Issues Involved In Speech To Text Conversion

Er. Jaspreet Kaur*, Er. Nidhi, Ms. Rupinderdeep Kaur*****

*Department of Computer Science, SVIET, Banur
**Department of Computer Science, SVIET, Banur
***Department of Computer Science, Thapar University, Patiala

Abstract

This document presents various Speech Recognition issues in Indian script. STT synthesis is an online application to convert the speech into the text form. The term "voice recognition" is sometimes used to refer to recognition systems that must be trained to a particular speaker—as is the case for most desktop recognition software. Recognizing the speaker can simplify the task of translating speech. Speech recognition is the process of converting an acoustic waveform into the text similar to the information being conveyed by the speaker. Speech is the most natural way of communication. It includes the fundamentals of speech recognition & different issues like commands by using hyperlink, effectiveness, overlapping speech, low signal to noise ratio, homonyms.

1) Introduction

Speech is a natural mode of communication for people. We learn all the relevant skills during early childhood, without instruction, and we continue to rely on speech communication throughout our lives. Speech recognition is the ability of a machine or program to identify words and phrases in spoken language and convert them to a machine-readable format. Generally, transfer of information between human & machine is accomplished via keyboard, mouse etc. But human can speak more quickly instead of typing. Speech input offers high bandwidth information & relative ease of use. Speech recognition applications include call routing, speech-to-text, voice dialing and voice search. Speech recognition is a solution which refers to technology that can recognize speech without being targeted at single speaker such as a call center system that can recognize arbitrary voices. Speech recognition applications include voice user interfaces such as voice dialing (e.g., "Call home"), call routing (e.g., "I would like to make a collect call")[1,2] .

2) Punjabi Language

Punjabi is an Indo-Aryan language spoken by inhabitants of the historical Punjab region (north western India and in Pakistan). Punjabi is the modern form of Gurmukhi (in India) or Shahmukhi (in Pakistan).Gurmukhi means "from the mouth of the Guru"[3]. According to the Ethnologue[4] 2005 estimate, there are 88 million native speakers of the Punjabi language, which makes it approximately the 10th most widely spoken language in the world. In India, Punjabi is one of the 22 languages with official status in India. It is the first official language of Punjab (India) and Union Territory State Chandigarh and the 2nd official language of Haryana, Himachal Pradesh and Delhi. In Pakistan, Punjabi is the provincial language of Punjab(Pakistan) the second largest and the most populous province of Pakistan.

2.1 Character Set For Punjabi

Punjabi fonts containing the entire Punjabi character set, with simple input. It includes Vowels, Consonants & Auxiliary signs as follows [5]:

2.1.1 Vowels

There are nine vowel phonemes in Punjabi. They are vowels making only one sound. All consonants use the vowel. Table 1 shows the vowels.

	Vowel Sign	Name of vowel	
1.	Invisible	ਮੁਕਤਾ	muktā
2.	◌ਾ	ਕੰਨਾ	kannā
3.	◌ਿ	ਸਿਹਾਰੀ	sihārī
4.	◌ੀ	ਬਿਹਾਰੀ	bihārī
5.	◌ੁ	ਔਂਕੜ	auṅkaṛ
6.	◌ੂ	ਦੁਲੈਂਕੜ	dulaiṅkaṛ
7.	◌ੇ	ਲਾਂਵਾਂ	lāṃvāṃ
8.	◌ੈ	ਦੁਲਾਂਵਾਂ	dulāṃvāṃ
9.	◌ੋ	ਹੋੜਾ	hōṛā
10.	◌ੌ	ਕਨੌੜਾ	kanauṛā

Table 1

2.1.2 Consonants

Punjabi language consists of 41 consonants. Consonants list of Punjabi language is written in Fig 2.

