ASL: Architectures for Speech and Language Processing

Wolfgang Menzel
Universität Hamburg, Fachbereich Informatik
Bodenstedtstraße, 16 D-2000 Hamburg 50
email: menzel@informatik.uni-hamburg.de

Abstract

Further advances in speech recognition heavily depend on the design of architectures which are flexible enough to accomodate very different requirements for the flow of data and hypotheses. These requirements result from the peculiarities of the speech understanding process, which is a complex decision procedure covering a number of different levels of language description.
The paper identifies some reasons for assuming additional communicative needs within the recognition system and points out two areas of ongoing research:

- the design of prototype speech processing modules, which basically rely on inter-module interaction and
- the development of software-tools for the creation and modification of different architectural layouts.

1 Introduction

Speech recognition has made considerable progress during the last decade. Recognition rates have increased with almost a constant ratio. Systems became to a certain degree speaker independent ones and nowadays even accept connected speech if it is produced at a reasonable speed (e.g. by slowly reading aloud a preformulated text). By being able to handle dictionaries in the order of some thousand word form entries, the performance of todays prototypes has reached a level which seems to be sufficient for practical applications in quite different areas:

- command interpreters for the control of simple technical devices
- voice activated typewriters
- query language interpreters in well structured domains (e.g. data bases)

What all these application domains have in common, is that speech recognition is used under rather controlled conditions: The number of speakers may easily be constrained to a few selected persons, which in addition may be specially trained in order to take care of the peculiarities of the recognition device (Speak slowly, pause between words, do not ask things outside the scope of the system, don't bother in case of quite unnatural system behaviour, etc.). Things change dramatically if current system solutions are going to be applied in public services where the system is confronted with casual users and their common speaking habits. The most notorious problems under such conditions are mostly connected with phenomena of spontaneously uttered speech which typically exhibits very strong "distortions" on all the levels of language

- *phonetic*: assimilation and reduction up to the almost complete omission of words
- *syntactic*: hesitations, interruptions, restarts, agreement violations etc.
- *semantic*: inconsistent use of terminology, use of metaphors
- *pragmatic*: violations of the Gricean maxims

With respect to all these aspects the performance of current speech recognition systems falls dramatically short of the human speech processing capabilities and usually the recognition rates drop to untolerably low figures even for otherwise rather stable systems.

2 Speech and Language

Speech and language processing technology traditionally became two almost distinct research areas, which both have developed their own theoretical and technological basis and use quite different kinds of models. Speech processing, on the one hand, is oriented on stochastic decicion making in highly noisy environments (usually based on the very popular paradigm of Hidden Markov Models). It uses sequential representations for language utterances which are basically one-dimensional and the linear concatenation of elementary units turns out to be the most prominent structure building operation. Language processing, on the other hand, is mostly concerned with the treatment of written and, hence, deterministically given language. It aims at highly structured representations of the content of an utterance in order to derive adequate responses in dialogue or translation systems.

The principal clash of methods and ambitions between the two areas arises from very different approaches to the modelling of language: traditional solutions for speech recognition tend to merge knowledge from different sources into a single, monolithic model:

- *word models* describe possible sequences of signal processing parameters without any explicit reference to phonological and morphological regularities

- a *language model* usually incorporates syntactic, semantic, pragmatic and domain specific knowledge within a single transition network.[1]

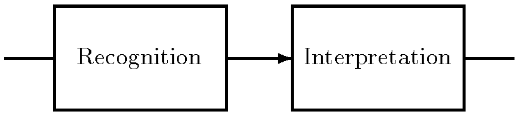
Of course it is not simply ignorance that has led towards such a biased development in favour of unstructured models. Experience has shown that monolithic models are connected with convincing advantages in the areas of efficiency, recognition reliability and the automatic aquisition of model components from speech sample collections.

