Speech-to-Speech Translation: A Review

Mahak Dureja
Department of CSE
The NorthCap University, Gurgaon

Sumanlata Gautam
Department of CSE
The NorthCap University, Gurgaon

ABSTRACT

This paper reviews the technology used in Speech-to-Speech Translation that is the phrases spoken in one language are immediately spoken in another language by the device. Speech-to-Speech Translation is a three step software process which includes Automatic speech Recognition, Machine Translation and voice synthesis.

This paper includes the major speech translation projects using different approaches for speech recognition, translation and text to speech synthesis highlighting the major pros and cons for the approach being used.

Keywords

Automatic Speech Recognition (ASR), Machine Translations (MT), Text-To-Speech synthesis (TTS).

1. INTRODUCTION

Speech translation is a process that takes the conversational speech phrase in one language as an input and translated speech phrases in another language as the output. The three components of Speech-to-Speech Translation are connected in a sequential order. ASR is responsible for converting the spoken phrases of source language to the text in the same language followed by machine translation which translates the source language next to target language text and finally the speech synthesizer is responsible for text to speech conversion of target language.

The following sections include the tools and methodologies of two projects: IBM's MASTOR and Verbmobil.

1.1 Basic model for Speech-to-Speech Translation

The high-tech in Speech-to-Speech Translation system that enable such multi-lingual communications is supporting a pipelined architecture of automatic speech recognition, machine translation system and speech synthesis or text-to-speech which primarily relies on lexical information and ignoring the other rich information which is present in speech and spoken discourse such as noise and human utterances. There is negligible interaction between the basic components ahead of the pipeline and also there is no involvement of humans in the loop for automatic learning, adapting and collectively managing the interaction. To overcome these elemental precincts requires mounting robustness at all the stages of the pipeline.

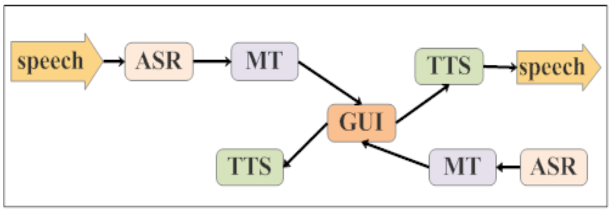

Fig1: Overall Speech-to-Speech translation system

The basis for the system is that it should utilize the rich context that is away from the dictated words, while being aware of and working with the different cultured humans for improving information transfer, communication efficiency and social co-presence for enabling successful multi-lingual interactions. In this paper different projects are capturing that model and transferring highly related information that is conveyed for the speech prosody, discourse and the user state behaviour to support robust translation and significant synthesis of the target language [6].

2. IBM's MASTOR

The IBM MASTOR shorthand for Multilingual Automatic Speech-to-Speech Translator is developed for the DARPA CAST and its mission is to develop technologies that facilitate rapid deployment of real-time Speech-to-Speech Translation of low-resource languages on mobile devices [1].

The general structure of MASTOR system has the components of ASR, MT and TTS. This pipelining approach allows system for the deployment of the existing speech and language handing out techniques, while taking care of unique problems in Speech-to-Speech Translation.

2.1 Methodology used for automatic speech recognition

Grapheme based acoustic models are used to overcome the problem of absence of short vowels [7]. Grapheme based acoustic model lead to unambiguous pronunciation of lexicons and hence facilitates the model training and decoding. Also, depending on its context the same grapheme may yield different phonetic sound and lead to less accurate acoustic models. For this reason two different approaches come into existence. The first one is to use short vowels known as full phonetic approach and the second one uses the context-sensitive graphemes in which two different phonemes are generated for the letter "A" (Alif) depending on its position in the word. The IBM ViaVoice product engine is a highly robust and efficient framework which is used for acoustic modelling by using rank based acoustic scores that are derived from tree-clustered context reliant Gaussian Models for both the desktop systems and hand-held systems.

Another concept for high performance ASR for open ended communication system is Language Modelling (LM) which is the probability of various word sequences. A huge amount of corpus data is needed for N-gram LM's for the representation of domain word usage distribution. Three approaches are used for generating LM's, first one is to obtain additional training material automatically second is to interpolate domain-specific LM's with other LMs and the third approach is to improve the distribution estimation for the robustness and accuracy with limited resources.

