SPEECHLEX - PHONOLOGICAL WORD MODELLING COMPONENT OF AN EXPERIMENTAL SPEECH RECOGNITION SYSTEM

Bernhard Kaspar, Bernd Lochschmidt

Research Institute of the Deutsche Bundespost, Darmstadt, FRG.

ABSTRACT

SpeechLex is a module of an experimental speech recognition system. The main features of this system are the rule-based approach for the acoustic-phonetic detection, the incorporation of Fuzzy Logic to handle uncertain decisions, and an overall top-down recognition strategy. The task of SpeechLex is to set up a phonotactical network describing the words of a given vocabulary with common variations in pronunciation and to drive the acoustic-phonetic rules of another module. Processing is, in principle, carried out left-right, i.e. one syllable after the other, but inside-out within a syllable. Since the whole system is of an experimental nature, emphasis was laid on providing it with graphical aids to trace the processing steps and to support its development.

1. INTRODUCTION

SpeechLex is a component of the experimental speaker independent speech recognition system SpeechStation, which is currently being developed at the Research Institute of the Deutsche Bundespost. SpeechStation is designed as a knowledge-based system, the incorporated knowledge originating from acoustics, phonetics, and linguistics and being represented explicitly, as opposed to more statistically motivated approaches like Hidden Markov Models (HMM) and Neural Networks ([1, 2, 3]) or intermediate approaches of tuning classical algorithms using linguistic knowledge ([4]).

Other main components of the system are SpeechLab, a module for signal processing and representation of speech parameters, SpeechRul, an acoustic-phonetic expert system, and SpeechBase, a relational database handling all global speech and model data. An additional module SpeechMap is planned, whose task will be statistical evaluation as well as support in the generation of new rules for the module SpeechRul. These modules are described in [5], the module SpeechRul in more detail in [6]. Components modeling syntax or dialogue are to be added in the future. See figure 1 for an outline of the architecture.

The different modules represent the conceptual separation of the different processing levels. For developing and testing purposes they are capable of working independently, whereas they operate in concert for the entire recognition process. Here a top-down cascaded strategy is followed (s. Woods in [7]) thus employing a simple control structure and avoiding a complicated data structure like a "blackboard" (s. e.g. [8]). A main feature of the system is the incorporation of "Fuzzy Logic" in order to handle weak decisions and to score alternative decisions. Fuzzy values enter the system via the low-level rules at the interface between physical data and symbolic representation. See [9] for a description of Fuzzy Logic, and [10] for applications in speech analysis.

Up to now, the function of SpeechStation has been the recognition of single words of a given vocabu-
lary. Nevertheless, its design allows for a gradual enlargement of the vocabulary and a development towards simple utterances of connected speech.

Since the system is of an experimental nature, importance was placed on providing it with graphical aids as "explanation components" in order to monitor parts of the recognition process as well as to support its development and carry out basic work like labelling.

2. DOMAIN OF SPEECHLEX
The task of SpeechLex within the system is to incorporate phonotactical knowledge. This comprises the admissible temporal alignment of phones and phone segments in an utterance including knowledge about segment durations, handling of variations in pronunciation, and treating of coarticulation effects.

Unlike in many other approaches (see e.g. [11]), no general phonotactic network of the entire language (in our case German) is provided. Rather the words of a given restricted vocabulary are modelled including effects caused by elision, assimilation, and other variations common in speech. The overall task of SpeechLex during the recognition process is to drive the acoustic-phonetic rules, thus applying only those which are relevant at a certain instant for a given time interval in the recognition process.

3. WORD AND PHONE MODELS
The respective levels of representation existing in the nets are words, syllables, phones, phone segments and special events (like "voice onset"). The syllables are regarded as units, where the main coarticulation effects may occur. Therefore consonants are informed about neighbouring phones in the same (demi)syllable, in order to have the potential (not the obligation), to employ context-dependent rules for acoustic-phonetic decoding, if necessary. For the sake of simplification, vowels and diphthongs are treated as autonomous, even if this may cause difficulties with short vowels.

Both the syllable inventory and the phone inventory are determined by the vocabulary, thus allowing the syllable models and the necessary acoustic-phonetic rules to be developed incrementally. In principle, the phone inventory contains not only the "phonemes" (as defined in linguistics), but also some allophonic variants common in speech. These variations have different sources and are thus represented in different ways:

Formant Pattern, Stylized

\[ \begin{array}{c}
\text{F1} \\
\text{F2}
\end{array} \]

Segment net

Figure 2: Segment net of a diphthong

- Different spectral shapes caused by different places of articulation (e.g. /c/ and /x/ are represented by a single phoneme in German, but are treated as different phones here).
- Different manners of articulation (e.g. rolled or flapped /r/).
- Different segment structures (e.g. aspirated plosives have an additional segment).

