PARSING SPEECH FOR STRUCTURE AND PROMINENCE

Dieter Huber

Department of Computational Linguistics
University of Goteborg

and

Department of Information Theory
Chalmers University of Technology

S-412 96 Gothenburg

Sweden

International Workshop on Parsing Technologies
Carnegie Mellon University
Pittsburgh, Pennsylvania
August 28-31, 1989

INTRODUCTION

The purpose of parsing natural language is essentially to assign to a linear input string of symbols a formalized structural description that reflects the underlying linguistic (syntactic and/or semantic) properties of the utterance and can be used for further information processing.

In most practical applications, this *delinearization* [4] is acheived by some kind of recursive pattern matching strategy which accepts texts in standard orthographic writing, i.e. composed of discrete symbols (the letters and signs of some specified alphabet) and blocks of symbols (words separated by blanks) as input, and rewrites them step by step, in accordance with (1) a lexicon and (2) a finite set of production rules defined in a formal grammar, into a *parse tree* or a *bracketed string*. This approach is commonly restricted to the domain of the sentence as maximal unit of linguistic processing, thus adhering to the traditional view that larger units like paragraphs, texts and discourse, are formed by mere juxtaposition of autarchic, independently parsed sentences.

Clearly, this kind of procedure developed for parsing written language material is not immediately applicable to speech processing purposes. For one, natural human speech does not normally present itself in the acoustical medium as a simple linear string of discrete, well demarcated and easily identifiable symbols, but constitutes a continuously varying signal which incorporates virtually unlimited allophonic variations, reductions, elisions, repairs, overlapping segmental representations, grammatical deficiencies, and potential ambiguities at all levels of linguistic description. There are no "blanks" and "punctuation marks" to define words or indicate sentential boundaries in the acoustic domain. Syntactic structures at least in spontaneous speech are often fragmentary or highly irregular, and cannot be easily defined in terms of established grammatical theory [26]. Last not least, important components of the total message are typically encoded and transmitted by nonverbal and even nonvocal means of communication [18].

On the other hand, human speakers organize and present their speech output in terms of well defined and clearly delimited *chunks* rather than as an unstructured, amorphous chain of signals. This division into chunks is represented among other parameters in the time course of voice fundamental frequency (F_0) where it appears as a sequence of coherent *intonation units* optionally delimited by pauses and/or periods of laryngealization [19], and containing at least one salient pitch movement [9],[20]. Human listeners are able to perceive these units as "natural groups" forming a kind of *performance structure* [12], which reflects the *information structure* of the utterance [14] and is used to decode the intended meaning of the transmitted message. This involves (1) chopping up the message into *information units* in accordance with the speaker's and listener's shared state of knowledge, (2) organizing these units both internally and externally in terms of given and new information, and (3) selecting one or at the most two elements in each unit as points of prominence within the message.

SYSTEM OVERVIEW

While written language input is generally presented to the parser with both the *terminal symbols* (i.e. words) and the *starting symbols* or *roots* (i.e. sentences) clearly delineated and set off from each other by spaces and/or punctuation marks, thus imposing the parsing algorithm with the task to identify som kind of intermediate structure(s) representation composed of *variables* from a finite set of *non-terminal symbols* or *categories* (i.e. the phrase structure, constituent structure, functional structure, etc), essentially the reverse applies when parsing connected speech input. That is, the continuously varying speech signal is presented to the analysis with some kind of intermediate structure(s) representation either immediately observable (e.g. the voiced-unvoiced distinction between individual speech sounds) or readily deducible (e.g. the prosodic structure expressed in patterns of intonation and accentuation) without prior knowledge of higher-level linguistic information, thus leaving the parser with the task to recognize (or rather support the recognition of) both the individual words and the full sentences.

This reverse relationship between text parsing on one side and speech parsing on the other is illustrated schematically in figure 1. It must be appreciated in this context that the intermediate structure(s) representations in text *versus* speech parsing are neither identical nor necessarily isomorphical.

