Natural Speech Dialogue Systems

Volker Steinbiss

Philips GmbH Forschungslaboratorium, Aachen, Germany
now with Philips Speech Processing, Aachen, Germany - steinbiss@acn.be.philips.com

Abstract: The application of speech recognition and understanding to real-life situations generates new scientific problems. To illustrate this, we give the example of an automatic switchboard system that can be used in a very natural communication style. It is based on a high end speech recognizer that outputs a word lattice which is subject to further evaluation by a speech understanding and a dialogue component. Although word error rates are above 20% for this task, the integration of additional knowledge sources like the telephone directory and the dialogue history can add to both the acoustic and the language model in order to achieve highly improved performance. In order to improve the perceived dialogue quality, the level of confirmation is based on the probability that the request was understood correctly.

INTRODUCTION TO SPEECH DIALOGUE SYSTEMS

Dialogue systems that allow a person to communicate with a computer just by natural speech have reached an impressive performance. Without changing the underlying large-vocabulary speech recognition system, which produces recognition results that are not error-free, it is possible to improve speech understanding accuracy by incorporating knowledge that has to do with the specific dialogue situation. We will line out a few of these techniques in this paper.

THE PROBABILISTIC FRAMEWORK FOR SPEECH UNDERSTANDING

In speech recognition, the statistical approach has been widely accepted. It is based on the observation that speech is highly variable and influenced by factors that we cannot model or control other than with a probabilistic description. The statistical framework is the appropriate one to tackle this "lack of knowledge" and takes advantage of a solid mathematical framework. In a nutshell, it can be described as follows: If a word sequence $W$ is to be recognized from an observation $O$ of acoustic vectors, that have been derived from a speech signal, Bayes' decision rule suggests to take the $\hat{W}$ that maximizes (over all $W$) the probability $P(\hat{W}|O) = P(O|\hat{W})P(\hat{W})/P(O)$ or, equivalently, $P(O|\hat{W})P(\hat{W})$. The scientific challenge of speech recognition is the estimation of the $P(O|W)$, the so-called acoustic model, from speech data, the estimation of the $P(W)$, the language model, from text data and maybe grammar knowledge, and the efficient optimization over the possible word sequences.

The most straight-forward approach to speech understanding would be feeding the output of a speech recognizer into a conventional parser for natural language. Unfortunately, the word recognition errors that occur would induce an unsatisfactory performance in the parsing path. On the other hand, to understand the meaning of a spoken sentence, it is not necessary to get each of the spoken words exactly right. It is thus desirable to let the understanding gain some robustness against recognition errors and to hand over to the parsing component some (maybe implicit) information on the lack of reliability and on alternative recognition hypotheses.

This can indeed be achieved by extending the probabilistic framework that had proven so useful in speech recognition to speech understanding. Without going into the details that can be found in [1,2], the model assumes that the speaker states a set of information items $I$ by uttering a word sequence $W$ that leads to an acoustic observation $O$ of the recognition system. The methods which estimate the model parameters (training phase) and retrieve the most likely information item set (understanding) are in the spirit of the statistical methods used in speech recognition. In particular, there is no unnecessary decision on the sentence level, i.e. it is not the speech recognizer which decides on the hypothesis. In contrast, it delivers several hypotheses that are labeled with probabilities, e.g. via a so-called word graph or just via an n-best list of sentence hypotheses. The understanding grammar, taking into account knowledge sources above the word level (e.g. the dialogue history), decides on the most likely interpretation. It is in our system...
quite tolerant against the semantically irrelevant parts of the sentence such that recognition errors in these “filler phrases” typically do not result in understanding errors.

**IMPROVING SPEECH UNDERSTANDING BY DATABASE CONSTRAINTS AND DIALOGUE HISTORY**

To illustrate how this concept can be used in practice, let us zoom into an example. Our research prototype switchboard system PADIS (Philips Automatic Directory Information System) has been used for internal call handling in our research laboratory since April 1996 and is described in more detail in [3]. PADIS provides telephone and room numbers, e-mail addresses and call completion and operates on a database with around 600 entries. It handles natural-language requests in fluently spoken German. The mixed-initiative dialogue allows maximum flexibility for the caller to express himself.

Although word error rates are above 20% for this task, the integration of additional knowledge sources like the telephone directory and the dialogue history can add to both the acoustic and the language model in order to highly improve performance. In our system [1], the attribute error rate, a measure for understanding quality, improved from 40.5% to 29.5%. Specifically, the application knowledge and the long-range constraints used in the system are:

**Consistency with the database.** As a cooperative user typically asks for valid first / last name combinations, recognition hypotheses that do not refer to at least one possible valid entry are discarded. (The same holds for query hypotheses that are being constructed from several inputs of the caller.)

**Consistency within the interpretation.** Self contradicting hypotheses like "Jones, not Jones" are being discarded.

**Match with the system prompt.** A correction of an item not contained in the system prompt is not accepted.

In each case, we apply knowledge sources that were not available to the speech recognition component. Discarding hypotheses is an extreme case of re-scoring, using zero probabilities.

**IMPROVING DIALOGUE QUALITY BY CONFIDENCE MEASURES**

As recognition is not perfect, the user has to confirm what the system understood. In order to avoid lengthy confirmations, it was requested to base the level of confirmation (repeated question, explicit or implicit confirmation) on the probability that the request was understood correctly. This was formulated within the statistical framework of speech understanding and successfully implemented into our prototype system [4]. The degree of reliability controls the further dialogue flow.

In the example of call completion, a call is transferred if recognition is quite secure, explicit confirmation is requested in less reliable cases, and the system repeats its question if the recognition is unlikely to be correct. This system behavior is highly appreciated by our users.

**ACKNOWLEDGEMENTS**

The speech dialogue system and the methods described above have been built by Andreas Kellner, Bernd Rueber, and Frank Seide. Many others in our speech recognition research group at Aachen have contributed to this system.

**REFERENCES**