2.2 Translation methodology used

In NLU/NLG-based Speech Translation for statistical machine translation methods a sentence T in one language that is translated into a sentence A which is in the another language by using statistical model which is estimating the conditional probability of A when T is given, i.e. P(A|T). Typically, P(A|T) is optimized on a set of two sentences that are translations of each another. Let C denotes the source language concepts and S denotes the target language concepts, then the statistical concept-based algorithm is going to select a word sequence A' as

$$A' = \arg\max P(A|T)$$

$$= \arg\max \{\sum\sum P(C|\ T)\ P(S|\ C,T)\ P(A|\ S,C,T)$$

Where the conditional probabilities P(C|W) is estimated by Natural Language Understanding (NLU), P(S| C,T) is estimated by Natural Concept Generation (NCG) and P(A| S,C,T) is estimated by the Natural Word Generation (NWG) procedures, A decision-tree based statistical semantic parser is estimating P(C|T) and P(S| C,T) and P(A| S,C,T) are estimated by the maximization function for conditional entropy [5].

2.3 Text-to-Speech Synthesizer

A text-to-speech engine synthesizes the translated utterance generated from the NLG module. As there are limited resources available for a mobile device such as memory space battery backup etc., the IBM's TTS system is based on IBM's formant TTS technology. An unlimited number of voice can be synthesized using the formant based TTS system which allows elastic customizations by modifying choral characteristics of a speech such as gender, pitch, volume and speed. Unlimited vocabularies are supported by this system. Also, the TTS software can pronounce any text which is given to it. More importantly, formant based TTS system has a small footprint that requires less memory (only about 3 MB for every language), which is appropriate for deployment of the software in embedded applications.

3. VERBMOBIL

Verbmobil is a two way Speech-to-Speech Translation system which does not depend on the speaker. It is used for translation of spontaneous dialogs in mobile situations. It firstly identifies the input and further analyses and translates it, and finally delivers the final translation. This is a multilingual system which handles dialogs delivery in three-business-oriented domains where the translation depends on the context between three languages (German, English and Japanese) [2].

This system deals with the spontaneous dialogs. In this case it doesn't mean just continuous speech like in the current dictation systems, but here rational disfluencies and repairing phenomena such as changing mid word, ums and arr, and some short words that are accidently left out in rapid speech are also included in the speech. For example, Verbmobil corpus has the chance that 20% of all dialog turns having at least one auto-correction and 3% also include false starts. A combined approach for deep and shallow analysis methods is used by this system to find out the slips in the speech and then translate it in accordance to what the person tried to say rather than what was actually said by him.

3.1 Methodology Used For Speech Recognition

The format used for the Verbmobil speech corpora writes out fifteen strata of interpretation: these are two transliteration variants, lexical orthography, canonical form of pronunciation, manual phonological segmentation, phonological based automatic segmentation, segmentation of word, prosodic segmentation, dialog acts, noise, superimposed speech, syntactic grouping, word type, syntactic function, and the prosodic boundaries [2]. For the monolingual data, the multi-language Verbmobil corpus incorporates two way dialogs delivery for in person dialogs with human interpreters, or the dialogs are interpreted by the various versions of Verbmobil that are aligned to bilingual transliterations. Three tree-banks for German, English and Japanese have been developed with interpretation on three stratums: phrase structure, morpho-syntax and predicate-argument structure, 3 tree-banks are created for English, Japanese and German to train neural nets, Hidden Markov Models, parsers, probabilistic automata, translation methods, rule-based systems, and plan recognizers different machine learning methods were used. The continuous evaluation for coverage, accuracy and robustness of a speech translation system for dialogs which aren't structures correctly majorly depends on the quality and quantity of the training corpora.

3.2 Methodology used for speech translation

The Verbmobil is using a multiengine approach in its translation module. Five synchronized translation engines are used by it: case-based translation approach, substring-based translation approach, dialog-act based translation approach, sematic transfer approach and statistical translation approach [3].

The statistical translation module starts with the hypothesis of the single best sentence of speech recognizer and other prosodic information about phrase margins and sentence form are utilized by its statistical translation module. This module yields a string of words in the target language along with a confidence measure used by the selection module for the final choices of the translation output. It includes two components for translation based on the case. (1) This method of substring-based translation is used for incremental synchronous interpretation. This method is primarily based on machine learning methods that are applied to a sentence-aligned bilingual corpus. The basic processing units of the system are those substrings whose contiguous piece of translation can be found in the text corpus. In the incremental translation algorithm for the speech sequence of input segments, we use the methods of combining the pairs of substring with patterns for word order switching and word cluster information. (2) Sentence-aligned corpus is another component for case-based translation which is based on 30000 translation templates. In a dialog-act based translation, the main propositional content of an utterance has been extracted by 19 dialog acts and a cascade of around 300 finite-state transducers. The statistical dialog classifier is dependent on N-grams and the process is self-learning that takes previous dialogs into account. The topic, propositional content and recognized dialog act are represented by a simple frame register including 49 nested objects with 95 possible attributes for covering the appointment scheduling and planning travel tasks. To transform the inter-lingual terms into the analogous target language a template based approach is used. The shallow inter-lingual representations of an utterance is stored along with a deep semantic representation encoded in the dialog memory as well as focus information and topic for further processing by the context and dialog evaluation component.