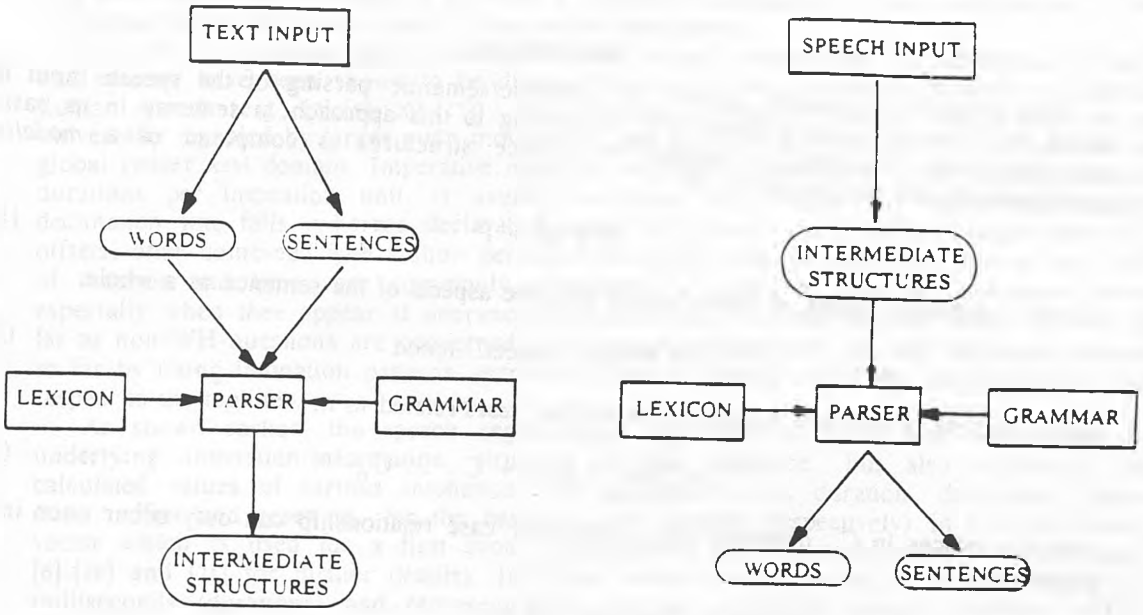


Figure 1 Parsing NL text *versus* parsing connected speech

The speech parsing algorithm presented in this study is thus initiated by a data-driven, left-right speech segmentation stage that exploits the prosodically cued *chunking* present in the acoustical speech signal and uses it to perform automatic, speaker-independent segmentation of continuous speech into functionally defined intonation/information units. For this purpose, two global declination lines are computed by the *linear regression* method, which approximate the trends in time of the peaks (topline) and valleys (baseline) of F_0 across the utterance. Computation is reiterated every time the *Pearson Product Moment Correlation Coefficient* drops below a preset level of acceptability. Segmentation is thus performed without prior knowledge of higher-level linguistic information, with the termination of one unit being determined by the general resetting of the intonation contour wherever in the utterance it may occur.

Earlier studies in the correlations between prosody and grammar have shown that the intonation units thus established time-align in a clearly defined way with units of linguistic structure that can be described in probabilistic terms with respect to three interlacing levels of analysis: constituent structure, linear word count and duration [1],[20]. Furthermore, once the extent of an intonation unit has been established both in the time and in the frequency domain, areas of prominence can easily be detected as overshooting or undershooting F_0 excursions that provide valuable points of departure for further linguistic analysis and island parsing strategies.

A detailed description of the segmentation algorithm together with an evaluation of its performance on three medium sized Swedish texts read by four native speakers (two female, two male) is presented in [21]. Problems of classification by means of hierarchically organized, non-parametric, multiple-hypothesis classifiers are discussed in [6]. A statistical evaluation and coarse classification of the time-alignment between the intonation units established by our segmentation algorithm and features of linguistic structure at the level of a complete sentence (S), clause (C), noun phrase (SUB), verb phrase (VP), adverbial modifier (ADV) and parenthetical construction (PAR) can be found in [20] and [21].

The present paper deals specifically with design aspects of a parsing algorithm that accepts the output of the speech segmentation stage as input and uses it

1 - to build a *case grammar* representation of the original speech utterance;

2 - to guide the word recognition process by generating expectations resulting from partial linguistic analyses.

In the following sections, the grammar formalism, the lexicon and the parser will be presented as separate modules. Problems of integration with other language models (linguistic and stochastic) will be discussed in the summary.

GRAMMAR

The grammar formalism adopted for syntactic/semantic parsing of the speech input is based on Fillmore's *case grammar* [11]. According to this approach, a sentence in its basic structure (deep structure as opposed to surface structure) is composed of a *modality* component M and the *proposition* P:

$$S \Rightarrow M + P \tag{1}$$

where M defines a series of modes which describe aspects of the sentence as a whole:

$$M \Rightarrow \text{tense, aspect...mood} \tag{2}$$

and P consists of the verb together with various *cases* related to it:

$$P \Rightarrow \text{Verb} + C_1 + C_2 \dots C_n \tag{3}$$

with the indices in C_i denoting that a particular case relationship can only occur once in a proposition.

