Speech To Speech Machine Translation System from Bengali To English

Introduction:

In this knowledge age the various sources of knowledge are available in multiple languages and we cannot consume all these knowledge as we are hindered by our limited linguistic knowledge. Machine Translation bridges this linguistic barrier. Direct verbal communication between two persons is also severely hampered if they do not know a common language. Speech to speech machine translation is the solution to this spoken language barrier. Speech to speech machine translation has of late been very popular in Europe due to its linguistic diversity. These days even smartphones have started providing speech to speech machine translation applications.

In the following sections we discuss related works on speech to speech translation. Then we discuss about the three components of speech to speech translation - automatic speech recognition, text-to-text machine translation and text-to-speech generation or speech synthesis. Then we discuss about the tools which are used in speech to speech translation. In the subsequent section, we discuss about different datasets that we require for speech to speech translation applicable for Indian languages.

Related Work:

Spoken Language Translator:

In 1992 a speech to speech translation project named Spoken Language Translator was launched. The Spoken Language Translator was based on a combination of state-of-art unification-based natural language processing technology developed by SRI International in Cambridge, England, and state-of-art Hidden Markov Model (HMM) – based speech recognition technology developed by SRI in Menlo Park, California. In computational linguistics, these include grammar specialization through explanation based learning and discriminant based statistical preferences for language processing. In speech recognition, the most important result is the development of discrete-mixture HMMs, leading to two to three times faster recognition than with the state-of-art continuous HMMs having similar recognition accuracy.

Verb mobil:

In 1993 “Verb mobil” was the project in the speech to speech translation involving European Languages. Verb mobil is a speaker-independent and bidirectional speech-to-speech translation system for spontaneous dialogs in mobile situations. It recognizes spoken input, analyses and translates it, and finally utters the translation.
The multilingual system handles dialogs in three business-oriented domains, with bidirectional translation between three languages (German, English, and Japanese).

Spectra:

AT&T Research Labs is developing another speech to speech translation system called “Spectra”. We are currently working on Spectra for automatic speech recognition in Hindi and Bengali. Spectra comprises of an HMM-based large vocabulary continuous speech recognition system (AT&T Watson Speech Recognizer), a phrase-based translation system and a unit selection text-to-speech synthesis system (AT&T Natural Voices TTS). Spectra is endowed with automatic language identification capabilities and can currently translate from/to English and six other languages (Chinese, French, German, Italian, Japanese and Spanish).

Motivation:

Very few works on speech recognition and synthesis have been carried out in Indian languages. In recent times, research and development of machine translation systems involving Indian languages are gradually picking up. No research and development activities in speech to speech machine translation systems involving Indian languages have been reported. Bengali is the second most popular language in India. It is placed in the seventh position in the world based on the number of native speakers. Bengali is the national languages of Bangladesh. The proposed methodologies will be used for developing the Bengali to English speech to speech machine translation system. The knowledge gained may be applied for developing similar systems involving other Indian languages.

The components of a Speech to Speech Translation system:

Speech to speech translation has three main tasks involved in it. The first one is automatic speech recognition (ASR) which recognizes the user’s speech input and converts it into source language text. The next one is text to text machine translation (MT) which translates the source language text into target language text. The final task is text to speech synthesis (TTS) which converts target language text to target language speech output. Each of these three components is an important research area in its own merit.

Automatic Speech Recognition (ASR):

There are two kinds of ASR models - context independent ASR model and context dependent ASR model. For context independent ASR model, the target phone does not depend on the previous or next phone, but in case of context dependent model, the target phone depends on the previous and next phones. ASR models are typically built on a big speech corpus of around 20–50 hours with its transcription in the concerned language. It also needs a pronunciation dictionary which contains the grapheme to phoneme model for the particular language. ASR also requires a large language model for modeling the target language. The phonemes are clustered into
classes like velar, vowels, consonants, sibilants, approximants, aspirated, un-aspirated, monophthong, diphthong, palatal, dental, labial, etc. This classification should be done carefully incorporating linguistic knowledge. We also need a larger language model in Bengali language to create a larger pronunciation model (Grapheme to Phoneme Model) to produce more accurate text output.

Machine Translation (MT):

There are two main approaches to MT – rule-based (or linguistics-based) MT and data driven MT. With the availability of huge amount of machine readable parallel training data and cheap and fast computation power, data driven approaches to MT have become the de facto standard. Data driven approaches to MT can again be broadly classified into two groups – example-based MT (EBMT) and statistical MT (SMT). SMT has taken the centre stage in MT research over the last decade. In SMT, the most dominant approaches are phrase-based SMT (PB-SMT) and hierarchical (or syntax-augmented) SMT. For the speech to speech translation project we will use the PB-SMT approach. We will use the standard log-linear PB-SMT model for this purpose with IBM model 4 for word alignment, phrase-extraction heuristics of (Koehn et al., 2003), minimum-error-rate training (Och, 2003) on a held-out development set, target language model with Kneser-Ney smoothing (Kneser and Ney, 1995), and Moses decoder (Koehn et al., 2007).