Linguistic research traditions differ from this approach in at least two ways:

- a strong tendency towards a strict classification of language phenomena as well as a clear modularization of knowledge sources and processing units. Linguistic theory distinguishes a number of levels within the system of language (namely the already mentioned levels of phonology, morphology, syntax, semantics and pragmatics). It will turn out that the adoption of these levels for use in a speech recognition system is not only recommendable from a methodological point of view, but may have very practical consequences for the overall performance of a speech communication system.

- the formal capacity of the models used. Whereas the traditional speech recognition approach describes speech signals in principle as a uniform sequence of atomic symbols, linguistic research considers language utterances as objects with rich internal structure, where structures on the different levels can be mapped onto each other and the linear ordering of units successively becomes less important as one progresses from one level up to the next.

Earlier, the principal gap between these two areas has been bridged by a rather simple *interface approach*: a recognition module is sequentially chained with a linguistic interpretation module, where the first component is responsible to deliver an - in principal - sequential description of the input utterances with an as high as possible reliability. The second module is expected to analyse the obtained description in order to arrive at a semantically and pragmatically adequate system reaction.



Under this schema the recognition component indeed constitutes a (feed forward) interface to a language processing module, whereby in principle the language processing module should be designed in a way, which ideally is independent of the input being written or spoken. The recognition interface passes its results (usually words) up to the linguistic interpretation and there is no possibility to influence the recognition procedure by previously obtained results of the linguistic interpretation. Since unique decisions for a single word form per time unit is inherently susceptible to errors, instead of a simple sequence usually a lattice of partially alternative recognition hypotheses is submitted to the linguistic interpretation.

3 Modularization

There are two major disadvantages connected with the traditional speech recognition approach as sketched above. The first one is connected with the simple interface solution, which does not provide a modularization properly reflecting the structure of

[1] Earlier approaches even attempted to assemble all the knowledge required for recognition into a single model (Lowerre and Reddy, 1980).

the problem domain. Obviously the artificial division into a recognition and an interpretation component involves knowledge which is needed in both components and which therefore has to be maintained twice. Such a duplication first of all is unsatisfactorily from a theoretical point of view. The problem occurs for instance for all the knowledge incorporated in the language model of the recognition module, which is required a second time in the course of the linguistic analysis within the second module.[2] From a more practical perspective this disadvantage is far less obvious: language models usually are trained on corpora of domain specific utterances. They therefore do not only contain syntactic and semantic constraints in just a different form, but typically reflect far more (implicit) domain specific constraints than can be expressed explicitly even in a very restrictive grammar model.

Much more important seems to be a second disadvantage, which is a direct consequence of the predominance of monolithic models. It primarily does not show up in case of normal system behaviour but becomes important only, when difficulties are encountered during recognition or interpretation of the input. Typical contemporary speech recognizers use global optimization criteria to find a decision for e.g. the string of word forms, which gives an optimal description of the input signal (with respect to the internal model). This means that the system will always be able to find a matching string in terms of its internal lexicon, even if the input occasionally turns out to be something else: an unknown word, an utterance in a foreign language or simply noise. It is not a trivial problem, however, to modify the underlying search strategies in such a way, that they allow the recognizer to *partially refrain* from a decision in case of e.g. poor input quality.[3] To leave parts of the speech signal unrecognized, however, is common practice in inter-human speech communication and in a dialogue situation it simply results in asking back to directly inquire the missing information.
What is required here, is the ability of the system to reflect explicitly upon its own performance and its own limitations. In this respect contemporary systems for speech recognition lack all the capabilities of their human counterpart. It is worth noting, that although the ability of selfreflexion may be an important one for a natural dialogue construction, it is *not* mirrored in the recognition rate, the sometimes one and only quality criterion for speech recognition systems!
The possibility to take non-decisions, however, is only a first step to arrive at a more natural communicational behaviour. Additionally, some capabilities to roughly diagnose the reason of a rejection would be useful, because the reaction upon a recognition failure should be dependent of the specific linguistic level on which it occurred. Hence, one should be able to distinguish at least between some of the following cases:

1. *When does the train to kxtscf leave?*
   Violates basic phonological rules, perhaps the input is distorted by noise: Ask for a (partial) repetition:
   Where do you want to go?