Each case is defined according to Simmons [28] as:

$$C \Rightarrow K + NP \tag{4}$$

where K (which may be null) stands for the preposition which introduces the noun phrase and defines its relationship with the verb:

$$K \Rightarrow \text{Prep} \tag{5}$$

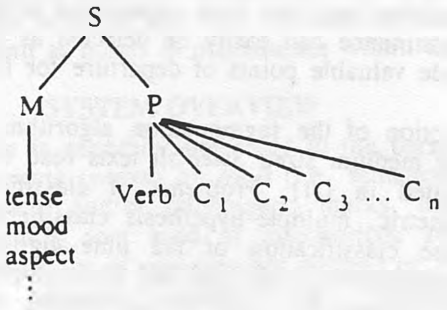
and the noun phrase NP is defined as:

$$NP \Rightarrow (\text{Prep})^* + (\text{Det})^* + (\text{Adj}|\text{N})^* + \text{N} + (\text{S}|\text{NP})^* \tag{6}$$

in which the parentheses denote optional elements, the asterix means that the element may be repeated, and the vertical bar indicates alternation.

A full case grammar representation can thus be described as a tree structure in the form:

(7)



Modality

Within the general framework of case grammar, the following modes and their respective possible values have been adopted:

TENSE	- present, past, future
ASPECT	- perfect, imperfect
ESSENCE	- positive, negative, indeterminate
FORM	- simple, emphatic, progressive
MODAL	- can, may, must
MOOD	- declarative, imperative, interrogative
MANNER	- adeverbial
TIME	- adverbial

The modality of the utterance as a whole is ultimately determined by the combination of the individual values assigned to each of the modes listed above.

At least five of these eight modes, i.e. form, mood, essence and the adverbials of time and manner have been shown to be directly reflected in the intonation contours of natural human speech (e.g. [2],[5],[20],[27]). For instance, emphatic pronunciation appears to be universally signaled by larger pitch movements both in the local (emphatic accent) and in the global (wider *key*) domain. Imperative mood, in addition to displaying on the average shorter durations per intonation unit, is usually associated with higher F_0 onsets and steeper declination line falls, whereas declarative mood is typically cued by low, target-value F_0 offsets, often combined with a short period of laryngealization or devoicing. Adverbials, both of manner and time, are commonly processed in terms of separate intonation units, especially when they appear at utterance-final positions. The interrogative mood, at least as far as non-WH-questions are concerned, is signaled intonationally in most languages studied so far by rising intonation patterns, terminally and/or globally (the latter predominantly with respect to the topline).

As shown earlier, the speech segmentation algorithm not only aims to unearth the underlying intonation/information structure of the utterance, but also represents the calculated values of various intonation unit parameters (i.e. duration, declination slope, onset, offset and resetting, for the baselines and toplines respectively) in a 10-parameter vector which is used for a first broad classification and hierarchization (see references [6],[20] and [21] for further details). Individual values are measured in Hz (F_0-values) or milliseconds (durations) and represented in separate probability density functions (PDF) which allows for (1) finer grain, (2) fast computation of average means, standard deviations and modal targets, and (3) direct comparison and categorization of individual intonation unit parameters reflecting *modality* by simple and robust VQ methods.

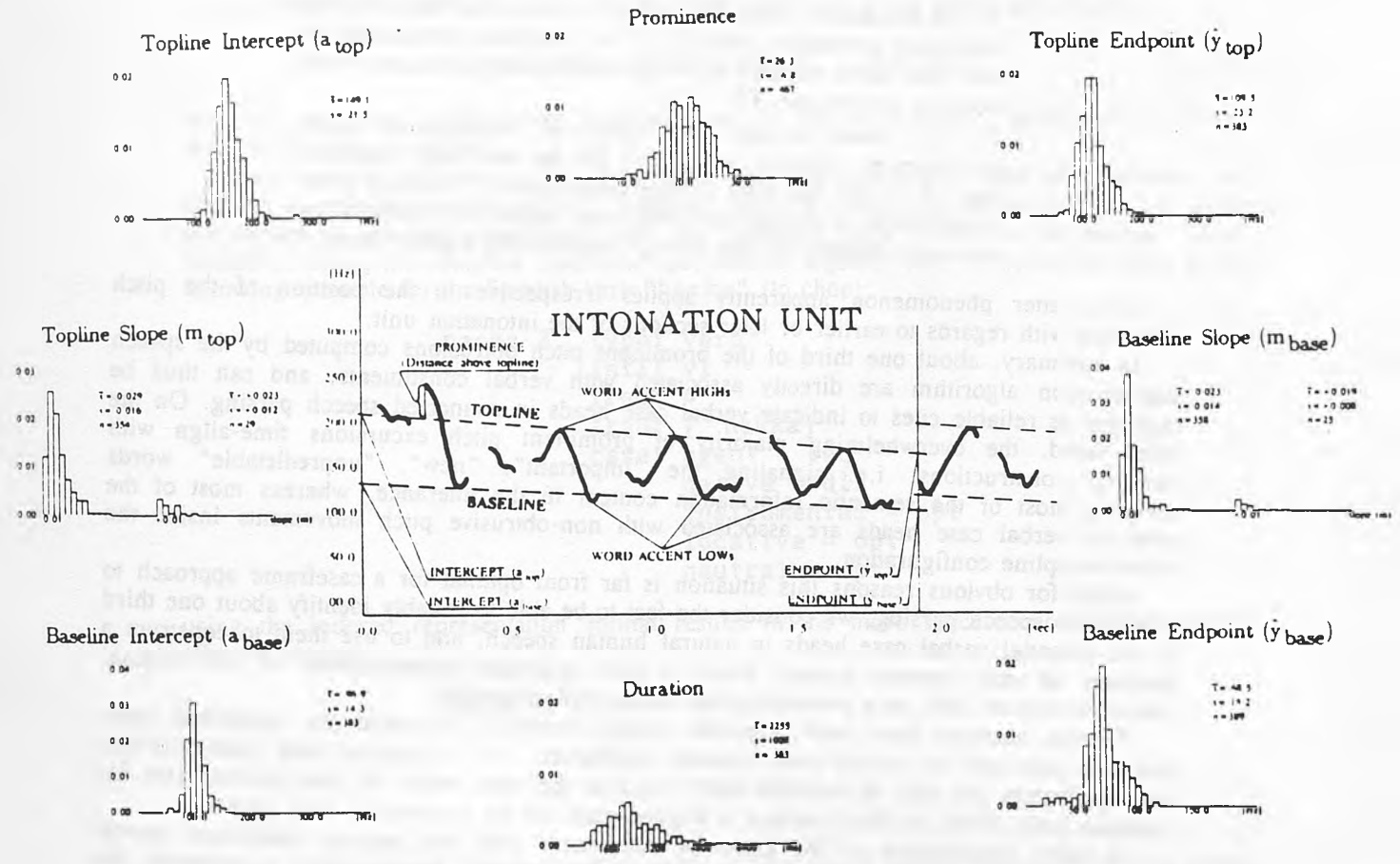


Figure 2 Intonation unit parameters for one male speaker

In summary, modality provides essential information about the propositional content of the utterance. It also provides valuable cues to word order (e.g. interrogative mood is often associated with inverted word order), word structure (e.g. imperative sentences usually lack a lexical expression for the subject, which is commonly understood to be the addressed person), and constituent identity. Determining the modality at an early stage of the parsing process by probabilistic evaluation of the intonational cues specified by the segmentation algorithm thus helps (1) to establish important aspects of the overall meaning of the utterance, and (2) to judge the plausibility of alternative word order hypotheses.

Proposition

In traditional case grammar, the main verb in the proposition constitutes the kernel to which the cases are attached, and the auxiliary verbs contain much of the information about modality. It is thus important to detect and identify the verbal elements of the utterance at an early stage of the parsing process.

It has been shown earlier that once the extent of an intonation unit is established both in the time and in the frequency domain, areas of prominence can easily be spotted as overshooting or undershooting pitch excursions that reach outside the F_0 range defined by the computed baseline-topline configuration. Unfortunately, only a small proportion of these prominent pitch obtrusions (less than one third, i.e. 31.6 %, in our accumulated Swedish material comprising 10440 running words and 704 sentences of read speech recorded by four native speakers) have been found to be directly associated with the verbal constituents in natural human speech, and thus provide an immediate cue for the detection and identification of the *case head*. On the other hand, these verb-prominence coincidences - at least in our Swedish material - have been found to be strongly related:

> 1 - to prominent pitch obtrusions in the *initial* parts of the individual intonation units (81.7 %), whereas prominence in the final parts appears to be predominantly associated with nominal constituents (77.1 %):
>
> 2 - to *lower* average F_0 values of overshooting pitch prominence (typically around 12 Hz for our male speakers and 17-20 Hz for their female counterparts), whereas pitch prominence in connection with focal accent or emphasis on nominals reaches on the average significantly higher values.

This latter phenomenon apparently applies irrespective to the position of the pitch obtrusion with regards to earlier or later sections of the intonation unit.

In summary, about one third of the prominent pitch obtrusions computed by the speech segmentation algorithm are directly associated with verbal constituents, and can thus be regarded as reliable cues to indicate verbal case heads in connected speech parsing. On the other hand, the overwhelming majority of prominent pitch excursions time-align with nominal constructions, i.e. signaling the "important", "new", "unpredictable" words carrying most of the semantic information content in the utterance, whereas most of the potential verbal case heads are associated with non-obtrusive pitch movements inside the baseline-topline configuration.

Albeit for obvious reasons this situation is far from optimal for a caseframe approach to continuous speech parsing, we consider the fact to be able to reliably identify about one third of the potential verbal case heads in natural human speech, and to use them to construct a skeleton of verb kernels around which a case grammar representation of the original utterance can be built, as a promising step in the right direction.

