Why is Speech Recognition Difficult?

Markus Forsberg
Department of Computing Science
Chalmers University of Technology
markus@cs.chalmers.se

February 24, 2003

Abstract

In this paper we will elaborate on some of the difficulties with Automatic Speech Recognition (ASR). We will argue that the main motivation for ASR is efficient interfaces to computers, and for the interfaces to be truly useful, it should provide coverage for a large group of users.

We will discuss some of the issues that make the recognition of a single speaker difficult and then extend the discussion with problems that occur when we target more than a single user.

1 Introduction

The problem of automatically recognizing speech with the help of a computer is a difficult problem, and the reason for this is the complexity of the human language. We will in this article try to sketch some of the issues that make ASR difficult. We will not give any solutions, or present other's solutions, we will instead try to give you a panorama of some of the potential difficulties.

We start by presenting the general setting - what we mean by ASR, why we are interested in performing ASR, and what speech actually is, and then we will list some of the problems we may encounter, and finally end with a discussion.

2 What is speech recognition?

Speech recognition, or more commonly known as *automatic speech recognition*(ASR), is the process of interpreting human speech in a computer.

A more technical definition is given by Jurafsky[2], where he defines ASR as *the building of system for mapping acoustic signals to a string of words*[2]. He continues by defining *automatic speech understanding*(ASU) as *extending the goal to producing some sort of understanding of the sentence.*

We will consider *speaker independent* ASR, i.e. systems that have not been adapted to a single speaker, but in some sense all speakers of a particular language.

3 Why do we want speech recognition?

The main goal of speech recognition is to get efficient ways for humans to communicate with computers. However, as Ben Schneiderman points out in [6], human-human communication is rarely a good model for designing efficient user interfaces. He also points out that verbal communication demands more mental resources than typing on a keyboard. So do we really want to communicate with computers via spoken language?

Mats Blomberg [5] enumerates some of the applications of ASR, and some of the advantages that can be achieved. For example, he mentions *personal computers* that can be voice-controlled and used for dictation. This can be an important application for physically disabled, lawyer etc. Another application he mention is *environmental control*, such as turning on the light, controlling the TV etc.

We feel that speech recognition is important, not because it is 'natural' for us to communicate via speech, but because in some cases, it is the most efficient way to interface to a computer. Consider, for example, people that have jobs that occupies their hands, they would greatly benefit from an ASR controlled environment.

4 What is speech?

When we as humans speak, we let air pass from our lungs through our mouth and nasal cavity, and this air stream is restricted and changed with our tongue and lips. This produces contractions and expansions of the air, an acoustic

wave, a sound. The sounds we form, the vowels and consonants, are usually called *phones*. The phones are combined together into words.

How a phone is realized in speech is dependent on its context, i.e. which phone is proceeding it and which phone is directly following it (the term *triphones* is used for a phone in context). This phenomenon is studied within the area of *phonology*.

However, speech is more than sequences of phones that forms words and sentences. There are contents of speech that carries information, e.g. the prosody of the speech indicates grammatical structures, and the stress of a word signals its importance/topicality. This information is sometimes called the *paralingustic* content of speech.

The term *speech signal* within ASR refers to the analog electrical representation of the contractions and expansions of air. The analog signal is then converted into a digital representation by sampling the analog continuous signal. A high sampling rate in the A/D conversion gives a more accurate description of the analog signal, but also leads to a higher degree of space consumption.

5 Difficulties with ASR

5.1 Human comprehension of speech compared to ASR

Humans use more than their ears when listening, they use the knowledge they have about the speaker and the subject. Words are not arbitrarily sequenced together, there is a grammatical structure and redundancy that humans use to *predict* words not yet spoken. Furthermore, idioms and how we 'usually' say things makes prediction even easier.

In ASR we only have the speech signal. We can of course construct a model for the grammatical structure and use some kind of statistical model to improve prediction, but there are still the problem of how to model *world knowledge*, the knowledge of the speaker and encyclopedic knowledge. We can, of course, not model world knowledge exhaustively, but an interesting question is how much we actually need in the ASR to measure up to human comprehension.

5.2 Body language

A human speaker does not only communicate with speech, but also with body signals - hand waving, eye movements, postures etc. This information is completely missed by ASR.

This problem is addressed within the research area *multimodality*, where studies are conducted how to incorporate body language to improve the human-computer communication.

