ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 1, January 2013

STUDY OF VARIOUS ISSUES IN VOICE TRANSLATION

Aakash Nayak[1], Santosh Khule[2], Anand More[3], Avinash Yalgonde[4],Dr. Rajesh S. Prasad[5]
[1,2,3,4]Student ,[5]Professor at Computer Department,
ZES's Dnyanganga College Of Engineering And Research,Pune-411041.

***Abstract*— In today's world the language translation plays an important role. Requirement of speech translation is increasing day by day. In this paper, we have reviewed various approaches to translate the actual voice/speech. The content of this paper includes different issues regarding to speech translation, and challenges related to speech translation. More so, these factors become important when we need to achieve real-time speech translation on mobile phones.**

***Keywords* - language interpretation, template matching, probability search, speech recognition, session- based cache.**

1. INTRODUCTION

Basically any form of written or speech communication is through various Languages. The communication among human computer interaction is called human computer interface. Nowadays, the presence of multiple languages has been a hindrance to effective communication. In India the language and dialect changes with region, the requirement of a middle translation layer that can eliminate the linguistic barriers becomes essential. [5]

Speakers from different regional identities should be able to interact with one another without the need to understand individual languages. There are two ways of Translation:
1: Text-based translation and
2: Voice-based translation.

Near about all translation methods make use of dictionaries to find the translated words. However, searching for required words and synonyms from such large dictionaries is very slow and very time-consuming. It also depends on the content of the sentence being translated. That means if you have larger sentence then it takes much time to convert into target language.

At First Text-based translation services mainly focus around capturing words and converting them to any target language and however, voice based translation services have remained few and slow. This is because most of these models concentrate mainly on language interpretation and language generation. They fail to take into consideration the large amount of back-end processing that takes place while translation.

In this paper, We are going to discuss the issues which are being faced while developing Speech to Speech Translation (S2ST) system in a client-server mode, in which the enabling components, including ASR (Automatic Speech Recognition), MT (Machine Translation) and TTS (Text to Speech), are implemented on the server. [6]

Client end must install with software that is mainly responsible for the user interface and interaction as well as Some audio preprocessing. This paper is aimed at mobile phone users who can then communicate with other users, irrespective of the other user's device (client) ability to understand the speaker's language and having support to internet use. This paper can find varied applications in different domains like in businesses, voice response systems and teaching system.

Voice translation method mainly divided into 3 issues….
1:- Speech to Text Translation
2:- Text to Text Translation
3:- Text to Speech Translation

2. SPEECH-TO-TEXT CONVERSION

In this method the actual speech to text translation is done by using 'Google API'. This method mainly having following factors that shows us what are the factors need to be considered and how actually this conversion will be takes place.

2.1. TIME

A good typist can type about 300 letters per minute and the average speaking rate is about 150 words per minute (with some variance between the speakers and the languages), even the professional typing rate is certainly not high enough to transfer a stream of spoken words into a readable form in real-time. As a consequence, the speed of typing has to be increased for a sufficient real-time speech-to-text transfer. [5]

2.2 MESSAGE TRANSFER

The goal of speech-to-text transfer is type into text of spoken Words. However, this process will be carried out at the backside (not visible to user). If children are not sufficiently exposed to spoken language, their oral language system may develop slowly and less effectively compared with their peers.

Result to this many people with an early hearing impairment are less used to the grammatical rules applied in oral language as adults and have a less elaborated mental lexicon compared with normal hearing people.[5]

If words are unknown and/or if sentences are too much complex to understand, the written form does not help their understanding. The result for intralingual speech-to-text conversion is that precondition 1, the language proficiency of the audience, also has to be addressed, i.e. the written transcript has to be adapted to the language abilities of the audience - while the speech goes on.

Speech-to-text service not only needs to know their audience, they also have to know which words and phrases can be exchanged by equivalents which are easier to understand, and how grammatical complexity will be reduced.

They need to know techniques of how to make the language in itself more accessible while the information is going to be transferred and preserved.

2.3 CHALLENGES OF SPEECH TO TEXT TRANSLATION

There are some challenges in speech to text translation. The text is written and converted almost simultaneously, and the control of the reading speed shifts at least partly over to the speaker and the speech-to-text converter. The text is not fixed already, instead new words are produced continuously and readers must follow this word production process very closely if they want to use the real-time abilities of speech-to-text transfer. Because of this interaction of writing and reading, the presentation of the written text must be optimally adapted to the words needs of the convertor. The challenges of real-time speech-to-text conversion can now be summarized as follows:

1. To be slow enough in producing written language (source/target).
2. It becomes possible to meet the expectations of the audience with respect to the characteristics of a written text. Word-by-word transfer done by a Google API using meaning from the dictionary.
3. Moreover, a successful real-time presentation must match the hearing abilities of the audience, i.e. the written words must be presented in a speech such way that is optimally recognizable and understandable for the audience.