Several attempts have been reported in the literature to extend the traditional case-theoretic approach to include even nominal caseframes, i.e. to construct case grammars that use caseframes not only to describe verbs but also the head nouns of noun phrases (see for instance [15]). Work in this direction is ongoing and will be reported in later papers.

A fuller presentation of the grammar component built for parsing continuous speech input, together with an implementation study for Swedish speech input is prepared for presentation at COLING 90.

LEXICON

The lexicon to be used with the parser is specially designed for speech processing applications (text-to-speech, speech recognition, speech coding, etc) and supports the caseframe approach to continuous speech parsing outlined in this study. Its format is defined as a Swedish monolingual dictionary which contains in addition to the standard entries (head, homograph index, part-of-speech, inflexion code, morphological form classes, etc) also:

1 - a narrow phonetic transcription reflecting standard pronunciation usage;

2 - the textual frequency rating based on a one-million word korpus of Swedish newspaper articles;

3 - an indexed caseframe description for each verb entry.

For the latter purpose, the following reduced set of cases has been adopted from Stockwell, Schachter and Partee [29], with definitions compiled by the author:

AGENT	- animate instigator of the action
DATIVE	- animate recipient of the action
INSTRUMENTAL	- inanimate object used to perform the action
LOCATIVE	- location or orientation of the action
NEUTRAL	- the thing being acted upon (combining the objective and the factive in Fillmore's original list of cases

A caseframe is thus defined as an ordered array composed of the entire set of cases

```
caseframe = array[agent...neutral]                    (8)
```

in which each case can be either required (`req`) or optional (`opt`) or disallowed (`dis`) and must be marked accordingly.

Since several different verbs often share the same particular kind of caseframe, we propose to store the entire set of 3^5 logically possible caseframes as an indexed list, using the indices as pointers (identifiers) with the respective verb entries in the lexicon. Thus, instead of listing the complete caseframe specification together with the lexical entry as in the following example for the Swedish verb "`hacka`" (to chop):

```
hacka 3   type: verb
          infl: v1
          freq: 4
          tran: [ 2 hakka]
          case: agent - req
                dative - dis
                instrumental - opt
                locative - opt
                neutral - opt
```

using the indexed representation format results in the more space-economic and search-effective structure:

```
hacka 3   type: verb
          infl: v1
          freq: 4
          tran: [ 2 hakka]
          case: 97
```

Observe that the entry "`type: verb`" might at first glance appear redundant in view of the fact that to begin with only the verb entries are listed with caseframes. As indicated in the previous section, however, we plan to include caseframe descriptions even for nouns and

other nominal constructions, with feature descriptions based on research on valency theory currently conducted at the department of computational linguistics. Further lexical work is also directed towards the extension of individual case states marked as "req" or "opt" with probabilistic lexical hypotheses derived from KWIC-studies of coherent speech.

PARSER

Given the potentially ungrammatical and often highly fragmentary nature of continuous speech input, the actual parsing of the prosodically segmented utterance is performed following a flexible, multiple-strategy, construction-specific approach as proposed among others by Carbonell & Hayes [8]), Kwasny & Sondheimer [24] and Weischedel & Black [31]. A fundamental objective associated with this kind of approach is to integrate general signal processing and natural language processing techniques (both linguistic and stochastic) in order to fully exploit the combination of partial information obtained at various stages of the analysis.

As shown earlier, the output of the speech segmentation algorithm and input to the parser is a linear sequence of parameter vectors representing the LPC-coefficients and pitch value estimates of the original continuous speech utterance at 16ms-intervals, with the F_0 contour segmented into prosodically defined intonation/information units. Typical prosodic structure representations are exemplified below in figure 3 for three short samples of Swedish (male speaker, high-quality digital recording), English (female speaker, poor-quality analogue recording) and Japanese (male speaker, toll-quality analogue recording) speech.