We have already developed a SMT based machine translation system for Language pair English to Bengali using Moses and Giza ++ and SRILMT tools. In a similar way we can develop a Bengali to English SMT system.

Text To Speech Synthesis (TTS):

For Text To Speech synthesis we need a good quality recording of an expert voice for around 20 – 50 hours. It is better to approach a news reader or a person in drama for this task. We will also apply unit selection method in this TTS system. Unit selection synthesis uses large databases of recorded speech. During database creation, each recorded utterance is segmented into some or all of the following: individual phones, diphones, half-phones, syllables, morphemes, words, phrases and sentences. Typically, the division into segments is performed using a specially modified speech recognizer set to a “forced alignment” mode with some manual correction afterward, using visual representations such as the waveform and spectrogram. An index of the units in the speech database is then created based on the segmentation and acoustic parameters like the fundamental frequency (pitch), duration, position in the syllable and neighbouring phones. At run time, the desired target utterance is created by determining the best chain of candidate units from the database (unit selection). This process is typically achieved using a specially weighted decision tree.

We will build up a English Text To Speech Synthesis system using tools which are popular in the market.
Tools required for Speech to speech translation:

For ASR:

For ASR, there are several tools available like HTK, Sphinx, Kaldi, etc. HTK is a toolkit for building Hidden Markov Models (HMMs). HMMs can be used to model any time series and the core of HTK is similarly general-purpose. However, HTK is primarily designed for building HMM-based speech processing tools, in particular recognizers. Thus, much of the infrastructure support in HTK is dedicated to this task. There are two major processing stages involved. Firstly, the HTK training tools are used to estimate the parameters of a set of HMMs using training utterances and their associated transcriptions. Secondly, unknown utterances are transcribed using the HTK recognition tools.

For MT:

The field of MT research has witnessed dramatic growth in the last decade. Many open-source tools have come up providing support for different paradigms and architecture. Statistical machine translation (SMT) is the state-of-the-art today in MT. Tools are available for most successful approaches to MT, viz. Phrase-based SMT (PB-SMT) and hierarchical phrase-based SMT (HPB-SMT). Moses (Koehn et al., 2007) has become the most popular SMT toolkit in the research community over the years. RWTH Aachen University’s Jane (Vilar et al., 2010; Wuebker et al., 2012) open-source SMT toolkit supports techniques for both PB-SMT and HPB-SMT. Ncode (Crego et al., 2011) is an open-source SMT toolkit for translation models estimated as n-gram language models of bilingual units (tuples). These SMT toolkits in turn use many other tools – GIZA++ (Och and Ney, 2003), Berkeley Word Aligner for word alignment; SRILM, KenLM, IRSTLM for language modelling; MERT (Zaidan, 2009) for minimum error rate training (Och, 2003), OpenFst library for weighted finite-state transducers (FSTs), etc.

For TTS:

For Speech Synthesis we need tools like HTS, Festival etc. HTS is HMM based Speech Synthesis tool. HMM-based approaches to speech synthesis can be categorized as follows:

1. Transcription and segmentation of speech database.
2. Construction of inventory of speech segments.
3. Run-time selection of multiple instances of speech segments.
4. Speech synthesis from HMMs themselves.

Another important tool is Festival. Festival offers a general frame – work for building speech synthesis systems as well as including examples of various modules. As a whole it offers full text to speech through a number of APIs: from shell level though
a Scheme Command interpreter as a C++ library and as Emacs interface. The system is written in C++ and uses the Edinburgh Speech Tools for low level architecture and has a Scheme (SIOD) based command interpreter for control.

Datasets:

Since almost all the components involved in speech to speech translation are reliant on statistical methods, we need to have good amount of data for each of them and the bigger the better. For Bengali To English speech to speech translation, we intend to use the SHRUTI Bangla speech corpus developed by Society for Natural Language Technology and Research (SNLTR), in association with IIT Kharagpur. It consists of about 22 hours’ speech data in Bangla. For MT, we will use the TDIL parallel corpus in tourism and health domain, developed in the ANUVADAKSH project (English to Indian Language Machine Translation project Phase I & II). The corpus contains around 41 thousand parallel sentences. For language modeling purpose we will use a monolingual corpus comprising of around six lakh words collected from the tourism domain.

References:


