ON THE USE OF PROSODY IN A SPEECH-TO-SPEECH TRANSLATOR

Volker Strom (1), Anja Elsner (1), Wolfgang Hess (1), Walter Kasper (4), Alexandra Klein (2), Hans Ulrich Krieger (4), Jörg Spilker (3), Hans Weber (3), Günther Görz (3)
e-mail: vst@asl1.ikp.uni-bonn.de

(1) Institute of Communications Research and Phonetics (IKP), University of Bonn,
(2) University of Wien, Austrian Research Institute of Artificial Intelligence
(3) University of Erlangen-Nürnberg, Computer Science Institute (AI)
(4) German Research Center for AI, DFKI GmbH, Saarbrücken

ABSTRACT

In this paper a speech-to-speech translator from German to English is presented. Beside the traditional processing steps it takes advantage of acoustically detected prosodic phrase boundaries and focus. The prosodic phrase boundaries reduce search space during syntactic parsing and rule out analysis trees during semantic parsing. The prosodic focus faciliates a "shallow" translation based on the best word chain in cases where the deep analysis fails.

1. INTRODUCTION

Integration of prosody into a speech-to-speech translator as an additional speech-language interface is a current topic of research. Within the VERBMOBIL project, which aims at translation of spontaneously spoken dialogues, a special experimental system called INTARC was designed which performs translation in an incremental manner. For this purpose time synchronous versions of traditional processing steps such as word recognition, parsing, semantic analysis, and transfer had to be developed. In part completely new algorithms had to be designed in order to achieve sufficient processing performance to compensate for the lack of right context in search. The use of prosodic phrase boundaries became essential to reduce search space in parsing and semantic analysis.

A further goal was robustness: If detailed linguistic analysis fails, the system should nonetheless be able to produce an approximately correct output. For this purpose, system has a second template-based transfer strategy besides the main data flow as a supplement, where a rough transfer is performed on the basis of prosodically focused words and dialogue act detection.

The material investigated consists of spontaneously spoken dialogues on appointment scheduling. A subset of 80 minutes speech was prosodically labelled: Full prosodic phrases (B3 boundaries) are distinguished from intermediate phrases (B2 boundaries). Irregular phrase boundaries are labelled with B9, and the default label for a word boundary is B0. The B2 and B3 boundaries roughly correspond to the linguistic concepts of phrase boundaries, but are not necessarily identical to those [11].

2. OVERVIEW OF THE SYSTEM

Like the VERBMOBIL prototype the INTARC system is a speech-to-speech translator from German to English.

While signal processing is performed by a gradient box, there are three modules that carry out traditional recognition tasks: the focus detector, the phrase boundary detector, and the word recognizer, which is a beam decoder based on the HTK[1] toolkit.

The word lattice is incrementally transferred to a search engine which decodes the n best trees out of the lattice. In this step the output trees are computed as a beam search maximization over a linear combination of five probabilistic values coming from the acoustic word model, the word bigram model, the prosodic phrase boundary model, the phrase boundary language model, and the probabilistic grammar model. The influence of the prosodic components on the search is discussed below.

The trees found are propagated to the semantic components which build the appropriate semantic representation for each tree. Syntactic and semantic analysis are based on a common integrated grammar specification in the HPSG framework. The distributed grammar processing model presupposed in