5.3 Noise

Speech is uttered in an environment of sounds, a clock ticking, a computer humming, a radio playing somewhere down the corridor, another human speaker in the background etc. This is usually called *noise*, i.e., unwanted information in the speech signal. In ASR we have to identify and filter out these noises from the speech signal.

Another kind of noise is the *echo effect*, which is the speech signal bounced on some surrounding object, and that arrives in the microphone a few milliseconds later. If the place in which the speech signal has been produced is strongly echoing, then this may give raise to a phenomenon called *reverberation*, which may last even as long as seconds.

5.4 Spoken language $\neq$ Written language

Spoken language has for many years been viewed just as a less complicated version of written language, with the main difference that spoken language is grammatically less complex and that humans make more *performance errors* while speaking. However, it has become clear in the last few years that spoken language is essentially different from written language. In ASR, we have to identify and address these differences.

Written communication is usually a *one-way communication*, but speech is *dialogue-oriented*. In a dialogue, we give feed-back to signal that we understand, we negotiate about the meaning of words, we adapt to the receiver etc.

Another important issue is *disfluences* in speech, e.g. normal speech is filled with hesitations, repetitions, changes of subject in the middle of an utterance, slips of the tounge etc. A human listener does usually not even

notice the disfluences, and this kind of behavior has to be modeled by the ASR system.

Another issue that has to be identified, is that the grammaticality of spoken language is quite different to written language at many different levels. In [4], some differences are pointed out:

- In spoken language, there is often a radical reduction of morphemes and words in pronunciation.
- The frequencies of words, collocations and grammatical constructions are highly different between spoken and written language.
- The grammar and semantics of spoken language is also significantly different from that of written language; 30-40% of all utterances consist of short utterances of 1-2-3 words with no predicative verb.

This list can be made even longer. The important point is that we can not view speech as the written language turned into a speech signal, it is fundamentally different, and must be treated as such.

5.5 Continuous speech

Speech has no natural pauses between the word boundaries, the pauses mainly appear on a syntactic level, such as after a phrase or a sentence.

This introduces a difficult problem for speech recognition — how should we translate a waveform into a sequence of words?

After a first stage of recognition into phones and phone categories, we have to group them into words. Even if we disregard word boundary ambiguity (see section 5.9.2), this is still a difficult problem.

One way to simplify this process is to give clear pauses between the words. This works for short command-like communication, but as the possible length of utterances increases, clear pauses get cumbersome and inefficient.

5.6 Channel variability

One aspect of variability is the context were the acoustic wave is uttered. Here we have the problem with noise that changes over time, and different kinds of microphones and everything else that effects the content of the acoustic wave from the speaker to the discrete representation in a computer. This phenomena is called *channel variability*.

5.7 Speaker variability

All speakers have their special voices, due to their unique physical body and personality. The voice is not only different between speakers, there are also wide variations within one specific speaker.

We will in the subsections below list some of these variations.

5.7.1 Realization

If the same words were pronounced over and over again, the resulting speech signal would never look exactly the same. Even if the speaker tries to sound exactly the same, there will always be some small differences in the acoustic wave you produce. The *realization* of speech changes over time.

5.7.2 Speaking style

All humans speak differently, it is a way of expressing their personality. Not only do they use a personal vocabulary, they have an unique way to pronounce and emphasize. The speaking style also varies in different situations, we do not speak in the same way in the bank, as with our parents, or with our friends.

Humans also communicate their emotions via speech. We speak differently when we are happy, sad, frustrated, stressed, disappointed, defensive etc. If we are sad, we may drop our voice and speak more slowly, and if we are frustrated we may speak with a more strained voice.

5.7.3 The sex of the speaker

Men and women have different voices, and the main reason to this is that women have in general shorter vocal tract than men. The fundamental tone of women's voices is roughly two times higher than men's because of this difference.

5.7.4 Anatomy of vocal tract

Every speaker has his/hers unique physical attributes, and this affects his/her speech. The shape and length of the vocal cords, the formation of the cavities, the size of the lungs etc. These attributes change over time, e.g. depending on the health or the age of the speaker.

5.7.5 Speed of speech

We speak in different modes of speed, at different times. If we are stressed, we tend to speak faster, and if we are tired, the speed tends to decrease. We also speak in different speeds if we talk about something known or something unknown.