2. *Is there a train to alphaville?*
   Phonologically sound, but no lexical entry found (unknown word, may be a name): Ask for the category and the meaning of the word:
   Is alphaville the name of a town or village?
   Where should it be situated?

3. *When will the London train arrive?*
   Semantically sound, but no unique reference in the world model:
   Do you mean the train from or to London?

4. *Will the train to London be here in time?*
   Referentially sound, but no appropriate information in the database:
   I have no information about time delays at all.

5. *Will Pete be on the midsummer night party?*
   Obviously outside the scope of the system (too much unknown words): Explain the possible options:
   There's information about railway time tables only.

Note, that this diversity of system responses essentially differs from a global standard rejection of the type *Rephrase your question!* or *I don't know.* To achieve this level of communicative competence, however, a proper modularization of the system and its knowledge base is an indispensable prerequisite. A system which maintains monolithic word models and which has no access to explicit phonological rules and constraints cannot distinguish between noisy input on the one hand and unknown words on the other (case 1 and 2). Additionally it

[2] PEREIRA tries to derive a language model for the recognition module automatically, starting from the grammar of the linguistic interpretation module.

[3] A similar behaviour could be achieved by the use of word spotting techniques, which focus exclusively on reliable content words during a first pass through the input signal (cf. MAST et al. 1992).

will have difficulties to cope with the problems of spontaneously spoken and thus extremely reduced speech (Church 1987).
Systems with a monolithic language model likewise force the recognition into a domain specific interpretation of the input utterance irrespective of a possibly very different intention of the user (case 5). To succesfully handle cases 3 and 4 it is necessary to provide the system with a dictionary and a world model which cover a slightly greater domain than would be required for a minimum solution for the task at hand.

4 The nonsequential nature of speech processing

At a first glance the hierarchical structuring of language into different levels and their proper reflection in a system architecture again might suggest a linear sequence of modules. In such a chain of modules each component may in principle serve as a kind of filter to the previous one.[4] There is no need, of course, to translate this structure directly into an equivalent modular design. Lexicalized expressions or performative idioms (*I would like to know ...*) as well as compound nouns are typical examples for a rather weak distinction between syntax and morphology.
A slightly closer look at the specifics of spoken language will show, however, that a strictly sequential layout will not be an appropriate on for speech recognition purposes. The most prominent exception, which requires nonsequential architectures to be taken into consideration, is the use of suprasegmental, i.e. prosodic information (rhythm, strees and melody of the speech signal). Since prosodic cues are not distinctive features e.g. in the germanic languages, recognition procedures usually try to abstract away as much as possible from prosodic variation. Prosody, on the other hand, carries a lot of relevant information about the utterance and therefore sometimes is even called the "punctuation of speech". Within a speech recognition system prosodic information can be put to use in at least two different ways:

- to control certain signal processing and classification procedures, since stressed syllables and words usually bear more information content and can be detected more reliably than unstressed ones.
- to disambiguate partial syntactic results, by looking for prosodic markers for phrase boundaries or sentence mood.

[4] With the exeption of the subsymbolic speech level, because the important step from subsymbolic to symbolic descriptions cannot be modelled as a filtering operation.

Especially for the second task sequential architectures are of limited use. Firstly, the bridge between the source and the destination of prosodic information crosses several modules in the classical sense (phonology, morphology, syntax and possibly even semantics) and it is not recommendable to percolate the necessary information up the whole chain of components. Secondly, phrase boundary markers are difficult to detect in the speech signal without strong support from the context. Hence, it is much easier in general not to search for them in a bottom-up fashion, but to ask for the presence of certain indicators only when this is triggered by an ambiguous parsing context. For this purpose a direct communication link between the parsing module and the prosodic component is needed.