[1] Hidden Markov Toolkit of Entropic Research Laboratory, Inc

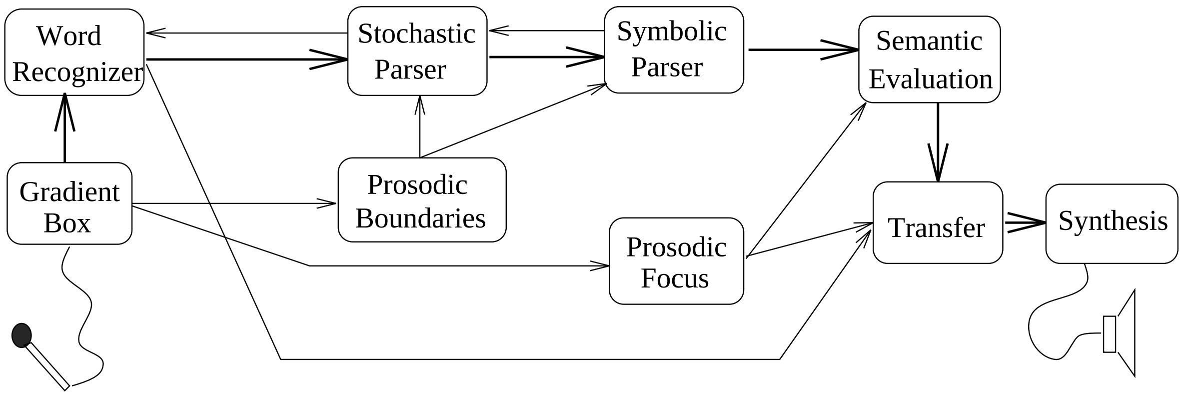

Figure 1: *Overview of the INTARC system: Bold arrows respresent the main data flow. Prosodic boundaries support stochastic and semantic parsing, the prosodic focus supports semantic evaluation. If deep analysis fails, a shallow translation is performed on the basis of the best word chain and the prosodic focus.*

the system is based on [3, 8]. The grammar integrates focus and speech act related information about utterance mode (e.g., interrogative) from prosody [7]. This information is exploited in semantic evaluation for speech act recognition and contextual reference resolution. Speech act recognition uses a finite state dialogue model with probabilistic preferences.

As a robust fallback of the system the decoder's best string is searched for cue words which match with the prosodic focus detector. If the detailed semantic analysis fails, the word sequence found in this "shallow" fallback is directly translated on the basis of templates associated with the cue words and focus information. In about 30% of the detailed analysis' failures the fallback provides an acceptable translation result.

3. PROSODY MODULE

The prosody module consists of two independently working parts: the phrase boundary detector [10] and the focus detector [9].

3.1. Phrase Boundary Detector

First, a parameterization of the fundamental frequency and energy contour is obtained by calculating eleven features per frame: F0 is interpolated in unvoiced segments and decomposed by three band pass filters. F0, its components, and the time derivatives of those four functions yield eight F0 features which describe the F0 contour at that frame globally and locally. Furthermore three bands of a short-time FFT followed by median smoothing are used as energy features.

The phrase boundary detector then views a window of (if possible) four syllables. Its output refers to the syllable boundary between the second and the third syllable nucleus (in the case of a 4-syllable window). Syllables are found by a syllabic nucleus detector based on energy features derived from the speech signal.

For each window a large feature vector is constructed: The mentioned 11 features at each of the 4 syllable nuclei in the window, plus 7 time features (the lengths of the four syllable nuclei and the distances between them). The 30 best features have previously been determined with a feature selection algorithm.

A Gaussian distribution classifier was trained to distinguish between all combinations of boundary types and tones. The classifier output was then mapped on the the four classes B0, B2, B3, and B9. The a posteriori probabilities are used as confidence measure. When taking the boundary with maximal probability the recognition rate for a test set of 30 minutes is 80.76%, average recognition rate is 58.85%.

3.2. Focus Detector

The focus detection module of INTARC works with a rule-based approach. The algorithm tries to solve focus recognition by global description of the utterance contour, in a first approach represented by the fundamental frequency F0.

A reference line is computed by detecting significant minima and maxima in the F0 contour. The average values between the maximum and minimum lines yield the global reference line. Focus accents occur mainly in the areas of steepest fall in the F0 course. Therefore, in the reference line the points with the highest negative gradient were determined first in each utterance. To determine the position of the focus the nearest maximum in this region has been used as approximation.

The recognition rate is 78.5%, and the average recognition rate is 66.6%. The focus detection module will send focus hypotheses to the semantic module and to the module for transfer and generation.

In a recent approach, phrase boundaries from the detector described above were integrated in the algorithm [4]. With help of the phrase boundaries the detection task can be split up so that focus accents for each phrase separately can be determined. The recognition rates are more than two percent points higher, depending on the dialogue. After optimization of the algorithm even higher rates are expected.