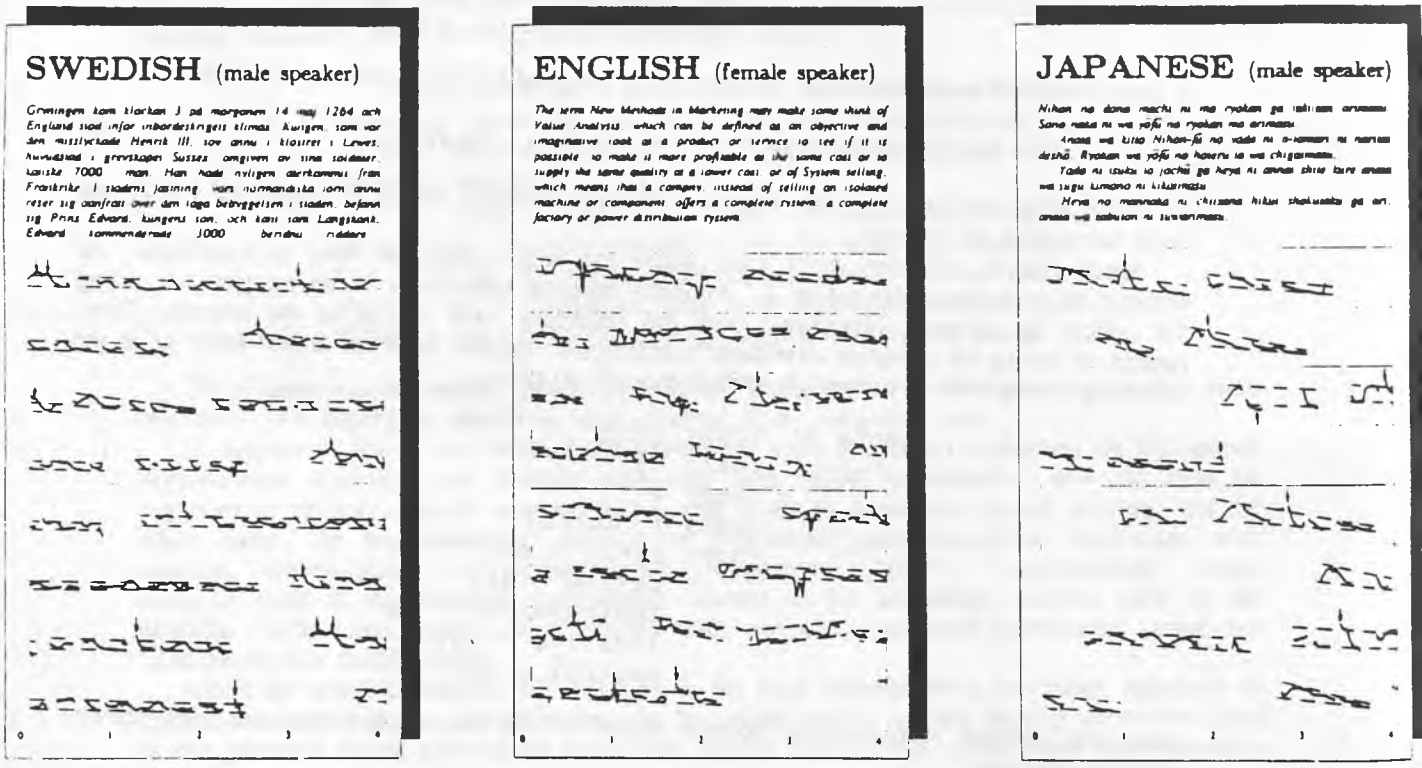


Figure 3 Prosodic structure representation for three short samples of Swedish, English and Japanese speech. Arrows indicate areas of prominence outside the F_0 range defined by the baseline-topline configuration.

The calculated values of the intonation unit parameters duration, declination, onset, offset and resetting, for the baselines and toplines respectively, are stored in a 10-parameter vector and used for a first broad classification and hierarchization of the material.

Once the speech segmentation algorithm has established the extent of an

intonation/information unit both in the time and in the frequeny domain, areas of prominence can be easily spotted as overshooting or undershooting pitch excursions reaching outside the F_0 range defined by the computed baseline-topline configuration. Prominence is measured by the Hz-distance above topline or below baseline respectively (compare figure 2).

Based on the probabilistic data for verb-prominence correspondences established in the previous section, the verbal components of the utterance are localized and used as points of departure for further linguistic processing. As shown among others by Waibel [30] for English and Bannert [3] for German, these pitch obtrusions provide the must reliable cue for the automatic detection if *stress* in continuous speech recognition, i.e. marking the "important" words carrying most of the semantic information content in the utterance. In addition, stressed syllables are commonly pronounced with longer durations and better articulation, which qualifies them as "islands of phonemic reliability", generally scoring better recognition rates than the unstressed (reduced, neutralized) parts of the utterance.

Parsing is run in parallel with the acoustic-phonetic classifier, following a hypothesis-driven island parsing strategy, i.e. using the areas of prominence (islands of reliability) as points of departure for inside-out processing. In other words, the classifier first forms a hypothesis about the phonetic identity of the speech segment(s) at the center of prominence. After that, the island is gradually expanded in both directions by verifying neighbour phone candidates using continuously variable hidden Markov models (HMM) [25] based on precompiled allophone/diphone/triphone statistics [16] and bounded by phonological constraints expressed in the form of finite state transition networks as proposed among others by Church [10].