There are of course additional reasons which speak in favour of a nonsequential architecture. One of them is the desire to combine very different information pieces during lexical access. In particular, this concerns partial recognition results, which might be combined with semantic hypotheses from the context (e.g. the name of a city starting with a voiceless plosive followed by the vowel "a"). Here, a highly interactive architecture could faciltate a flexible cooperation of components following the principles of mutual information enrichment, which seems to be essential for the processing of spontaneously spoken input.

5 Human speech understanding

It is not just for performance reasons that a modular structure of the speech recognition process should be assumed, which follows classical linguistic distinctions. Obviously, the human speech understanding system is based on very similar principles. There is for instance neuro-physiological evidence for even locally disjoint areas (the so called Broca- and Wernicke-areas) in the human cortex, which are responsible for the syntactic and semantic processing of speech, respectively (Frederici, 1992).
Cognitive oriented theories argue that human speech understanding is a basically hypotheses driven process. Partial recognition results are passed as early as possible to the respective modules in order to perform a syntactic, semantic and

pragmatic verification. This leads to an incremental processing of the input signal. Essential features of such a hypothesis driven architecture can be found e.g. in the speech recognition system EVAR (Mast et al 1992) where keywords trigger semantic and syntactic hypotheses about the surrounding context and the system performs an island driven analysis.
Moreover, ressource limitations of the human brain and the fundamental role of inhibition mechanisms lead to the assumption, that in the course of speech processing only one (or possibly very few) competing hypotheses are investigated. Hence, human speech processing can be characterized as an incremental, time-synchronous and quasi-deterministic process (Görz 1992).
It is quite obvious that cognitive insights of this kind should not be considered for their own sake. There is hope, however, that cognitive adequate architectures may be connected with certain advantages and thus can provide a longtime perspective for the design of speech processing systems with competitive performance characteristics.

6 Integrated Architectures

Arguments from very different areas (communicative performance, the neuronal and cognitive structure of the human analogon, technological considerations etc.) speak in favour of a strict modularization of a speech recognition system. At the same time there is a necessity to provide for a close interaction between the modules in a non-sequential array. Interaction is not only required in order to pass (partial) recognition results along, but additionally to allow a mutual adaptation of different processing parameters to the specific needs of the current recognition situation. This is achieved by requesting and delivering hypotheses and hints to other components. It therefore becomes possible to focus the analysis on the most essential questions in the analysis process and provide for an economical (since goal oriented) allocation of system resources.

7 The ASL-Project

ASL (Integrated Architectures for Speech and Language Systems) is a coordinated research effort of academic and industrial partners funded by the Federal Ministry of Research and Technology. It consists of two subprojects concerned with the optimization of existing stochastics-based solutions and with innovative approaches, respectively. Partially, the ASL-research will be continued in the follow-up project Verbmobil (a German national project for speech-to-speech translation).
Those investigations which aim at innovative proposals for speech recognition architectures[5] follows two basic lines of research:

1. to show possibilities for a well modularized nevertheless strongly interactive system architecture.

This work is based the concept of acoustic events, where explicit phonological constraints are used to reduce the search space (Hauenstein and Hübener 1992, Hauenstein and Hübener 1993). The choosen approach facilitates

- an early verification of recognition results on different levels: segmentation hypotheses are verified by classification results which in turn are subject to a verification by the phonological processing
- a hypothesis-driven approach even at the subphonemic level, where phonological hypotheses control the classification of acoustic events.

Additionally, the prospects of integrating connectionist approaches into a speech recognition system are investigated in detail (Wermter 1992).

The second goal in the search for advanced architectures for speech recognition consists in

2. defining the basic necessities for the design of flexible integrated architectures on the software level.

For this purpose a heterarchical software architecture is considered, which is built of homogeneous modules, having the same structure and the same set of parameters (Pyka 1991). Modules are linked by a number of well defined communication channels. There is no predefined overall direction of information dissemination. Communication links may be defined as required and even modules can be added, omitted or splitted.
Communication links are used to pass messages to other modules. Messages may be hypotheses or request for hypotheses, where each hypothesis is effectively time stamped (begin and - if already available - end point in the speech signal) and has a confidence value attached to it.

