

Roles for Speech And Language Technology in The Information Society

Torbjørn Svendsen
Department of Telecommunications
Norwegian University of Science and Technology
Trondheim, Norway

Introduction

I assume that nearly every reader of this paper uses speech technology every day. Every time you use a GSM phone (or call anyone on their GSM phone), speech coding is used to compress the speech signal. Because of this, the bit rate required for the speech transmission is reduced by a factor of approximately five compared to an ordinary telephone call. Also, language technology in terms of spelling and grammar checkers are in common use, although there might be differing views on their quality and reliability. Yet, speech and language technology can be much more.

The electronic infrastructure of the Information Society provides the citizen with access to enormous amounts of information. The problem of navigating through the information jungle in order to retrieve some required information is one of the problems each of us face frequently. The challenge of facilitating information access for all citizens is partly reliant upon the design of user interfaces that are intuitive and easy to use. Speech and language technologies can provide some of the necessary components for the accessible Information Society.

Intelligent applications will be facing us in our everyday life. Ambient intelligence is a term that is used to denote the ever-present ICT applications that will surround the citizen of the Information Society. Will we in the near future carry out a conversation with the refrigerator to decide on today's dinner menu? Or discuss the correct washing cycle with the washing machine?

In this paper I will address some of the possible roles for speech and language technologies in the Information Society. But first it is necessary to give a brief introduction to some of the core technologies.

The technologies

Speech and language technology comprises technologies and methods for computer processing of spoken and written language. As the spoken and written language tends to differ (at least to a degree) in terms of grammar, structure and vocabulary, a distinction is often made between technologies for speech and text. Language technology as a term covers both spoken and written language.

Speech synthesis is computer speech. There are two basic methods for obtaining computer generated speech: phrase concatenation and text-to-speech synthesis (TTS). In phrase concatenation pre-recorded sentence fragments and words are concatenated to create spoken messages. The speech quality can be good, but the message vocabulary (i.e. the number of messages the system can generate) is limited. TTS can in principle generate speech from unrestricted written text, and is what most would mean by the term speech synthesis. TTS includes a linguistic analysis of the text in order to transform it to symbolic sequence of sounds, annotated with prosodic information, and the transformation of the annotated symbol sequence to the acoustic waveform that is speech. Older TTS systems employed a model of the human speech production apparatus to produce the synthetic speech, and although this gave considerable freedom in controlling the voice quality, the resulting speech could be rather machine-like. Current systems use concatenation of (short) speech segments, typically diphones, for the waveform generation. The speech segments is produced from pre-recorded, natural speech, resulting in much more natural sounding speech. Still, the synthesized speech is easily distinguishable from natural speech.

Speech recognition is the transformation from the speech waveform to text (or some other symbolic representation in the computer). In Command and Control applications of speech recognition, the recognized speech pattern is associated with an action controlled by the computer. In other applications such as dictation systems, the purpose is to transform the speech into written words. In yet other applications, the intention is not a verbatim transcription of the speech, but an interpretation of the speaker's intent. The performance of speech recognizers is dependent on the vocabulary size, the speaking style (e.g. isolated utterances vs. continuous speech, dictation type speech vs. spontaneous conversational utterances), and on whether a particular user's speech patterns are known to the recognizer or not. The large degree of variability in the speech signal makes speech recognition a very demanding task. The current technology is based on extensive use of statistical modelling

techniques, both of the acoustic realization that is the speech waveform, and of the linguistic structure of the language. The performance of large vocabulary speech recognizers are far from perfect today, but the technology is sufficiently developed to achieve a reasonable performance, particularly in limited domains.

In a *Spoken dialogue system* the human-machine communication is totally speech based. The user speaks his/hers requests and computer responds and prompts for information using synthetic speech. Speech recognition and speech synthesis are vital elements of the dialogue system, which also needs an “intelligent” dialogue manager and natural language processing to extract the meaning from the user’s input. Below is an example of a spoken dialogue system that we have developed in Trondheim. Any user requiring information about bus traffic in the city can call the computerized system. Through a spoken dialogue between the user and the machine, it is possible to get information on how (and when) to travel by bus between any two bus stops in the city.