4. SYNTAX PARSER

One of the main benefits of prosody in the INTARC system is the use of prosodic phrase boundaries inside the word lattice search. The incremental probabilistic search engine based on [6] receives word hypotheses and phrase boundary hypotheses as an input.

The input is represented as a chart (i.e. a well formed substring table) where frames correspond to chart vertices and word hypotheses are edges which map to pairs of vertices. Word boundary hypotheses (WBHs) are mapped to connected sequences of vertices which lie inside the time interval in which the WBH has been located. The search engine tries to build up trees according to a probabilistic context free grammar derived from the HPSG grammar and supplied with higher order Markov probabilities. Partial tree hypotheses are uniformly represented as chart edges. The search for the n best output trees consists of successively combining pairs of edges to new edges guided by an overall beam search strategy.

The overall score of a candidate edge pair is a linear combination of three factors which we call decoder factor, grammar factor and prosody factor. The decoder factor is the well known product of the acoustic and bigram scores of the sequences of word hypotheses covered by the two connected edges. The grammar factor is the normalized grammar model probability of creating a certain new analysis edge given the two input edges. The prosody factor is calculated from the acoustic WBH scores and a class based tetragram which models sequences of words and phrase boundaries.

When calculating a prosody factor for an edge pair, we pick the WBH associated with the connecting vertex of the edges. This WBH forms a sequence of WBH's and word hypotheses if combined with the portions already spanned by the pair of edges. A score for this sequence can be calculated using the tetragram model. Since word boundary hypotheses are distributions defined over a space of 4 different word boundary instances, we use a local Viterbi maximization to compute the prosody factor.

Tests for the contribution of the prosody factor to the overall search lead to the following results: The same recognition performance in terms of n best trees could be achieved using 20% less edges on the average. A lot of edges are constant in a given search space – namely those used for the representation of the original set of word hypotheses and the empty active rule edges which have a zero span. Counting only those edges which are built up dynamically by the search process a reduction of 65% was measured.

5. SEMANTICS CONSTRUCTION

The Semantic Construction Component (Sem-Parser) is a bottom-up chart parser that uses the semantic information of the HPSG grammar for dialogue turns. Its primary input are the syntax tree hypotheses derived by the syntax parser. It operates essentially by doing chart re-construction. In case of non-applicable rule combinations, a failure is reported back to the syntax parser, thus narrowing down its overall space of hypotheses.

Since the grammar describes full dialogue turns and not just sentences, one major problem is that of segmenting a turn into the correct utterance segments, for reasons of efficiency of parsing as well as for correct grammatical analysis. Therefore, the Sem-Parser exploits information from the phrase boundary detector to reduce its search space. This is achieved by informing the parser which grammar rules are segment connecting and which are only segment internal. Clearly, segment-connecting rules enforce a boundary between segments, whereas segment-internal rules require the opposite. For ideal phrase boundary hypotheses derived from the hand-labelled data, we achieved a reduction of parsing hypotheses by 65.4%. This ruled out 41.9% of the analysis trees.

Since prosodic information is not always reliable, and also because prosodic boundaries do not completely coincide with *grammatical* phrase boundaries, the Sem-Parser was extended by a recovery mechanism, making it possible to reactivate hypotheses excluded by boundary information, thus enabling the derivation of otherwise lost readings [7]. By using the recovery mechanism, real boundary hypotheses reduced the average number of readings of a turn by 24.7%.

6. TRANSFER

In INTARC the transfer module performs a dialogue act based translation. In a traditional deep analysis it gets its input (dialogue act and feature structure) from the semantic evaluation module. In an additional path a flat transfer is performed with the best word chain (from the word recognition module) and with focus information [5].

During shallow processing the focus accents are aligned to words. If a focus is on a content word a probabilistically selected dialogue act is chosen. This dialogue act is then expanded to a translation enriched with possible information from the word chain.

