

Speech Input and Output Technology - State of the Art and Selected Applications

Klaus-Rüdiger Fellbaum

Chair of Communication Engineering
Brandenburg University of Technology Cottbus
Universitaetsplatz 3-4
D-03044 Cottbus
fellbaum@kt.tu-cottbus.de

Abstract: This paper deals with speech communication aspects with an accent on speech dialogues, in which one communication partner is a computer. After some basic definitions, we discuss the speech components of a man-machine dialogue, namely speech recognition and speech synthesis. It will be shown that a computer is far away from human's ability to recognise or understand speech.

1. Introduction

The most important form of communication is speech communication. This explains why humans want to have speech as communication/interaction medium with computers as well. However, today we are still far away from a human-like speech dialogue [WAHL97], [JAYA97], [TAYL89]. As we will see in the next sections, systems for a reliable and robust speech recognition of a large vocabulary are still in the state of research. Concerning speech output of an unlimited vocabulary (speech synthesis), the systems suffer from a machine-like sound. On the other hand, for many useful applications of the real life, restricted forms of speech input or output are sufficient. There are at least two key points for a successful use of speech components: a clear focus on the concrete application and a careful system integration.

2. Communication by voice

Fig.1 shows a rough classification of the man-man communication and the man-machine communication (more precisely, the human-computer interaction), which will be used in this paper.

The man-man communication is the area of telecommunication services, above all speech telephony. Key issues are a good speech quality and low costs. Both factors are strongly influenced by the coding techniques. We will not go into details here because our main subject is more related to the human-computer interface. A very detailed and useful description of speech coding principles is presented in [Jaya84].

Concerning the man-machine communication, we have to distinguish two directions: if the machine speaks, we call it *speech output*, otherwise *speech input*. In a normal conversation, the partners change their roles between speaking (active phase) and listening (passive phase); they perform a dialogue.

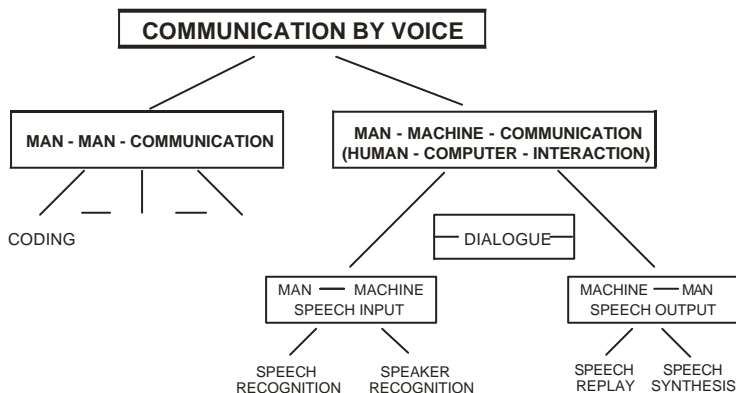


Fig. 1: Classification of voice communication

2.1. Speech input

As shown in figure 1, we have two different forms of speech input, namely speech recognition and speaker recognition. While speech recognition aims at the content of the speech input, speaker recognition tries to identify the speaker. For our topic, speech recognition is by far more important and thus we will restrict here on it.

There are many useful applications for speech recognition, above all in the technical area. Some arguments for speech recognition are listed here:

- Speech is the most important form of human communication,
- hands and eyes are free for other activities,
- communication with a machine and other humans is simultaneously possible,
- system can be used by blind and other disabled people (e.g. motorically disabled),
- freedom of movement and orientation,
- information in dark and dusty rooms,
- the user is not bound to a fixed place,
- compatible with radio and telephone (remote control by voice).

However, there are also drawbacks and problems:

- Speech input can be disturbing for the environment,
- problems might arise with privacy,
- recognisers are extremely sensible against environment noise, above all, background speaker,
- for some applications (e.g. those with high security requirements) the recognition accuracy might be insufficient,
- usually high efforts for system training are necessary.