2.4 MATCHING TECHNIQUES

I. Complete-word matching: - The search engine compares the incoming audio signal against a prerecorded set of the words.

This technique requires much less processing time than sub-word matching, but it requires the user (or someone) prerecord every word that will be recognized - sometimes several hundred thousand words which is so complicated. Complete-word templates also require large amounts of storage memory and are practical only if the recognition vocabulary is known when the application is developed [8].

II. Sub-word matching: - The engine looks for sub-words – usually phonemes and then performs further pattern recognition on those.

This technique requires more processing time than complete-word matching, but it requires much less storage area. In addition, the pronunciation of the word can be guessed from English text without requiring the user to speak the word beforehand. [3] [9]

3. METHODS OF REAL-TIME SPEECH-TO TEXT CONVERSION

There are three methods of realizing real-time speech-to-text transfer: speech recognition, computer assisted note taking (CAN) and communication access (or computer aided) real-time translation (CART).

The methods differ:

1. in their ability to generate exact real-time transcripts.
2. With respect to the conditions under which these methods can be properly applied and
3. With respect to the amount of training which is needed to become a good speech-to text service provider.

3.1 SPEECH RECOGNITION

Automatic speech recognition (ASR) technologies today can correctly recognize and write down more than 90% percent of a long series of spoken words for many languages. However, even this high percentage is not sufficient for speech-to-text a service, since 95%+ correctness is needed to provide a sufficient message transfer [2].

Moreover, even the 90%+ accuracy in automatic speech recognition does not occur by itself. In order to be recognizing, the speaker has to train the speech recognition system in advance with him/is voice and speaking characteristics. Some regional speaking characteristics are generally only poorly recognized, even after extensive training. Physical changes in voice quality (e.g. from a flu) can result in poorer recognition results.

The reason for this is that the speech recognition process is based on a match of physical parameters of the actual speech signal with a representation which was generated on the basis of a general phonetic model of language and the phonetic and voice data from the individual training sessions. If the individual physical parameters differ from those of the training sessions, recognition is less successful. Moreover, if background noise decreases the signal-to-noise ratio, accuracy might go down to below 80 percent.

Speech recognition systems can meet challenge writing speed under good circumstances. In this case, the recognition rate of ASR would in principle be high enough to transfer every spoken word into written text at the back end in real-time. But there are limitations which have to be taken into account. The most restrictive factor is that automatic speech recognition systems are not (yet) capable of recognizing phrase- and sentence boundaries. [3] Therefore, the output from an automatic speech recognition system is a stream of words without any comma or full stop. Moreover, the words would not be assigned to the different speakers.

ASR system today fails as far as challenge 3 is concerned. Although the single words are readable, the output of automatic speech recognition systems is almost not understandable for any reader.

Re-speaking has advantages though. It makes it possible to adapt the spoken language for an audience with limited oral language proficiency. This would not be possible with automatic speech recognition. Real-time speech-to-text conversion with speech recognition systems does not require special technical knowledge or training except for the fact that the SR- system has to be trained. For the user it is sufficient to speak correctly. However, linguistic knowledge and a kind of "thinking with punctuation" are necessary to dictate with punctuation marks. [7]

3.2 COMPUTER-ASSISTED NOTE TAKING (CAN)

With computer-assisted note taking (CAN), a person writes into an ordinary computer what a speaker says. However, as was discussed earlier, even professional writing speed is not sufficient to write down every word of a speech. To enhance writing speed, abbreviation systems are used in computer-assisted note taking which minimize the amount of key strokes per word. The note taking person types abbreviations or a

mixture of abbreviations and long forms. An abbreviation-to-long-form dictionary translates the abbreviations immediately into the corresponding long form. On the screen, every word appears in its long form. [5]

On the other hand, there are much elaborated and well developed systems like e.g. C-Print which has been developed at the National Technical Institute for the DEAF at Rochester Institute of Technology (RIT 2005). This system uses phonetic rules to minimize the key strokes for every word. After a period of training with the system, the captions is able to write with a higher speed. This allows for a high quality message transfer. However, the writing speed is still limited so that word-for-word transcripts are rather unusual, even with C-Print. With CAN-systems like C-Print, a message-to-message rather than a word-for-word transfer is produced.

The efficiency of CAN systems is mainly determined by the quality of the dictionary which translates the short forms into the corresponding long forms. Better the dictionary, the higher the typing speed potential.

