ISSN: 2321-7782 (Online)

Volume 2, Issue 2, February 2014

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Paper / Case Study

Available online at: www.ijarcsms.com

Speech to Text Technology: A Phonemic Approach for Devanagari Script

Nanekar Priyanka S[1]
Computer department
MIT College of Engineering
Pune – India

Kohli Urvashi[2]
Computer department
MIT College of Engineering
Pune – India

Bhosale Komal[3]
Computer department
MIT College of Engineering
Pune – India

Yeolekar Rishikesh[4]
Prof. Computer department
MIT College of Engineering
Pune – India

Abstract: This paper presents a brief study of Automatic Speech Recognition (ASR) and discusses the phonemic approach for Devnagari Script. Despite years of research and impeccable progress of the accuracy of ASR, it remains one of the most important research challenges, which calls for further research. The Devnagari Script inspires the work presented. It arranges 46 phonemes based on the process of its generation. Devnagari is based on phonetic principles, which consider the articulation of the consonants and the vowels they represent.

The basic unit of any language is a phoneme. These phonemes are divided into two types: vowel phonemes (swara varna) and consonant phonemes (vyanjan varna). This paper discusses the fundamentals of speech recognition, the understanding of various kinds of speech and further aspects relating to the Devnagari script.

Keywords: Automatic Speech Recognition (ASR), Devnagari Script, Phonemes, Swara Varna, Vyanjan Varna.

I. Introduction

The speech recognition is the translation of spoken words into text. ASR is the process in which the acoustic speech signal is mapped to text. This process is highly complex since the spoken words has to be matched with stored sound bites on which further analysis has to be done because not all sound bites match with pre-existing sound pieces. Determination of undetermined piece of sound requires computing power. The recognition program is designed to enable the computer to recognize the input and form text. The text stored is based on a huge set of parameters and grammar, which defines human speech. Thus the process begins when a speaker(user) speaks a sentence. The software then produces a speech waveform, which manifests the words of the sentence as well as the background noise and pauses in the spoken input. The software then attempts to decode the speech into the best estimate of the sentence. Thus the three features necessary for an ASR include large vocabularies, continuous speech capabilities, and speaker independence.

Phonemes are the basic unit of speech of any language. Each language has its own set of phonemes, typically numbering between 30 and 50. The Devnagari script has a set of 46 phonemes. Voice is produced when different phonemes are articulated.

The speech uttered by different persons may vary but the way it is uttered is the same, i.e., the signal extracted is the same because the vibrations produced must be similar for the same phoneme.

Every letter and its pronunciation is unique and can’t be represented or pronounced by using other letters which gives a unique representation of every word. This feature is absent in non-phonetic languages like English in which one pronunciation can be done in more than one way, e.g. there, their bot are pronounced similarly.