Flat transfer is only used when deep analysis fails. First results show that the 'focus-driven' transfer produces correct (but sometimes reduced) results for about 50% of the utterances (including those where the deep analysis secceeds). For 45 % of the utterances information is not sufficient to get a translation; only 5% of the translations are absolutely false.

7. CONCLUSION

Acoustic information about prosodic phrasing is used in two ways inside INTARC: The stochastic lattice parser uses phrase boundary hypotheses in conjunction with a probabilistic word-phrase boundary model to restrict the search when traversing the word lattice from left to right. When using prosodic phrase boundaries the search space is reduced by 20%, or by 65% when only counting edges built up dynamically by the search process. This compares well with results achieved by other groups [1, 2] when taking the different architectures into account.

The symbolic parser uses the same phrase boundaries to rule out 49.1 % of the analysis trees when using ideal boundaries hypotheses. By using the recovery mechanism, real boundary hypotheses reduced the average number of readings of a turn by 24.7%.

The prosody of illocutionary focus is detected by an own acoustic classifier. It is used inside the semantic evaluation and, if the deep analysis fails, for a template-based translation strategy. In this cases a translation using the best word string coming from the word recognizer with single words marked by the focus detector is performed. 30% of those translations are acceptable.

While the deep analysis uses prosody to reduce search space and disambiguate in cases of multiple analyses, the 'shallow focus based translation' can be viewed as directly driven by prosody.

8. ACKNOWLEDGEMENT

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the Verbmobil Project under Grant 01 IV 101 G. The responsibility for the contents of this study lies with the authors.

9. REFERENCES

1. G. Bakenecker, H.U. Block, A. Batliner, R. Kompe, E. Nöth, and P. Regel-Brietzmann. Improving Parsing by Incorporating 'Prosodic Clause Boundaries' into a Grammar. In *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 1115–1118, Yokohama, September 1994.
2. A. Batliner, A. Feldhaus, S. Geißler, T. Kiss, E. Nöth, and P. Regel-Brietzmann. Improving parsing by incorporating 'prosodic clause boundaries' into a grammar. In *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 1115–1118, Philadelphia, 1996.
3. Abdel Kader Diagne, Walter Kasper, and Hans-Ulrich Krieger. Distributed parsing with HPSG grammars. In *Proceedings of the 4th International Workshop on Parsing Technologies, IWPT-95*, pages 79–86, 1995.
4. A. Elsner. Focus detection with additional information of phrase boundaries and sentence mode. In *Proc. European Conf. on Speech Communication and Technology*, Rhodes, 1997.
5. A. Elsner and A. Klein. Erkennung des prosodischen Fokus und die Anwendung im dialogaktbasierten Transfer. internal Verbmobil Memo Nr. 107, Univ. Bonn, Univ. Hamburg, 1996.
6. G. Görz, Marcus Kesseler, Jörg Spilker, and Hans Weber. Research on architectures for integrated speech/language systems in verbmobil. In *Proc. Int. Conf. on Computational Linguistics*, Kopenhagen, 1996.
7. Walter Kasper and Hans-Ulrich Krieger. Integration of prosodic and grammatical information in the analysis of dialogs. In *Proceedings of the 20th German Annual Conference on Artificial Intelligence, KI-96*, 1996. Springer: Lecture Notes in Computer Science, Berlin.
8. Walter Kasper and Hans-Ulrich Krieger. Modularizing codescriptive grammars for efficient parsing. In *Proceedings of the 16th International Conference on Computational Linguistics, COLING-96*, pages 628–633, 1996.
9. A. Petzold. Strategies for focal accent detection in spontaneous speech. In *Proc. Int. Conf. on Phonetic Sciences*, volume 3, pages 672 – 675, Stockholm, 1995.
10. V. Strom. Detection of accents, phrase boundaries and sentence modality in German with prosodic features. In *Proc. European Conf. on Speech Communication and Technology*, volume 3, pages 2039–2041, Madrid, 1995.
11. V. Strom and C. Widera. What's in the "pure" prosody? In *Proc. Int. Conf. on Spoken Language Processing*, Philadelphia, 1996.