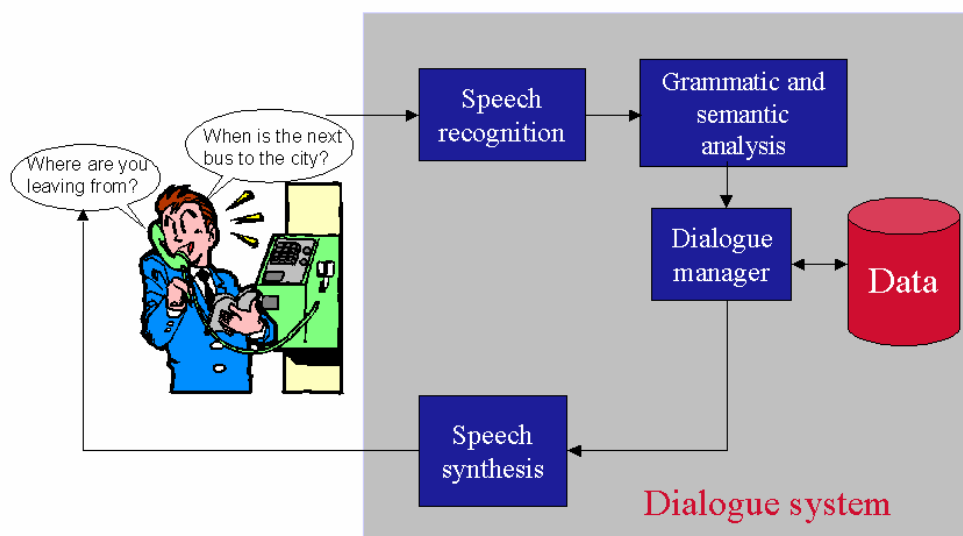


Figure 1. A spoken dialogue system for bus traffic information retrieval

Speaker recognition concerns the task of automatically determining a speaker's identity from an analysis of the speech patterns of an individual. It is common to distinguish between speaker *identification* and speaker *verification*. The former aims at determining the identity of a speaker from a set of known speakers (including the possibility as labeling the speaker as "unknown"). The latter is the simpler task of verifying a claimed identity, e.g. the user speaks (or uses a keyboard to enter) an identity claim, and is prompted to speak an utterance that is used to accept or reject the identity claim.

A text in a foreign language can be automatically translated to a familiar language using *Machine translation*. Free translation services can be accessed on the web, e.g. Alta Vista's *Babel Fish*. It is an open question whether perfect machine translation of general text is theoretically possible, but at least in limited task domains performance can be quite good. Often, the need is not necessarily for a perfect translation, but to facilitate the extraction of the most interesting information in the source document.

A common method for demonstrating the quality of machine translation is by using an MT system to first translate a source text to another language, and then let the system translate the result back to the source language. An example is shown in fig. 2, where we have used the *WorldLingo* machine translation system for the demonstration. The source text is taken at random from the *WorldLingo* web page. As can be seen, the translation is far from perfect, and the syntax is strongly flavored by the intermediate language. Yet, much of the content of the original text is preserved.

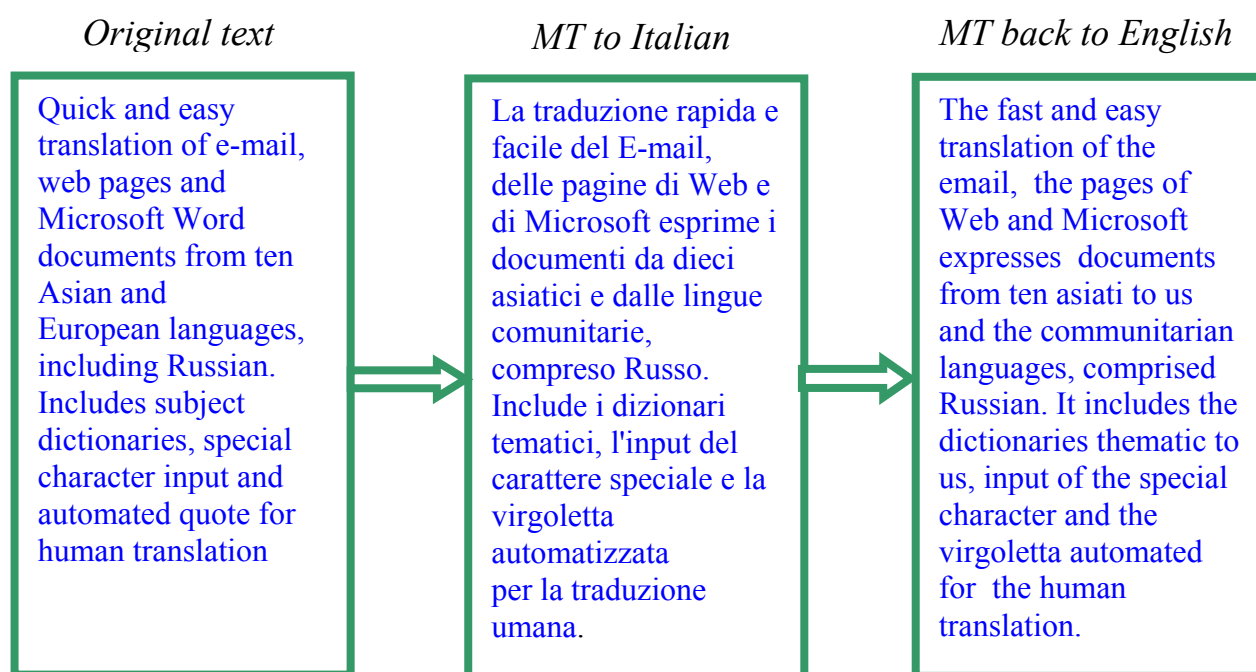


Figure 2. Machine translation of an original English text to Italian and then back to English.