II. SYSTEM WORKING

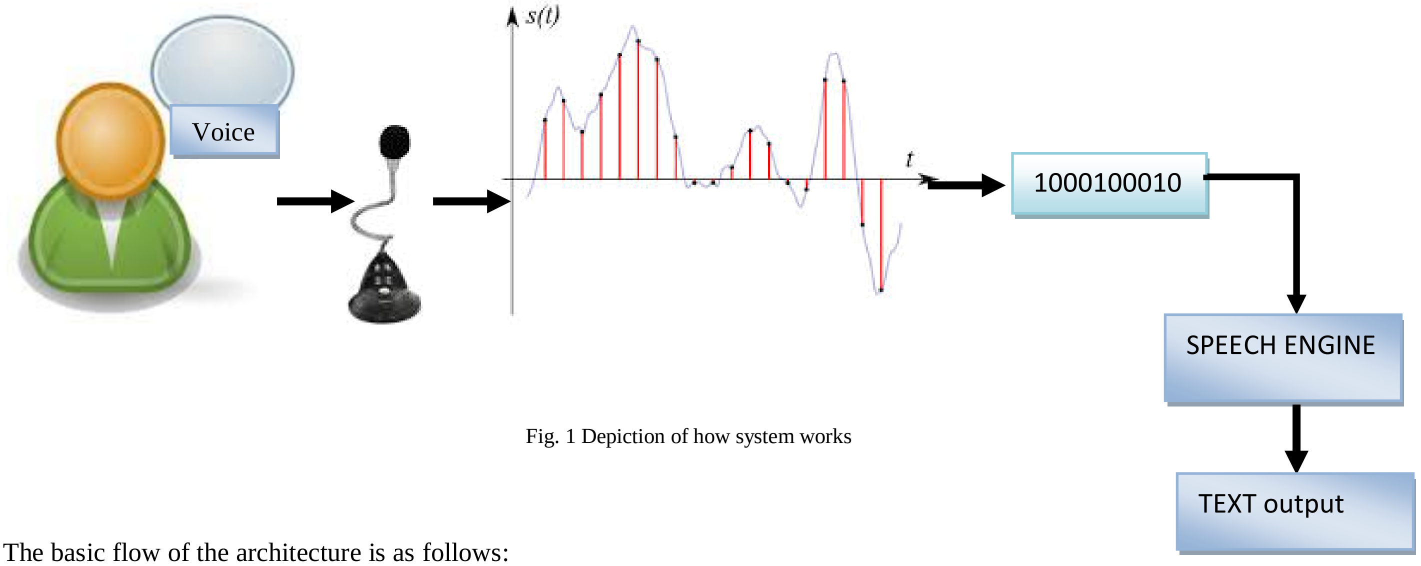


Fig. 1 Depiction of how system works

The basic flow of the architecture is as follows:

A. VOICE INPUT:

The voice is taken as an input with the help of either the speaker or the headphone. Whatever the user speaks is to be captured and correctly identified.

B. CONVERSION:

The sound card captures the sound waves and produces the equivalent digital representation of audio that was received with the help of a microphone/speaker.

C. DIGITIZATION:

The process of converting the analog signal into a digital form is known as Digitization. It involves the both sampling and quantization processes. Sampling is converting a continuous signal into discrete signal, while the process of approximating a continuous range of values is known as quantization.

D. SPEECH ENGINE (ASR):

The job of speech recognition engine is to convert the input audio into text. To accomplish this it uses all sorts of data, software algorithms and statistics. Once audio signal is in proper format (after digitization) it then searches the best match for it. It does this by considering the words it knows, once the signal is recognized it returns its corresponding text string.

CLASSIFICATION OF SPEECH:

The following are the categories of speech recognition system based on their ability to recognize words. These categories are based on the various kinds of issues that are faced by the ASR while determining the sentence that the speaker speaks.

ISOLATED SPEECH: involves words a pause between two utterances. The recognizer requires a single utterance at a time, although it can intake more than one word at a time. Often, these systems require the speaker to wait between utterances (usual processing during the pauses). This is one way to determine the words spoken by the ASR.

CONNECTED SPEECH systems are similar to isolated words, but allow separate utterances with minimal pause between them. An utterance is a spoken word, statement, or vocal sound. Utterance can be a single word, or a collection of a few words, a single sentence, or even multiple sentences.

CONTINUOUS SPEECH allows the user to speak almost naturally. This is a difficult task as compared to the other two since the utterance boundaries are to be determined.

SPONTANEOUS SPEECH is natural sounding and unrehearsed. An ASR System with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run together.

APPROACHES TO SPEECH RECOGNITION

A. *The acoustic-phonetic approach*:

Small Size Vocabulary	Medium Size vocabulary	Large Size vocabulary	Very-large vocabulary (Out of Vocabulary)
1 to 100 words	101 to 1000 words	1001 to 10,000 words	More than 10,000 words

The acoustic phonetic theory proposes that finite, smallest unit of speech is a phoneme in any spoken language. All the phonetic units, i.e., phonemes manifest into a speech signal.

B. The pattern recognition approach

In this approach the speech patterns are determined without explicit determination or segmentation as in case of the phonetic approach. It involves two steps:

- Training of speech patterns.
- Recognition of patterns

MATCHING TECHNIQUES

Speech-recognition engines use one of the following techniques to match the word.

A. WHOLE-WORD MATCHING: The analog speech signal is converted into digital form. The engine then maps this digital speech signal against a prerecorded template of the word. This technique takes much less processing than sub-word matching, but it requires that the user to prerecord every word that will be recognized - several hundred thousand words. Whole-word matching also requires large amounts of storage.

B. SUB-WORD MATCHING: The engine looks for phonemes and then performs pattern recognition on them. This technique takes more processing than whole-word matching, but it requires much less storage.