Concerning applications for speech recognition, different classes can be distinguished:

- Acoustic Data Collection:* Speech input of data and text lists, mobile data collection, parcel sorting, sorting of baggage (e.g. airport), acoustic dialing, quality control,
- Voice Control of Devices and Systems:* Robot control, control of tool machines, hi-fi devices, car functions, operation microscopes,
- Voice-controlled Information Systems:* Flight or train information, data base information, consultation systems,
- Automatic Translation Systems.*

There exist different types of speech recognition. One classification criterium is the presentation form of speech (fig.2).

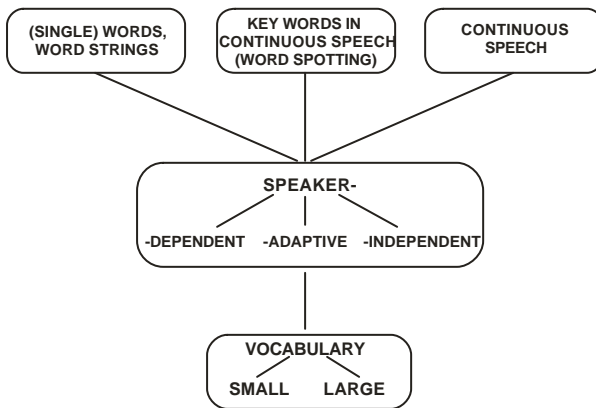


Fig. 2: Forms of speech presentation

In general, the recognition process is based on comparisons between patterns which are stored in a memory and patterns which are pronounced in an actual situation. For more details about the recognition procedure see for example [NEY99].

A very important distinction of recognition forms is *ordinary speech recognition* and *speech understanding*. Recognition in its fundamental form means that all details of a spoken utterance have to be recognized, while speech understanding aims at the *meaning* or the *semantic content* of the speech.

Fig. 3 explains how the recognition process has to be extended for speech understanding facilities. Key components are the syntax and semantic components. In many cases, the recognition procedure generates ambiguities (equally sounding words etc.) which result in word errors. If, for a recognised word, several alternatives exist, in some cases a grammar analysis can exclude some of them. The same holds for a semantic analysis which erases words which do not make any sense in a certain constraint.

The last step, the so-called pragmatic analysis, can decide, if a recognition result (e.g. a spoken command) is useful in an actual situation.

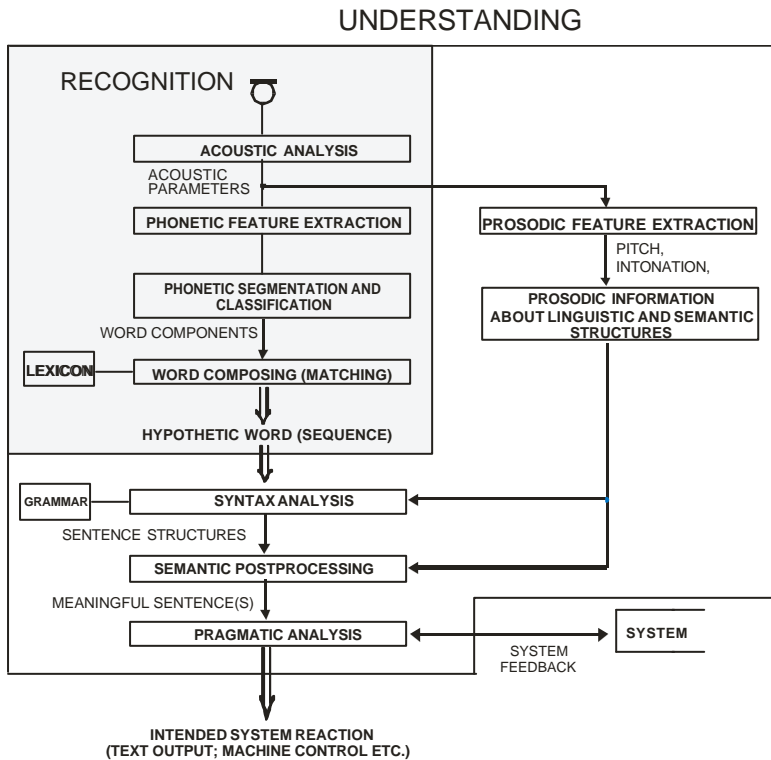


Fig. 3: Speech recognition and speech understanding [FELL99]

We will now summarise the state of the art in speech recognition/understanding.