ੳ ਅ ੲ ਸ ਹ
ooraa airaa eeree sassaa haahaa
ਕ ਖ ਗ ਘ ਙ
kakkaa khakkhaa gaggaa ghaggaa ganggaan
ਚ ਛ ਜ ਝ ਞ
chachchaa chhachhchhaa jajjaa jhajjaa yannyaan
ਟ ਠ ਡ ਢ ਣ
tainkaa thatthaa daddaa dhaddaa naanaa
ਤ ਥ ਦ ਧ ਨ
tattaa thatthaa daddaa dhaddaa naanaa
ਪ ਫ ਬ ਭ ਮ
pappaa phapphaa babbaa bhabbaa mammaa
ਯ ਰ ਲ ਵ ੜ
yayyaa raaraa lallaa vavvaa raaraa
ਸ਼ ਖ਼ ਗ਼ ਜ਼ ਫ਼
shashshaa khakkhaa gaggaa zazzaa faffaa

Fig 2

2.1.3 Auxiliary Signs

It serves to add a nasal sound to a particular vowel. These signs are represented in Fig 3.

ੰ ੱ

ਂ ੁ ੍

Fig 3

3) Fundamentals of Speech Recognition

Speech recognition is a multileveled pattern recognition task, in which acoustical signals are examined and structured into a hierarchy of subword units (*e.g.* phonemes), words, phrases, and sentences. Each level may provide additional temporal constraints, *e.g.* known word pronunciations or legal word sequences, which can compensate for errors or uncertainties at lower levels. This hierarchy of constraints can best be exploited by combining decisions probabilistically at all lower levels, and making discrete decisions only at the highest level[6]. The structure of a standard speech recognition system is illustrated in Fig 4.

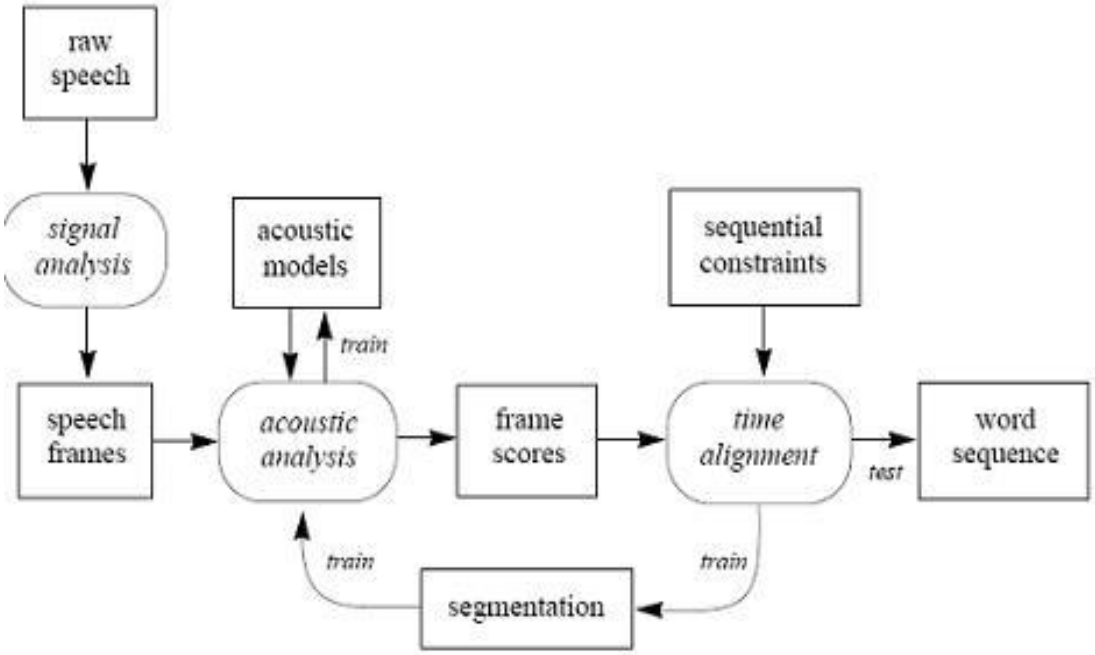

Fig 4. Structure of a standard speech recognition system.