5.7.6 Regional and social dialects

Dialects are group related variation within a language. Janet Holmes[3] defines regional and social dialects as follows:

Regional dialect Regional dialects involves features of pronunciation, vocabulary and grammar which differ according to the geographical area the speaker come from.

Social dialect Social dialects are distinguished by features of pronunciation, vocabulary and grammar according to the social group of the speaker.

In many cases, we may be forced to consider dialects as 'another language' in ASR, due to the large differences between two dialects.

5.8 Amount of data and search space

Communication with a computer via a microphone induces a large amount of speech data every second. This has to be matched to group of phones (monophones/diphones/triphones), the sounds, the words and the sentences. Groups of groups of phones that build up words and words builds up sentences. The number of possible sentences are enormous.

The quality of the input, and thereby the amount of input data, can be regulated by the number of samples of the input signal, but the quality of the speech signal will, of course, decrease with a lower sampling rate, resulting in incorrect analysis.

We can also minimize our lexicon, i.e. set of words. This introduces another problem, which is called *out-of-vocabulary*, which means that the intended word is not in the lexicon. An ASR system has to handle out-of-vocabulary in a robust way.

5.9 Ambiguity

Natural language has an inherent ambiguity, i.e. we can not always decide which of a set of words is actually intended.

This is, of course, a problem in every computer-related language application, but we will here discuss what kind of ambiguity that typically arises within speech recognition. There are two ambiguities that are particular to ASR, *homophones* and *word boundary ambiguity*.

5.9.1 Homophones

The concept *homophones* refers to words that sound the same, but have different orthography. They are two unrelated words that just happened to sound the same. In the table below, we give some examples of homophones:

one analysis	alternative analysis
the **tail** of a dog	the **tale** of the dog
the **sail** of a boat	the **sale** of a boat

Figure 1: Examples of homophones

How can we distinguish between homophones? It's impossible on the word level in ASR, we need a larger context to decided which is intended. However, as is demonstrated in the example, even within a larger context, it is not certain that we can choose the right word.

5.9.2 Word boundary ambiguity

When a sequence of groups of phones are put into a sequence of words, we sometimes encounters *word boundary ambiguity*. Word boundary ambiguity occurs when there are multiple ways of grouping phones into words. An example, taken from [1], illustrate this difficulty:

```
It's not easy to wreck a nice beach.
It's not easy to recognize speech.
It's not easy to wreck an ice beach.
```

This example has been artificially constructed, but there are other examples that occurs naturally in the world. This can be viewed as a specific case of handling the continuous speech, where even humans can have problems with finding the word boundaries.

6 Discussion

In this paper, we have addressed some of the difficulties of speech recognition, but not all of them. But one thing is certain, ASR is a challenging task. The most problematic issues being the large search space and the strong variability.

We think that the problems are especially serious, because of our low tolerance to errors in the speech recognition process. Think how long you would try to communicate verbally with a computer, if it understood you wrong a couple of times in a row. You would probably say something nasty, and start looking for the keyboard of the computer.

So, there are many problems, but does this mean that it is too hard, that we actually should stop trying? Of course not, there have been significant improvements within ASR, and ASR will continue to improve. It seems quite unlikely that we will ever succeed to do perfect ASR, but will surely do good enough.

One thing that should be investigated further, is if humans speak differently to computers. Maybe it isn't natural for a human to communicate in the same way to a computer as to a human. A human may strive to be unambiguous and speak in a hyper-correct style to get the computer to understand him/her. Under the assumption that the training data also is given in this hyper-correct style, this would simplify the ASR. However, if not, it may be the case that hyper-correct speech even make the ASR harder. And if this is not the case, we may investigate how we as human speaker can adapt to the computer to increase the quality of the speech recognition. As pointed out before, our goal is not a 'natural' verbal communication, we want efficient user interfaces.

References

[1] N. M. Ben Gold. *Speech and Audio Signal Processing, processing and perception of speech and music*. John Wiley & Sons, Inc., 2000.

[2] J. H. M. Daniel Jurafsky. *Speech and Language Processing, An introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, Upper Saddle River, New Jersey 07458, 2000.

[3] J. Holmes. *An introduction to sociolinguistics*. Longman Group UK Limited, 1992.

[4] E. A. Jens Allwood. Corpus-based research on spoken language. 2001.

[5] K. E. Mats Blomberg. Automatisk igenkänning av tal. 1997.

[6] B. Schneiderman. The limits of speech recognition. *Communications of the ACM*, 43:63–65, 2000.