3.3 Methodology Used For Speech Synthesizer

The speech synthesizer for German and American English follows a parallel approach based on large corpus of annotate speech data. The basic unit of concatenation is a Word, so that if a word is not found in the database then only sub-word units are used [2]. A graph based unit selection procedure is applied to the synthesizer and the best existing synthesis segments matching the prosodic and segmental constraints of the input and the synthesizer exploit the syntactic, discourse and prosodic information that are provided by the previous processing stages. It provides concept-to-speech synthesis and for the deep processing stream, whereas it operates more likely to traditional text-to-speech system that results in a lower quality of the output for the shallow translation threads.

4. CONCLUSION

In this paper methods and technologies used in building two independent software; one is IBM MASTOR and the other Verbmobil has been studied. Both have one common thing that these are the Speech-to-Speech Translation systems but different in many aspects such as languages, robustness of the system and technologies used for the pipeline structure of Speech-to-Speech Translation system. Through this paper we tried to check the insights of two systems and for the future aspect we can study some more software to design an efficient, accurate and robust translator using integrating approaches for translation system.

TABLE 1. Comparison for Ibm's Mastor and Verbmobil

VARIOUS DOMAINS	IBM MASTOR	VERBMOBIL
LANGUAGES	ENGLISH-MANDARIN CHINESE AND ENGLISH-ARABIC	GERMAN-ENGLISH-JAPANESE
METHODOLOGY FOR SPEECH RECOGNITION	IBM'S VIAVOICE FOR ACOUSTIC MODELLING	TRANSCRIBES FIFTEEN STRATA OF ANNOTATIONS
METHODOLOGY FOR TRANSLATION	NLU/NLG BASED SPEECH TRANSLATION	USES FIVE SYNCHRONISED TRANSLATION ENGINE APPROACH
METHODOLOGY FOR SYNTHESIS	IBM'S FORMANT TTS TECHNOLOGY	CONCATENATIVE APPROACH WITH A GRAPH-BASED UNIT SELECTION PROCEDURE
ERROR RATE (FOR ENGLISH LANGUAGE RECOGNITION)	11.06%	7.3%
ADVANTAGES	SELF DESIGNED TOOLS SPEAKER INDEPENDENT LEAD TO RAPID DEVELOPMENT OF SYSTEMS	COMBINATIONAL APPROACHES USED HIGHLY ROBUST REAL TIME SYSTEM

5. REFERENCES

[1] Gao, Yuqing, et al. "IBM MASTOR: Multilingual automatic speech-to-speech translator." *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. Vol. 5. IEEE, 2006.

[2] Wahlster, Wolfgang. "Mobile Speech-to-Speech Translation of spontaneous dialogs: an overview of the final Verbmobil system." *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer Berlin Heidelberg, 2000. 3-21.

[3] Zhang, Ying. "Survey of current speech translation research." *Found on Web: http://projectile. is. cs. cmu. edu/research/public/tal ks/speechTranslation /sst-survey-joy. pdf* (2003).

[4] Boitet, Christian, et al. "Evolution of MT with the Web." *Proceedings of the Conference" Machine Translation 25 Years On*. 2009.

[5] Gu, Liang, et al. "Concept-based Speech-to-Speech Translation using maximum entropy models for statistical natural concept generation." *Audio, Speech, and Language Processing, IEEE Transactions on* 14.2 (2006): 377-392.

[6] Chelba, Ciprian, et al. "Large scale language modeling in automatic speech recognition." *arXiv preprint arXiv:1210.8440* (2012).

[7] Narayanan, Shrikanth, et al. "Speech recognition engineering issues in speech to speech translation system design for low resource languages and domains."*Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. Vol. 5. IEEE, 2006.

[8] Yun, Seung, Young-Jik Lee, and Sang-Hun Kim. "Multilingual Speech-to-Speech Translation system for mobile consumer devices." *Consumer Electronics, IEEE Transactions on* 60.3 (2014): 508-516.