, researchers began speaker independent speech recognition experiments. A large number of clustering algorithms were used to find the number of distinct patterns required to represent words to lexical model for non-native speech recognition. This research has been refined so that the techniques for speaker independent patterns are widely used. Carnegie Mellon University’s Harphy system recognize speech with vocabulary size of 1011 words with reasonable accuracy. It was the first to make use of finite state network to reduce computation and determine the closest matching strings efficiently.

In 1980, the key focus of research was on connected words speech recognition. In the beginning of 1980, Moshey J. Lasry studied speech spectrogram of letters and digits and developed a feature based speech recognition. There is a change in technology in 1980 from template based approaches to statistical modelling approach specially HMM in speech research.

V. CONCLUSION

Speech is the most prominent and primary mode of communication between human beings. Over the past five decades, research in the area of speech is a first step towards ordinary man-machine communication. We also have encountered some limitations. What we know about speech processing is very limited. This paper attempts to give a comprehensive survey of research in speech recognition and some year-wise progress to this date and its current status. Although significant amount of work has been done in the last two decades but there is still work to be done.

At present research is focusing on creating and developing systems that would be much more robust against variability and shift in acoustic environment, speaker characteristics, language characteristics, external noise sources etc. It has been found that HMM is the best technique in developing language model. Speech recognition is very fascinating problem. It has attracted scientists and researchers and created a technological bang on society and is expected to boom further in this area of man machine interaction.

REFRENCES

[1] G. Tur and R. DeMori, Spoken Language Understanding: Systems for Extracting Semantic Information from Speech. New York,NY,USA: Wiley, 2011, ch. 15, pp. 417–446.

[2] C. Chelba, T. Hazen, and M. Saraclar, “Retrieval and browsing of spoken content,” IEEE Signal Process. Mag., vol. 25, no. 3, pp. 39–49, May 2008.

[3] M. Larson and G. J. F. Jones, “Spoken content retrieval: A survey of techniques and technologies,” Found. Trends Inf. Retr., vol. 5, pp. 235–422, 2012.

[4] L.-s. Lee and Y.-C. Pan, “Voice-based information retrieval–how far are we from the text-based information retrieval?,” in Proc. ASRU,2009.

[5] L.-s. Lee and B. Chen, “Spoken document understanding and organization,”IEEE Signal Process. Mag., vol. 22, no. 5, pp. 42–60, Sep. 2005.

[6] D.B.Fry, 1959, Theoritical Aspects of Mechanical speech Recognition , and P.Denes, The design and Operation of the Mechanical Speech Recognizer at Universtiy College London, J.British Inst. Radio Engr., 19:4,211-299.

[7]. K.Nagata, Y.Kato, and S.Chiba, 1963, Spoken Digit Recognizer for Japanese Language, NEC Res.Develop., No.6.

[8]. T.Sakai and S.Doshita, 1962 The phonetic typewriter, information processing 1962, Proc.IFIP Congress.

[9]. L.R.Rabiner, S.E.Levinson, A.E.Rosenberg, and J.G.Wilpon, August 1979, Speaker Independent Recognition