The system is WorldLingo (http://www.worldlingo.com/microsoft/computer_translation.html).

Spoken language translation will of course be an even more challenging task. Here, a spoken utterance in the source language is to be translated to a corresponding spoken utterance in the target language. A speech recognizer transforms the utterance to text, which is then subjected to machine translation. The translated text is then output using TTS. The speech recognizer will introduce recognition errors. The syntax and grammar of spoken utterances need not correspond to the rules of written language. Both these issues will make the task more difficult. Thus far, spoken language translation has only been demonstrated in very limited task domains, e.g. hotel room reservations and banking services.

In *Natural language understanding* the aim is to analyze a text in order to extract the meaning (or the meaning relative to a certain task). The meaning will often be expressed formally, e.g. in a form suitable for database queries when the application is information retrieval. As with machine translation, the differences between spoken and written language and the deficiencies of speech recognition technology complicate matters when trying to extract the semantic content of speech.

Automated abstracting aims at extracting the essential information in a larger document. This is useful for getting a quick overview of a document, and can also be an important feature in speech interfaces to email services. The user can telephone the service, and have the system read email messages using synthetic speech. For messages exceeding a certain length, the user can select to only have an abstract read. The quality of the abstract can be varying.

Some possible roles

Some possibilities for the use of speech and language technology are implied in the above description of the core technologies. In this section I will describe some scenarios for the application of language technology in the Information society. Not all are currently realistic; certainly it is an open issue whether all of them are desirable.

Mobility

The penetration of mobile devices has been one of the dominating successes of the “Information age”. Starting out as a “simple” mobile telephone, enhancements such as WAP, GPRS and in the near future UMTS, provides access to the information wealth of the web from the mobile telephone. The diminishing size of the mobile terminal provides a poor user interface, both for displaying information and for entering textual requests and messages. Exceptions exist, e.g. Nokia Communicator, but the size (and cost) of these terminals makes them an alternative only for a smaller group of users. Some producers of mobile telephones already have predictive dictionaries for SMS message creation. More advanced linguistic tools can simplify the task of creating textual input further. Also, speech recognition for information access and speech synthesis for reading (shorter) pieces of information is an interesting alternative for improving the user interface.

Domestic

Already, more and more domestic appliances feature some “intelligence”. A smart house is a concept where the domestic environment adapts to the inhabitants’ habits and needs, and anticipates the family’s requirements. Now, anyone who has tried to program a VCR knows that the design of the user interface of consumer electronics has a significant potential for improvement. Imagine being able to control the domestic environment by natural, spoken commands (or, optionally, by a universal remote control if the stereo is on too loud), or to be able to call home from the grocery store and have the refrigerator inform you whether or not

you will need to purchase milk. The possibilities are numerous, and to a large extent the technology is already developed. The ultimate question will of course be what kind of interaction we wish to have with our domestic appliances.

Workplace

You no longer have a key, or a key card that can be lost to gain access to your workplace. Your name, a numeral PIN-code and speaker verification is the “Open Sesame” that will unlock the doors. Telephone calls are made by voice dialing, obviating the need for remembering telephone numbers. At your workstation, navigation on the computer desktop can be done using voice commands, and documents can be generated by dictation. In addition, linguistic tools are employed for document generation, creating phrases, passages and in some cases complete documents from a condensed concept. This will help increase productivity also for those who are in an environment where they do not wish to disturb their co-workers by speaking,.

Medical

Medical technology is continually progressing. Speech interfaces is used to aid in many instances where it is important that the doctor needs the full use of both hands (and eyes) on the main task. An example is minimal invasive surgery, where both hands are used to control the micro-surgical tools, and the eyes are intently observing the position of the tools and organs. Voice commands are used for controlling what is shown on the monitor. After the operation, the surgeon dictates the report directly into the computer, and after verifying the transcription; the report is immediately added to the patient record in the hospital’s patient database.

In another ward, another doctor is on his way to the next patient. Using his wireless PDA-like terminal, he accesses the patient database and asks for an update of the patient’s condition. After examining the patient, he again queries the system for possible negative effects a new type of medication might have when used in combination with the current medicines the patient is given. Satisfied that there is no hazard, he prescribes the new medication and dictates the addendum to the patient’s medical record before proceeding to the next patient.