Island expansion proceeds to the beginning and end of the respective intonation/information unit, thus constructing a phone lattice that spans the entire duration of the IU. A word lattice of the input utterance is hypothesized on the basis of information about (1) the most probable number of words predicted for the respective intonation/information unit as derived from the broad classification [21], (2) the language specific knowledge about the phonotactic properties within words and across words defined by the phonology-constrained diphone and triphone models, (3) the expected *case* identities generated by the *caseframe* entries in the lexicon, and (4) the lexical identities listed in a Swedish pronunciation lexicon [17]. Syntactic (including morphological) constraints are only weakly defined in a constituent-based context-free grammar formulation (CFG), which is aimed to permit successful parses even for fragmentary and/or grammatically deficient speech input and is expected to support the pruning of "unpromising" parses at an early stage of the analysis.

It must be appreciated in this context that only about one fifth of all intonation/information units unearthed by the speech segmentation algorithm (18.2% in our Swedish material) align in a simple one-to-one fashion with full sentences, while the majority (81.8% in the Swedish material) aligns with features of constituent structure in the sub-sentence domain. This implies that the overwhelming majority of full sentences (grammatical as well as ungrammatical) contained in continuous speech is processed in terms of several intonation/information units. Empirical study of our accumulated Swedish speech material revealed an average of 2.36 IUs per sentence with three clearly defined modes at 2, 3 and 5 IUs [20]. It must be appreciated in this context that sentences composed of 4 or more intonation/information units typically contain parallel structures such as enumerations, appositions, parentheticals and rhetorical repetitions.

Given the limited number of actually occurring IU-per-sentence constellations represented by the combination of (1) the most probable number of IUs per sentence, (2) the internal properties of each individual IU specified in a 10-parameter vector containing duration, onset, offset, slope and resetting values for the baseline and topline respectively, and (3) the scored lattice of constituent label(s) derived from the coarse-classification procedure, the sub-problem of sentence generation by intonation unit concatenation can be conveniently solved by a finite-state parsing strategy such as proposed by Gibbon [13], i.e. using a *finite-state automation* (FSA) with transition probabilities attached to each arc.

SUMMARY AND CONCLUSIONS

The speech segmentation, classification and hierarchization components have been developed for Swedish speech input. Testing the algorithm for English and Japanese speech input is ongoing and shows promising results. Further research focuses on improvements in the definition of the linguistic description format (*i.e.* incorporating nominal caseframes, attaching probability scores for *cases* in the "opt" state, including lexical hypotheses with

the caseframe entries, integrating the case grammar with a functional grammar component, etc).

We like to believe that the approach presented in this study shows promise not only for spoken input parsing in general, but for a number of practical applications in the field of speech processing including telecommunication, interpreting telephony, automatic keyword extraction, and text-to-speech synthesis. Linear regression lines are easily calculated and require only little computational effort, which makes the segmentation algorithm a fast, robust and objective technique for computer speech applications. Modulating voice for increased informativity exploits a natural strategy that human speakers use quite automatically in communicative situations involving channel deficiency (e.g. due to static, transmission noise, or masking effects) and/or different kinds of ambiguity [22]. Prominent pitch excursions (together with greater segmental durations) constitute a universally used feature of language that is employed to signal new *versus* given, contrastive *versus* presupposed, thematic *versus* rhematic information in connected speech utterances [7] and can thus be used as a reliable cue to quickly identify the semantically potent keywords in the message. In addition, the frequency range covered by voice phenomena (intonation, accentuation, laryngealization) lies safely within the normal band limits of telecommunication, which qualifies F_0 as a natural, versatile, and accessible code for human-computer interaction via telephone.

Finally, text-to-speech systems using standard syntactic parsers designed to find "major syntactic boundaries" at which the intonation contour needs to be broken into separate units that help the listener to decode the message, invariably come up with the same two kinds of problems [23]:

1 - they tend to produce not one (the most probable, semantically most plausible) but several alternative parses;
2 - they produce too many boundaries at falsely detected or inappropriate sentence locations.

Perceptual evaluation of these synthesized contours reveals that listeners get distracted and often even plainly confused by too many prosodically marked boundaries, while too few prosodic breaks just sound like as if the speaker simply is talking too fast. These findings not only show that the amount of segmentation and the correspondence between syntactic and prosodic units are dependent on the rate of speech, but that listeners apparently neither expect, nor need, nor even want prosodically cued information about all the potential richness in syntactic structure described by modern syntactic theories, in order to decode the intended meaning of an utterance.

