**Notes of “Speech Translation - Wikipedia”**

**Type**

Web Article

**Topic**

Concept

**Notes**

Broken down into 3 steps, Automatic Speech Recognition (ASR), Machine Translation (MT) and voice synthesis (TTS)

Speech -> Text -> Translated Text -> Speech

Idfferent from “phrase translation”

**Usable Quotes or Figures**

Most recently, the field has benefited from advances in deep learning and big data.

Systems that do not use training are called "speaker independent"systems

Neural networks and other AI algorithms such as Hidden Markov models are also being incorparted, improving the performace of the tool.

Speech Recognition

**Notes**

Some speech recognition systems require "training"

**Limitations**

**High levels of stress and noise both affect the results**

**Error rates increase as vocabulary list grows in size**

**Vocabulary is hard to recognize if it contains confusable words**

e.g. the 26 letters of the English alphabet are difficult to discriminate because they are confusable words (most notoriously, the E-set: "B, C, D, E, G, P, T, V, Z"); an 8% error rate is considered good for this vocabulary

**Speaker dependence vs. independence**

A speaker-dependent system is intended for use by a single speaker.  
A speaker-independent system is intended for use by any speaker (more difficult)

**Isolated, Discontinuous or continuous speech**

With isolated speech, single words are used, therefore it becomes easier to recognize the speech.  
With discontinuous speech full sentences separated by silence are used, therefore it becomes easier to recognize the speech as well as with isolated speech.  
With continuous speech naturally spoken sentences are used, therefore it becomes harder to recognize the speech, different from both isolated and discontinuous speech

**Task and language constraints**

e.g. Querying application may dismiss the hypothesis "The apple is red."  
e.g. Constraints may be semantic; rejecting "The apple is angry."  
e.g. Syntactic; rejecting "Red is apple the."  
Constraints are often represented by a grammar.

**Read vs. Spontaneous Speech**

When a person reads it's usually in a context that has been previously prepared, but when a person uses spontaneous speech, it is difficult to recognize the speech because of the disfluencies (like "uh" and "um", false starts, incomplete sentences, stuttering, coughing, and laughter) and limited vocabulary.

**Adverse conditions**

Environmental noise (e.g. Noise in a car or a factory)  
Acoustical distortions (e.g. echoes, room acoustics)  
Speech recognition is a multi-levelled pattern recognition task.

**Acoustical signals are structured into a hierarchy of units;**

e.g. Phonemes, Words, Phrases, and Sentences;

Translation

**Notes**

The human translation process may be described as:

Decoding the meaning of the source text; and

Re-encoding this meaning in the target language.

**Types**

Dictionary-based – Uses a dictionary style translation where each word is translation by finding the word with the identical meaning in the other language. Requires a large corpus of information. Loses context fast.

Rule based-based – The rule-based machine translation paradigm includes transfer-based machine translation, interlingual machine translation and dictionary-based machine translation paradigms. Unlike other methods, RBMT involves more information about the linguistics of the source and target languages, using the morphological and syntactic rules and semantic analysis of both languages.

Interlingual – Takes the text of one language and translates into a neutral language known as “interlingual”. The translation is then output from the interlingua, a library of past known translations.

Statisical – Uses statiscical methods based upon bilingual text corpuses. This is limited by the number of langauge pairs that share bilingual corpuses. Many translation softwares use patterns in a large database of data. From those patterns, translations can be made using stastics which uses the phrase with the highest likelihood.

Neural MT – The method which has received the highest amount of development recently. Google now uses this method instead of their standard statistical method. This is favoured for it’s ability to retain the context of a sentence for its value in providing a successful translation.

**Major Issues**

**Word-sense disambiguation** is when a sentence can have more then one meaning. There are numerous ways of attempting to solve it. These are seperated into two types of approaches named shallow and deep. Shallow approaches assume no knowledge of the text. They simply apply statistical methods to the words surrounding the ambiguous word. Deep approaches presume a comprehensive knowledge of the word. So far, shallow approaches have been more successful.

**Non Standard Speech** such as slang and structures such as euphamisms, metaphoresm, similies atc.

Speech Synthesis

Notes

Speech synthesis is the artificial production of human speech.

A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech.Re-encoding this meaning in the target language.

Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database.

The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood clearly.

A text-to-speech system (or "engine") is composed of two parts: a front-end and a back-end. The front-end has two major tasks. First, it converts raw text containing capitalised words and symbols like numbers and abbreviations or into the equivalent of written-out words known as normalisation. The front-end then assigns phonetic transcriptions to each word, and divides and marks the text into prosodic units, like phrases, clauses, and sentences.  The back-end—often referred to as the synthesizer—then converts the symbolic linguistic representation into sound. In certain systems, this part includes the computation of the target prosody (pitch contour, phoneme durations),which is then imposed on the output speech.

Speech synthesis systems use two basic approaches to determine the pronunciation of a word based on its spelling, a process which is often called text-to-phoneme or grapheme-to-phoneme conversion (phoneme is the term used by linguists to describe distinctive sounds in a language). The simplest approach to text-to-phoneme conversion is the dictionary-based approach, where a large dictionary containing all the words of a language and their correct pronunciations is stored by the program. Determining the correct pronunciation of each word is a matter of looking up each word in the dictionary and replacing the spelling with the pronunciation specified in the dictionary. The other approach is rule-based, in which pronunciation rules are applied to words to determine their pronunciations based on their spellings. This is similar to the "sounding out", or synthetic phonics, approach to learning reading.

Each approach has advantages and drawbacks. The dictionary-based approach is quick and accurate, but completely fails if it is given a word which is not in its dictionary. As dictionary size grows, so too does the memory space requirements of the synthesis system. On the other hand, the rule-based approach works on any input, but the complexity of the rules grows substantially as the system takes into account irregular spellings or pronunciations. (Consider that the word "of" is very common in English, yet is the only word in which the letter "f" is pronounced [v].) As a result, nearly all speech synthesis systems use a combination of these approaches.