PROSODIC PROCESSING AND ITS USE IN VERBMOBIL

H. Niemann[1] *E. Nöth*[1] *A. Kießling*[1] *R. Kompe*[1] *A. Batliner*[2]

[1]Lehrstuhl für Mustererkennung (Informatik 5), Universität Erlangen–Nürnberg, Martensstr. 3, 91058 Erlangen, Germany
[2] Institut für Deutsche Philologie, L.–M. Universität München, Schellingstr. 3, 80799 München, Germany
e–mail: noeth@informatik.uni-erlangen.de, www: http://www5.informatik.uni-erlangen.de/

ABSTRACT

We present the prosody module of the VERBMOBIL speech–to–speech translation system, the world wide first complete system, which successfully uses prosodic information in the linguistic analysis. This is achieved by computing probabilities for clause boundaries, accentuation, and different types of sentence mood for each of the word hypotheses computed by the word recognizer. These probabilities guide the search of the linguistic analysis. Disambiguation is already achieved during the analysis and not by a prosodic verification of different linguistic hypotheses. So far, the most useful prosodic information is provided by clause boundaries. These are detected with a recognition rate of 94%. For the parsing of word hypotheses graphs, the use of clause boundary probabilities yields a speed–up of 92% and a 96% reduction of alternative readings.

1. INTRODUCTION

Already Lea [17] and Vaissière [26] have proposed the use of prosodic analysis in automatic speech understanding systems; illustrations for this use are given in the examples below. Even though the number of research projects on prosody in the context of automatic speech recognition/understanding has increased steadily over the past ten years, VERBMOBIL is world wide the first complete speech understanding system, where prosody is really integrated. Moreover with VERBMOBIL it can be demonstrated that prosody leads to drastic performance improvements. We see the following reasons for this gap between the amount of research on prosody and its use in complete systems:

The major role of prosody in human–human–communication is segmentation and disambiguation. In systems for restricted tasks the user utterances might be so short that these segmentation capabilities of prosodic information cannot lead to system improvement. For example, the average user utterance length in a field test with a travel information system was 3.5 words [9].

In the speech–to–speech translation task of VERBMOBIL the communication form is human–(computer)–human vs. human–computer in almost all other ASU application. Thus, in VERBMOBIL spontaneous, "real–life" utterances have to be processed. A corpus analysis of VERBMOBIL data, which were collected in simulated human–human dialogs, showed that about 70 % of the utterances contain more than a single sentence [25]; an utterance comprises about 20 words on the average. Furthermore, spontaneous speech phenomena like elliptical constructions and interruptions or restarts are frequent and increase the amount of ambiguities a lot.

We believe that the most important contribution of prosody lies in the understanding rather than in the recognition phase. This shows up clearly in a system like VERBMOBIL which is one of the first systems where the end–to–end performance (including a deep linguistic analysis) is the optimization criterion. The current version of the VERBMOBIL research prototype translates more than 70% approximatively correct [27].

2. THE VERBMOBIL SYSTEM

VERBMOBIL is a speech–to–speech translation project [28, 3] in the domain of appointment scheduling dialogs, i.e., two persons try to fix a meeting date, time, and place. Currently the emphasis lies on the translation of German utterances into English. In October 1996 a research prototype was successfully presented to the public; an overview of the architecture of this VERBMOBIL prototype is shown in Figure 1. After the recording of the spontaneous utterance a word hypotheses graph (WHG) is computed by a standard HMM word recognizer and enriched with prosodic information (cf. Section 3.). The WHG is parsed by one of the two alternative syntactic modules, i.e., the best scored syntactically correct word chain together with its different possible parse trees (readings) is passed to the semantic analysis. Also governed by the dialog module, the utterance is translated on the semantic level (transfer module) and an English utterance is generated and synthesized. Parallel to the *deep* analysis performed by these modules, the dialog module conducts a *shallow* processing, i.e., the important dialog acts are detected in the utterance and are roughly translated. A more detailed account of the architecture can be found in [7].

Figure 1 shows the interaction of the prosody module with the other modules in the VERBMOBIL architecture. The solid lines point out interfaces and the dashed lines mark additional flow of information. For the time being, the following modules use the prosodic information: syntactic analysis, semantic construction, dialog processing, transfer, and speech synthesis. In the remainder of this paper, we will first describe the computation of prosodic information and then discuss how this information is used by the other modules.

3. THE COMPUTATION OF PROSODIC INFORMATION

Input to the module is the word hypotheses graph and the speech signal. Output is a prosodically scored word hypotheses graph [16], i.e., to each of the word hypotheses, probabilities for prosodic accent, for prosodic clause boundaries, and for sentence mood are attached. The computation of prosodic information is described in more detail in [11, 12]. The use of this information on the basis of word graphs in the VERBMOBIL system is described in detail in [13].