Isolated word recognition up to a vocabulary in the order of 100,000 words is on the market. Most of the systems are *speaker-adaptive*, that means, at the beginning the recognition accuracy is very moderate, but after an intensive use it improves more and more. The adaptation process is in some cases painfully slow and requires a lot of patience. (This is normally not mentioned in advertising material and it leads sometimes to frustration.). Nevertheless, after a certain time, the accuracy can reach up to 98..99%.

Continuous speech recognition is still in a research or development state. Only few systems are available and the recognition accuracy still leaves to desire. One of the crucial problems has to do with the *segmentation* procedure. Since speech is a fluent medium, very often word boundaries are either extremely difficult to detect or they do not exist for coarticulation reasons. Thus, most of the systems start their work with the recognition of elementary sound elements (like phonemes), these are transcribed into graphemes and then, with the aid of a lexicon, the system tries to concatenate the graphemes to words and finally the words to sentences.

2.2. Speech output

For the most technical applications we only need a limited vocabulary. The system has to pronounce error, alarm or confirmation messages, control instructions, standardized question phrases, help functions and so on. A limited vocabulary can be spoken by a human and stored in a memory. From here it can be replayed and thus we call such a system *replay system*. Another term, usually applied in the American literature, is *voice store and forward system*.

The speech of replay systems sounds naturally and its quality depends on the coding technique and the bit rate. At least, the speech quality is a question of the price which the user is willing to pay.

In some cases we need an *unlimited* speech vocabulary (example: a 'reading machine' for blind persons). Since any kind of text may come into question, it is impossible to record it in advance. We therefore use another technique, namely *speech synthesis*. Its principle will shortly be explained here.

Synthesis means that speech is concatenated from short phonetic elements like phonemes, diphones (double phonemes) or others. It can be shown, that a restricted number of these phonetic elements is sufficient for the generation of an unlimited speech vocabulary.

The synthesis process in its most general form consists of three steps (fig. 4). Starting point is paper material with text, graphics, tables and so on.

In a first step, the paper material has to be pre-processed in a form which is readable for a computer. For ordinary text, the procedure is easy: a scanner and a text recogniser 'translate' the text letters into an electronically readable form, e.g. into ASCII characters. Compared to text, a translation of graphics and structured text (tables etc.) can be extremely difficult or even impossible. Therefore, we will restrict the input here on plain text.

In the second step, the ordinary text is transformed into phonetic symbols which describe the pronunciation much more precisely than orthographic text. We call this *linguistic-phonetic transcription*.

Finally, in the third step, called *phonetic-acoustic transcription*, the symbols are distributed to the related phonetic elements and the phonetic elements are concatenated to continuous speech.

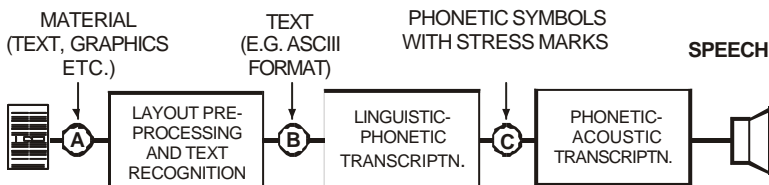


Fig. 4: Principle of text -to-speech synthesis

Although the phonetic elements are taken from natural speech, the generated speech sounds synthetically. The reason for that is twofold: firstly, the various sound transitions, which are typical for natural speech, are replaced here by more or less standardized deterministic transitions and secondly, the phonetic elements are neutral, i.e. they do not have a stress or speech melody.