REFERENCES

[1] B Altenberg, "Prosodic Patterns in Spoken English", Lund University Press, 1987

[2] H Altmann, "Zur Problematik der Konstitution von Satzmodi als Formtypen", in: J Meibauer (ed) Satzmodus zwischen Grammatik und Pragmatik, Max Niemeyer Verlag, Tübingen 1987

[3] R.Bannert, "From prominent syllables to a skeleton of meaning: a model of prosodically guided speech recognition", *Proceedings of the XIth International Congress of Phonetic Sciences*, Tallinn (Estonia) 1987

[4] A Barr and E A Feigenbaum, "The Handbook of Artificial Intelligence", Stanford 1981

[5] A Batliner, "Produktion und Prädiktion. Die Rolle intonarischer und anderer Merkmale bei der Bestimmung des Satzmodus", in: H Altmann (ed) Intonationsforschungen, Max Niemeyer Verlag, Tübingen 1988

[6] H Broman, P Brauer, E Eliassen, P Hedelin, D Huber and P Knagenhjelm, "Classification: A Problem of Optimization or Organization?", *Proceedings of the STU-Symposium "Digital Communication"*, Stockholm 1989

[7] G Brown, "Prosodic structure and the given/new distinction", in: A.Cutler and D.R.Ladd (eds), Prosody: Models and Measurements, Springer-Verlag, 1983

[8] J G Carbonell and P J Hayes, "Robust parsing using multiple construction-specific strategies", in: L Bolc (ed), Natural Language Parsing Systems, Springer-Verlag, Berlin 1987

[9] W L Chafe, "Givenness, contrastiveness, definiteness, subjects, topics, and points of view", in: Charles Li (ed), Subject and Topic, Academic Press, New York 1976

[10] K W Church. "Phonological Parsing in Speech Recognition". Kluwer Academic Publishers. 1987

[11] Ch Fillmore. "The case for case". in: E Bach and R T Harms. Universals in Linguistic Theory. Holt. Rinehart and Winston. Chicago 1968

[12] J P Gee and F Grosjean. "Performance structures: a psycholinguistic and linguistic appraisal". *Cognitive Psychology* **15**. 1983

[13] D Gibbon. "Finite state processing of tone systems". *ACL Proceedings. 1987*

[14] M A K Halliday. "Theme and Information in the English Clause". Oxford University Press. 1976

[15] Ph J Hayes. A G Hauptmann. J G Carbonell and M Tomita. "Parsing spoken language: a semantic caseframe approach". *ACL Proceedings*. 1986

[16] P Hedelin. D Huber and A Leijon. "Probability distribution of allophones. diphones and triphones in phonetic transcriptions of Swedish newspaper text". Chalmers Report 8. 1988

[17] P Hedelin. A Jonsson and P Lindblad. "Svensk Uttalslexicon (Swedish Pronunciation Lexicon)". Chalmers Report 4. 1989

[18] D Huber. "On the Communicative Function of Voice in Human-Computer Interaction". STIMDI 2. Stockholm 1988

[19] D Huber. "Laryngealization as a Boundary Cue in Read Speech". *Proceedings of the Second Swedish Phonetics Conference*. Lund 1988

[20] D Huber. "Aspects of the Communicative Function of Voice in Text Intonation". PhD Dissertation. Göteborg 1988

[21] D Huber. "A statistical approach to the segmentation and broad classification of continuous speech into phrase-sized information units". *Proceedings ICASSP 89*. Glasgow 1989

[22] D Huber. "Prosodic Contributions to the Resolution of Ambiguity". *Proceedings of the Conference NORDIC PROSODY V*. Åbo (Finland). 1989

[23] D H Klatt. "Review of text-to-speech conversion for English". *Journal of the Acoustical Society of America* 82(3). 1987

[24] S C Kwasny and N K Sondheimer. "Ungrammaticality and extra-grammaticality in natural language understanding systems". *Proceedings of the 17th Annual Meeting of the Association for Computational Linguistics*. La Jolla. Cal. 1979

[25] S E Levinson. "Continuously variable hidden Markov models for automatic speech recognition". *Computer Speech and Language* **1**. 1986

[26] J Löfström. "Repliker utan Gränser (Boundless Conversational Exchanges)". PhD Dissertation. Göteborg 1989

[27] W Oppenrieder. "Intonarische Kennzeichnung von Satzmodi". in: H Altmann (ed) Intonationsforschungen. Max Niemeyer Verlag. Tübingen 1988

[28] R F Simmons. "Semantic networks: their computation and use for understanding English sentences". in: R C Schank and K M Colby (eds). Computer Models of Thought and Language. W H Freeman & Co. San Francisco 1973

[29] R P Stockwell. P Schachter and B H Partee. "The Major Syntactic Structures of English". Holt. Rinehart and Winston. New York 1973

[30] A Waibel. "Prosodic knowledge sources for word hypothesization in a continuous speech recognition system". *Proceedings ICASSP 87*. Dallas 1987

[31] R M Weischedel and L Black. "Responding to potentially unparseable sentences". *American Journal of Computational Linguistics* **6**. pp.97-109. 1980