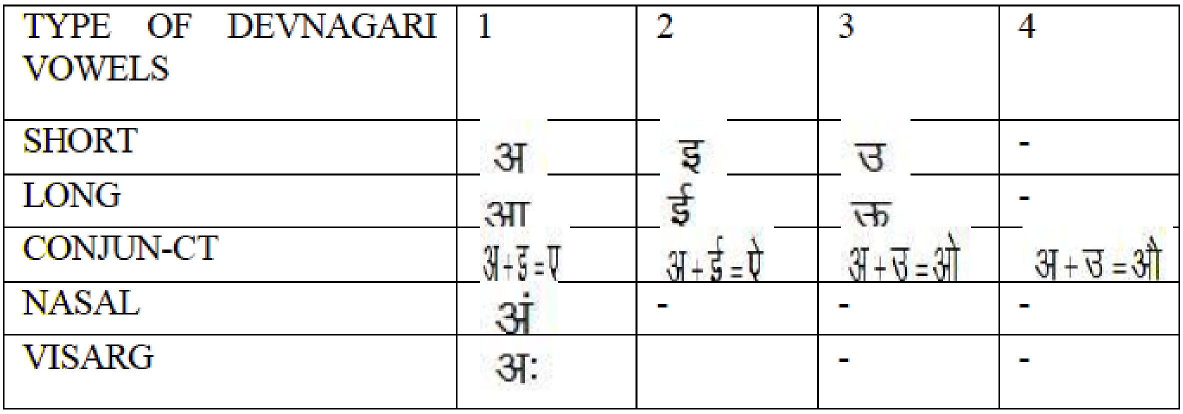
III. The Devanagari Script

The basic unit of any language is a phoneme. These phonemes are divided into two types: vowel phonemes (swara varna) and consonant phonemes (vyanjan varna). They together constitute the Varnamala, which has been referred as a varna-samamnaya.

The combination of consonant phoneme and a vowel phoneme produces a syllable (akshara).The basic Devnagari characters can be combined to indicate combinations of sounds. Conjunct consonants can be divided into six different groups depending on the type of modification of a consonant that takes place. The combination of two forms (C&V) into a syllable, at times creates a new integrated shape or retains partial identity of both the forms.

The syllables formed by adding vowel phonemes /a/, /aa/, /i/, etc. to consonant phoneme are written by creating aksharas. All swara phonemes are added to one consonant phoneme one by one. This concept is called a baaraakhadi.

The combination of two forms (C&V) into a syllable at times creates a new integrated shape or retains partial identity of both the forms. Adding vowel phonemes to a sequence of more than one consonant phoneme can also form syllables.



TYPE OF DEVNAGARI VOWELS	1	2	3	4
SHORT	अ	इ	उ	-
LONG	आ	ई	ऊ	-
CONJUN-CT	अ+इ=ए	अ+ई=ऐ	अ+उ=ओ	अ+उ=औ
NASAL	अं	-	-	-
VISARG	अः		-	-

Fig. 2 Devnagari vowels classified into five types

IV. Conclusion

Now a day's speech recognition system is an very important task. Special use of speech recognition is in cellular phone for typing free messaging and calling. We can also dictate a document and emails and our system will type them for you.

Thus, our main aim is to produce a typing free system. It is actually a very difficult task to implement it and requires a lot of research. Maintaining recognition accuracy and reducing a noise from voice are two important tasks to be carried out.

With the use of phonemes we can create a small database so, that we cannot require internet connection to map the voice signals with words.

It will overcome the requirement of internet connection which is used in other software example, Dragon.

Acknowledgement

We would like to express our sincere gratitude for the assistance and support of a number of people helped us.

We are thankful to Prof. Rishikesh Yeolekar, Department of Computer Engineering, MIT College of Engineering, and our internal guide for his valuable guidance that he provided us at various stages throughout the project work. He has been a source of motivation enabling us to give our best efforts in this project.

We are also grateful to Prof. Dr.Prasanna Joeg, Head of Computer Department, MIT College of Engineering and other staff members for encouraging us and rendering all possible help, assistance and facilities.

References

1. "Android developers", http://developer.android.com
2. J. Tebelskis, Speech Recognition using Neural Networks, Pittsburgh: School of Computer Science,Carnegie Mellon University, 1995.
3. S. J. Young et al., "The HTK Book", Cambridge University Engineering Department, 2006.
4. K. Davis, and Mermelstein, P., "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. Acoust., Speech, Signal Process. vol. 28, no. 4, pp. 357-366,1980.
5. L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains", Ann. Math. Statist., vol. 41, no. 1, pp. 164-171, 1970.

AUTHOR(S) PROFILE

Priyanka Nanekar MIT College Of Engineering.

Komal K. Bhosale MIT College Of Engineering

Urvashi Kohli MIT College Of Engineering

Prof. Yeolekar Rishikesh MIT College of Engineering