[5] With participants at Siemens AG as well as at the Universities of Bielefeld, Berlin, Bonn, Düsseldorf, Erlangen, Hamburg and Saarbrücken

Modules on the other hand consist out of three basic parts

- the processing unit itself
- a module specific space of hypotheses
- a parameter block

The communicational behaviour of a module is completely determined by the current content of the space of hypotheses and the actual values of the communication parameters, e.g.

- the Sending Threshold gives a minimum value for the confidence of an hypothesis to be transmitted to the receiving component.
- the Local Signal Time is the maximum of all end points of hypotheses which the module knows about
- the Actual Temporal Delay is the difference between the latest endpoint of any hypothesis in the system and the Local Signal Time. This delay obviously should be minimized.
- the Maximal Temporal Delay constrains the memory of a module by giving an upper threshold for the individual delay of an hypothesis. Hypotheses which exceed this threshold are suppressed.

The basic module design renders possible a completely local decision about communication activities and makes a central control instance superfluous. Crucial for this purpose, however, is the ability of the modules to valuate its own results.
The communicational infrastructure can be used to integrate rather diverse modules into systems of arbitrary complexity. It offers the possibility for an easy redefinition of existing architectures, that is the architectural layout itself may become subject of experimental investigations and the behaviour of modeles in different environments can be scrutinized.

8 Summary

The current state of speech recognition efforts suggests to look for possibilities to overcome the rather unnatural constraints on the dialogue setting and the communicative behaviour of systems based on contemporary technology. Insights from different sources leads to the design of systems with a more elaborated internal structure than is standard today. To provide for the necessary interaction of components even in the presence of highly modularized system layouts, a new approach to the problems of inter-module communication and architectural design of integrated systems is being investigated.

9 References

Briscoe, E. J. (1987) *Modelling Human Speech Comprehension. A Computational Approach.* Ellis Horwood, Chichester.

Church, K. W. (1987) *Phonological Parsing in Speech Recognition.* Kluwer, Dordrecht.

Friderici, A. (1992) *Natürliche Sprachverarbeitung: Funktionen und Dysfunktionen.* Künstliche Intelligenz, vol. 6, no. 2., p.20-27.

Görz, G. (1992) *Kognitiv orientierte Architekturen für die Sprachverarbeitung.* ASL-Memo.

von Hahn, W. (1992) *Von der Verknüpfung zur Integration: Kontrollstrategie oder kognitive Architektur.* Proc. KONVENS'92, Springer, Berlin. p. 1-10.

Hauenstein, A. and Hübener, K. (1992) *Towards Cognitive Processing of Speech Signals.* Proc. KONVENS'92, Springer, Berlin, p. 100-108.

Hauenstein, A. and Hübener, K. (1993) *Controlling Search in Segmentation lattices of Speech Signals.* to appear: Proc. Eurospeech 1993, Berlin.

Lowerre, B. T. and Reddy, D. J. (1980) *The HARPY Speech Recognition System.* in: Lea, W. A. (ed.) Trends in Speech Recognition, Prentice Hall, Englewood Cliffs, p. 340-360.

Mast, M., Kummert, F., Ehrlich, U., Fink, G., Kuhn, T., Niemann, H. and Sagerer, G. (1992) *A Speech Understanding and Dialogue System with a Homogeneous Linguistic Knowledge Base.* Memo Univ. Erlangen-Nürnberg, IMMD 5.

Pyka, C. (1991) *Deterministische, incrementelle und zeitsynchrone Verarbeitung und die Architektur von ASL-Nord.* Report ASL-TR-22-91/UHH

Wermter, St. (1992) *A Hybrid and Connectionist Architecture for Scanning Understanding.* Proc. ECAI'92, Vienna.