Individually made dictionaries are mostly a collection of abbreviations like 'hv' for 'have' and 'hvt' for 'have to' etc. However, this kind of dictionary is limited insofar as the user has to know every abbreviation. Consequently, the amount of time which is needed for people to learn and to prevent them from forgetting the abbreviations once learned increases with the increase in the size of the dictionary. [5]

Elaborated systems like C-Print use rule-based short-to-long translations. Here, the captions has to learn the rules of transcription. One rule could be that only consonants but not vowels are written down. The resulting ambiguities (e.g. 'hs' for 'house' and 'his') have to be resolved by a second rule. However, orthographic transcription rules turned out to be rather complicated – at least in English. Therefore, systems
like C-Print are often based on a set of rules which are in turn based on a phonetic transcription of the spoken words. On the basis of a set of shortening rules, the note taking person does not write certain graphemes but phonemes of the spoken words.

4. TEXT TO TEXT TRANSLATIONS

In text to text conversion method, the text is going to be change from one language to target language. This method will be implemented by using Google. As we have internet connection on our mobile phone, the text translation will be done using an internet. So for that we need to select proper target language in which we want to modify source language. This method gives us language converted text representation. The output of this method will be given to next method i.e. text to speech translation.

4.1 ELOQUENCE COMMAND INTERFACE (ECI)

The Eloquence Command Interface (ECI) is a proprietary, platform independent API that allows direct access to all the functionality and power of the IBM Text-to-Speech. This API:

• Is supported on a variety of operating systems.

• Allows customization of speech output both through function calls and textual annotations.

• Does not use the Windows Registry to find components, allowing developers to include a private copy of the text-to-speech engine with their application that is less likely to be accidentally modified by later installations or by other applications.

See the sections The ECI Application Programming Interface and ECI Reference for details on how to use this API. See the section Annotations for details on the use of ECI annotations to customize speech output. [7]

CONCLUSION:

Real-time speech-to-speech transfer is a powerful tool which provides people with a hearing impairment access to oral communication. However, elaborated dictionaries as they are needed for efficient CAN- or CART-systems are not yet developed for many languages.

Without those dictionaries, the systems cannot be used. Linguistic research has to find easy but efficient strategies for the real-time adaptation of the wording in order to make talking understandable also for an audience with limited language proficiency.

Finally, the optimized speech translation method is done in this way.

REFERENCES

1. Schlenker –Schulte, 1991; Perfetti et al. 2000 with respect to reading skills among deaf readers (Stinson et al.1999: Accuracy). Leitch et al.2002.

2. Zaidi Razak, Noor Jamaliah Ibrahim, Emran Mohd Tamil, Mohd Yamani Idna Idris "Quarnic Verse Recitation Feature Extraction Using Mel-Frequency Cepstral Coefficient(Mfcc)" Department Of Al-Quran & Al-Hadith, Academy Of Islamic Studies, University Of Malaya.

3. D. R. Reddy, An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave, Tech. Report No.C549, Computer Science Dept., Stanford Univ., September 1966.

4. Cremer, Inge (1996): "Prüfungstexte verstehbar gestalten", Hörgeschädigtenpädagogik 4, 50 Jahrgang, Sonderdruck.

5. Susanne Wagner - Intralingual speech-to-text-conversion in real-time: Challenges and Opportunities.

6.B Zhou, Y Gao, J Sorensen, et al. , "A hand held speech to speech translation system", Automatic Speech Recognition and Understanding, 2003.

7.IBM (2010) online IBM Research Source: - http://www.research.ibm.com/Viewed 12 Jan 2010.

8. S.katagiri, Speech Pattern recognition using neural networks.

9. L.R.Rabiner and B.H.jaung, Fundamentals of Speech Recognition Prentice-Hall, Englewood Cliff, New Jersey 1993.

Authors

AKASH NAYAK
Pursuing B.E. (Computer Engineering)
University of Pune, Dnyanganga College of Engineering & Research, Survey No.39, Narhe, Pune 411041, India.

SANTOSH KHULE
Pursuing B.E. (Computer Engineering)
University of Pune, Dnyanganga College of Engineering & Research, Survey No.39, Narhe, Pune 411041, India.

ANAND MORE
Pursuing B.E. (Computer Engineering)
University of Pune, Dnyanganga College of Engineering & Research, Survey No.39, Narhe, Pune 411041, India.

AVINASH YALGONDE
Pursuing B.E. (Computer Engineering)
University of Pune, Dnyanganga College of Engineering & Research, Survey No.39, Narhe, Pune 411041, India.

Dr. RAJESH PRASAD
Professor, ZES's Dnyanganga College of Engineering & Research, Narhe, Pune