In order to take account of the temporal structure of a phone, phone segments are employed as subphone units. The segmentation is motivated by changes in the manner of production (as described in [12]), e.g. voiced/voiceless, stationary, glide, etc.). The phones are classified according to their respective manner of articulation. Each class is then characterised by a special network of segments. Examples are given in figure 2. The individual classes are:

- diphthongs, vowels, nasals and lateral, fricatives, stops, flaps, affricates, and pause.

The phones of each class are then subclassified by the features "place of articulation" and "phonation" for consonants, and by a rough partition of the F1-F2-plain for vowels. In principle, the classification is similar to that exhibited in [13].

4. PROCESSING STRATEGY
With respect to the time dimension, syllables are processed one after the other, left to right, that is, in the natural order. Within a syllable, processing is done inside-out starting with the kernel, that is, the vowel, the diphthong or a syllabic consonant like /m/ or /l/. The underlying reason is, that these phones are normally easier to detect, and are, in general, less context-dependent than those at the syllable boundaries. (In fact, this is
a restricted but more controlled type of "island parsing", which we call "bridge parsing"). As regards the different levels (words, syllables, phones, segments), processing is expectation driven, top-down, cascaded. Alternatives are followed in "best first" order depending on the scoring by the fuzzy values. This is explained by a "snapshot" of the recognition process in figure 3, showing the verification of the German digit "sieben" (seven): The first syllable /si:/ has already been detected. The system has then hypothesized /bun/ as the most likely pronunciation variant of the second syllable and has already detected the vowel /a/. Now it is looking for a /b/ in a short time interval by searching for a voiced burst segment after a pause segment. The advantages of such a strategy are:

- Different levels can work in a quasi-parallel fashion, in this way making use of already recognized parts of the utterance as soon as possible when analyzing the rest. In this manner, ambiguities arising from the analysis at a certain level, could be ruled out at once by the information of a different level and thus need not be carried for a long distance.
- A wrong hypothesis can be rejected as soon as possible; e.g. hypothesis "sieben" (seven) is rejected, if no i-like vowel can be detected with sufficient certainty near the beginning of the utterance.
- Analysis (application of rules) can be performed in a controlled way and only at time intervals, wherever necessary.

It is important to note that no segmentation is needed as a precondition for classification. Rather, in our approach, segment boundaries appear as a by-product of the recognition process. As pointed out, up to now recognition has been carried out as a hypothesis/verification process. In order to tune the system for discrimination with respect to a given vocabulary, two modifications can easily be implemented:

- Features can be generalized if the vocabulary does not require the special cases, e.g. a phone /s/ in a special position within the net can be replaced by "voiceless fricative" if this is sufficient.
- Superfluent branches in the overall network can be cut, e.g. the vowel of a syllable might already be discriminating, the additional consonants being of no importance. This cutting may, of course, conflict with the overall fuzzy classification and has therefore to be handled with care.

5. IMPLEMENTATION

The respective nets are represented in a mechanism similar to the "Definite Clause Grammars" (DCG), as in some systems employed to represent syntax (s. [14]). Nevertheless, there are some important differences:

- The time parameter has to be handled explicitly because speech does not enter the system in discrete items (like words in written language).
- Fuzzy Logic is integrated, so that uncertain decisions can be treated and alternative decisions be scored.
- Multiple evaluation, as inherent in DCG, is reduced by storing results at phonetically relevant levels. (It is being considered to employ a general network representation altogether with a chart parser instead).

SpeechLez is implemented in PROLOG and the graphic system GKS. It is interfaced to the other modules of SpeechStation via mailboxes. The system runs on VAX-Workstations under VMS.

6. SUMMARY

The system SpeechStation represents an attempt to integrate analytic and symbolic processing in speech recognition and to incorporate speech knowledge explicitly at different levels into the recognition process. By employing an expectation driven strategy, it tries to avoid some well known problems occurring in pure bottom-up methods. It also seeks to avoid complicated control and data structures as found in Hearsay-like architectures by restricting itself to a fixed recognition strategy. The module SpeechLez is responsible for the phonotactical modelling of words. Of special interest is the representation of common variations in pronunciation, that are better comprehended structurally than by some general distance measure. In connection with the module SpeechRuI the task of SpeechLez is to drive the respective rules for acoustic-phonetic decoding. It controls the top-down cascaded strategy which attempts to interlock different levels of processing without merging them and to carry out processing in the natural time sequence, at least at the syllable level.

The system SpeechStation does not aim at covering speech processing up to higher linguistic levels.

EUROSPEECH '89, Paris, France, September 1989 1532
in full generality. Rather, it contains the information (e.g. rules) for a restricted domain and a limited vocabulary, but with the potential to enlarge both incrementally, as well as to add higher level components. As a (more) knowledge-based system, SpeechStation is opposed to more statistically modelled systems, e.g. those based on HMM's. We do not regard this opposition as exclusive; rather a synthesis of both methods might be worth considering in the future.

7. REFERENCES