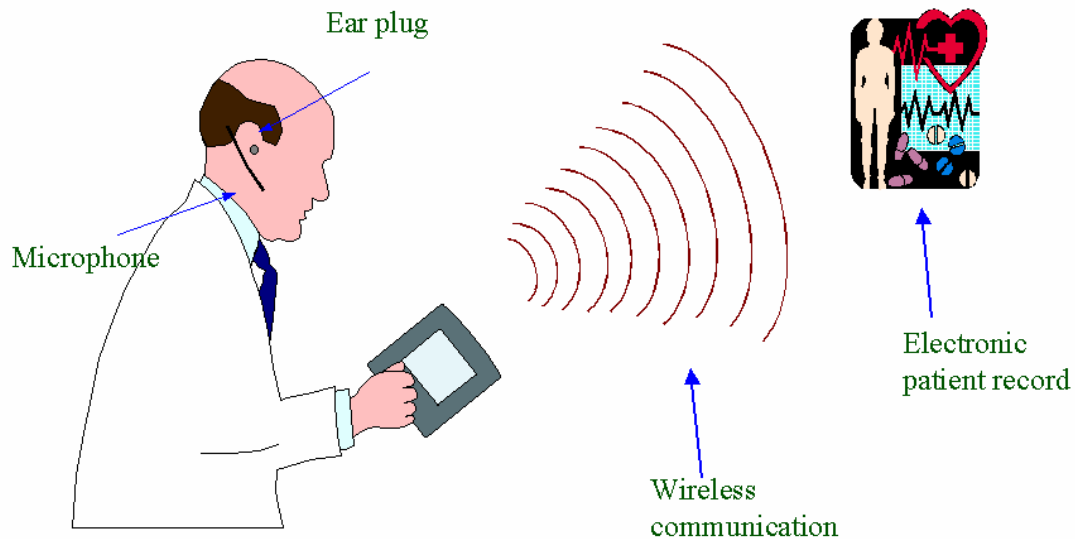


Figure 3. Mobile electronic medical record system

Car and travel

The use of mobile telephones while driving has already been outlawed in a number of countries. Improvements in speech recognition in noisy environments, and the fact that several car manufacturers equip their cars with a microphone array in the sun visor has made voice dialing a standard feature for those who wish to use the telephone when driving.

For the road warrior, the geographical information system in the car is invaluable. Combined with GPS navigation, the system calculates the best route to the destination the driver tells that he wishes to reach, bearing in mind the additional instruction that he wishes to avoid highways. In order to minimize traffic hazard, the display blanks out once the car speeds up, and all driving instructions are given by synthetic speech.

Aids for the disabled

The possibilities for using speech and language technology as aids for the disabled are enormous. Among the applications are reading machines for the blind and people with reading

disabilities, environment control for the physically impaired, speech transcribers for the deaf and dictation systems for blind and dyslectics. For mildly dyslectic people, spell checkers that take into account the typical spelling errors made by dyslectics take much of the trauma associated with text generation away. In addition to improving the general quality of life, aids like these help enabling valuable people to enter the working life rather than being restricted to a life on disability pension.

In many societies, the number of elderly people is steadily growing. Although a significant proportion of the elderly are in good health, many will need care and/or aids to be able to live a good life. Speech and language technology tools, e.g. some of the aids for the disabled, can improve the life quality of people with failing eyesight and hearing, and help elderly with reduced ability of movement to be more self-sufficient and secure.

Concluding remarks

As exemplified above, speech and language technology have many uses that can simplify and speed up everyday tasks for the average citizen. It can also lower the threshold for using information technology for many people. For the elderly and disabled the technology has a potential for improving the quality of life, in some instances significantly. Another important aspect is the concern about the Digital divide, i.e. a society where a significant portion of the population is not able to use the modern technology, and not able to access and utilize the knowledge base provided by information and communication technology. Authorities, both locally, nationally and internationally are increasingly providing information for the citizen on the web. Finding and utilizing the information is the responsibility of the individual. Making access to information simpler for all groups of society is vital for the democracies and for helping individuals to learn about their rights. Speech and language technology can contribute towards that goal.

Finally, a few words of caution. Current speech technology has a very limited understanding of dialects, and in most cases requires a normalized pronunciation. Speech synthesis is mostly provided using a “normal” pronunciation. Document generation tools make style suggestions that conform to a standardized written language. These factors might contribute towards reducing the richness of both spoken and written language. If machine translation becomes sufficiently proficient, much of the motivation for foreign language learning can vanish.

Knowing a language is an important factor for understanding culture, which in turn also contributes to international cooperation and peaceful coexistence.