4) Different issues

The various key issues were seen in speech to text convertor for English language using some online softwares. While speaking, the vocal tract of human being can vary widely in terms of their accent, pronunciation, articulation, roughness, nasality, pitch, volume, and speed. All these sources of variability make speech recognition, even more than speech generation, a very complex problem. So, these issues or the problem may be avoided in our STT for Punjabi language. We can work out for improving our STT convertor by taking into consideration following points :

4.1 Commands By using Hyperlink

We can apply the commands by using the hyperlinks, when we speak any command as given in our database, then it helps to open the link by navigating to that link.

4.2 Speaker dependence vs. independence

By definition, a speaker dependent system is intended for use by a single speaker, but a speaker independent system is intended for use by any speaker. Speaker independence is difficult to achieve because a system's parameters become tuned to the speaker(s) that it was trained on, and these parameters tend to be highly speaker-specific.

4.3 Effectiveness

The effectiveness of the speech recognition system can be described when a group of people speak the same speeh and we get the same output as matched by some of people *i.e.* if we consider 10 people to speak the same speech and only 6-7 of them are matched and give the same result. It shows that how effective our speech recognition system is.

4.4 Adverse Conditions

A system's performance can also be degraded by a range of adverse conditions. These include environmental noise (*e.g.* noise in a car or a factory); acoustical distortions (*e.g.* echoes, room acoustics); different microphones (*e.g.* close-speaking, omnidirectional, or telephone); limited frequency bandwidth (in telephone transmission); and altered speaking manner (shouting, whining, speaking quickly, *etc.*).

4.5 Low signal-to-noise ratio

The program needs to "hear" the words spoken distinctly, and any extra noise introduced into the sound will interfere with this. The noise can interrupt the user from a number of sources, including loud background noise in an office environment. Users should work in a quiet room with a quality microphone positioned as close to their mouths as possible. Low-quality sound cards, which provide the input for the microphone to send the signal to the computer, often do not have enough shielding from the electrical signals produced by other computer components.

4.6 Overlapping of speech

There are so many systems which have difficulty separating simultaneous speech from multiple users(like in group discussion). If we try to employ recognition technology in conversations or meetings where people frequently interrupt each other or talk over one another, you're likely to get extremely poor results.

4.7 Homonyms

Homonyms are two words that are spelled differently and have different meanings but sound the same. "There" and "their," "air" and "heir," "be" and "bee", “hair” and “hare”, “bear” and “beer” are all examples. There is no another way for a speech recognition program to differentaite between these words based on sound alone. However, extensive training of systems and statistical models that take into account word context can greatly improved their performance.

5) Conclusion

In order to solve these issues we have to improve our speech to text convertor and improve recognition rate for Punjabi words. In some cases we have found some ambiguity problems and efficient methods should be followed to remove the ambiguity problem. To get best speech synthesis rate, database of system should be modified so that our speech recognition engines suits our requirements. A number of methods can be followed to build an effective and efficient database of proposed system.

6) References

1. *Kuldeep Kumar, R. K. Aggarwal “* HINDI SPEECH RECOGNITION SYSTEM USING HTK” , International Journal of Computing and Business ResearchISSN (Online) : 2229-6166 Volume 2 Issue 2 May 2011

2. Jurafsky, D and Martin, J H (2009) *Speech and Language Processing,* Pearson Education, New Delhi, India.

3. Dinesh Kumar, Neeta Rana “Speech Synthesis System for Online Handwritten Punjabi Word: An Implementation of SVM & Concatenative TTS”, *International Journal of Computer Applications (0975 – 8887)Volume 26– No.2, July 2011*

4. In India And 70 million in Pakistan Punjabi language at *Ethnologue* (16th ed., 2009)

5. Rajiv K . Sharma, Dr. Amardeep Singh “Segmentation of Handwritten Text in Gurumukhi Script” , International Journal of Image Processing

6. Fundamental Of Speech Recognition, “http://www.learnartificialneuralnetworks.com/speechrecognition.html"