In summary, the main problems in the speech synthesis area are as follows:

- No consistent relation between written and spoken elements (*graphemes* and *phonemes*),
- generation of a natural intonation (prosody),
- generation of a natural intonation which is consistent to the *semantic content* of a text,
- generation of natural transitions between speech elements (concatenation problem).

3. Dialogue systems

Many problems of an optimal man-machine communication are still unsolved. Users are different, have different preferences and, what is very important, they might have handicaps because they are disabled or elderly. It must be clearly expressed that human factors in voice technology urgently need more attention, although there is a number of useful publications (see e.g. [GARD99], [FELL99]).

In general, a speech-based user interface requires both, speech input (recognition) and speech output (speech synthesis). It is important to state that input and output have narrow links to each other and a dialog is by far more than the sum of both. The dialog manager plays a key role in the system, and its design, comfort and ‘intelligence’ is the main factor for the user’s acceptance of the whole application. Fig. 5 shows such a dialog system. It might be the interface of a telephone-based inquiry system (e.g. for train or airplane directories), a diagnosis system for medical experts or it could be used for advanced Internet applications.

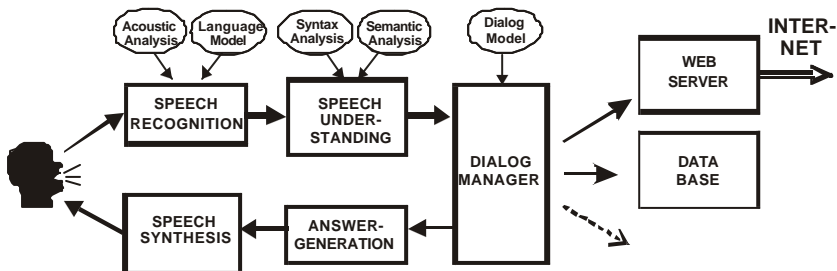


Fig. 5: Speech dialog components for an intelligent user interface [KUNZ00]

In complex applications, the dialog manager performs a user modeling with which, for example, trained and untrained users can be distinguished and the dialog is flexibly adapted to the individual dialog step, the user’s reaction and the requirements of the task.

4. Conclusions

Several research areas have been discussed before. Main issues were, for speech recognition systems, the improvement of robustness and a reliable recognition of spontaneous speech. Also the development of automatic translation systems are a real challenge[W AHL97].

If we now come back to a human-human dialogue, it is obvious that a human who speaks, does not restrict the activity only on the speech production. He completes the conversation by gestures, body movements and so on. If we think of a very noisy environment (manufactory hall etc.), gestures can strongly support the conversation. There are first success-promising research activities, in which a video analysis (for gesture recognition) is combined with speech recognition and in which the recognition results could be remarkably improved [JAYA97].

5. References

- [FELL99] *Fellbaum, K.*: Human-Human Communication and Human-Computer Interaction by Voice. Lecture on the Seminar "Human Aspects of Telecommunications for Disabled and Older People". Donostia (Spain), 11 June 1999
- [GARD99] *Gardner-Bonneau, D.*: Human Factors and Voice Interactive Systems. Kluwer Academic Publishers, Boston (1999)
- [JAYA84] *Jayant, N.; Noll, P.*: Digital Coding of Waveforms. Prentice-Hall, New Jersey 1984
- [JAYA97] *Jayant, N.*: Human Machine Interaction by Voice and Gesture. Proc. ICASSP 97, Munich 1997
- [KUNZ00] *Kunzmann, S.*: Applied speech processing technologies -our journey. The ELRA Newsletter, January-March 2000.
- [NEY99] *Ney, H.*: Speech Recognition-Where do we Stand? Proc. ESCA 99, Budapest, Sept.9, 1999
- [TAYL89] *Taylor, M.M.*: The Structure of Multimodal Dialogue. North Holland, Amsterdam etc. (1989)
- [WAHL97] *Wahlster, W.*: The Verbmobil Project. Project Description. Deutsches Forschungszentrum für künstliche Intelligenz GmbH, Saarbrücken 1997