[1]This work was funded by the German Federal Ministry of Education, Science, Research and Technology (*BMBF*) in the framework of the VERBMOBIL Project under Grants 01 IV 102 H/0 and 01 IV 102 F/4. The responsibility for the contents of this study lies with the authors. We wish to thank all VERBMOBIL partners who integrated the prosodic information into their analysis modules.

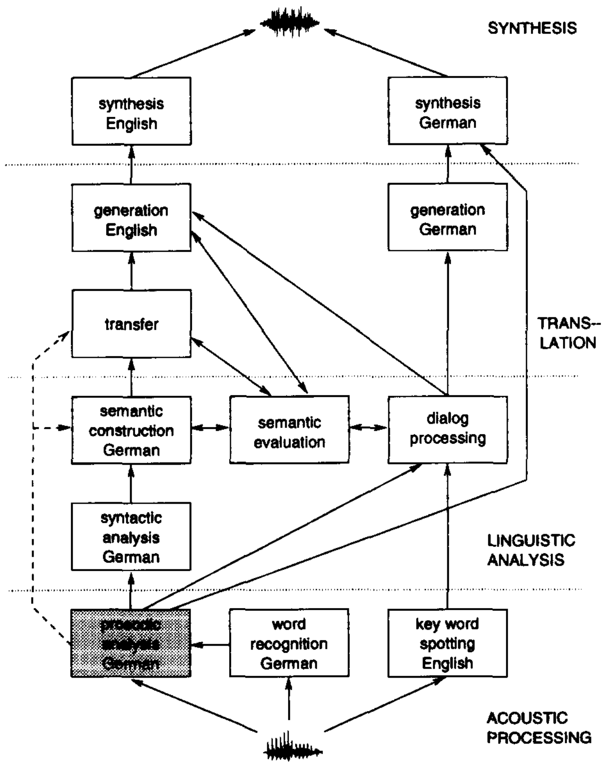

Figure 1. The VERBMOBIL architecture at a glance.

Based on the speech signal, the F0 and loudness contours are computed. Then for each of the word hypotheses a time–alignment of the corresponding phonemes according to the standard pronunciation is performed. This also results in a segmentation of the speech signal into syllable segments given a specific word hypothesis. For the computation of prosodic features for each word hypothesis pointers to the optimal predecessor/successor are established using Viterbi search. Then, for each word hypothesis the following types of features are computed based on the surrounding context (± 2 words as well as $\pm$ syllables and syllable nuclei with respect to the word final syllable): the relative duration [30]; features describing F0 and energy contours like regression coefficients, minima, maxima, and their relative positions; the length of the pause (if any) after and before the word; the speaking rate; flags indicating word finality and lexical word accent. For an evaluation and a more detailed description of the different types of features cf. [11, 12]. The optimal set of features for each of the classification tasks has been determined using feature selection methods [19]. For boundary classification 121 features and for accent classification 113 features are used. For the moment, sentence mood classification is done using up to 14 features derived from the F0 contour. For each of the three tasks a separate multi–layer perceptron (MLP) is used. We were able to achieve better recognition rates with an MLP than with more traditional classifiers like Gaussian or polynomial classifiers as used in [14].

In the case of boundary recognition, the MLP classifier is combined with a category based n-gram which models the probability of a clause boundary given a few words in the context [16]. Currently, a trigram is used so that the context is limited to ± 2 words. A drastic improvement in recognition rate could be achieved by using syntactic–prosodic labels instead of perceptually created labels for training [2]. This is due to the fact that the syntactic–prosodic labels are based only on the transliteration of the utterances and therefore large amounts of training data could be made available.

Other researchers used classification trees (CT) for the classification of prosodic boundaries based on word chains [29]; we achieved better results with n-grams than with CTs [20].

4. THE USE OF PROSODIC INFORMATION

In the following we will describe the use of prosodic information by some of the other modules in the VERBMOBIL system.

Syntactic analysis:
There are two reasons, why syntax heavily depends on prosody: First, to ensure that most of the spoken words are recognized, for spontaneous speech a large word hypotheses graph has to be generated. Currently, word hypotheses graphs of about 10 hypotheses per spoken word are generated. Finding the correct (or approximately correct) path through a word hypotheses graph is thus an enormous search problem. Second, even if the spoken word sequence has been recovered by word recognition correctly, there still might be many different parses possible, due to the high number of ambiguities contained in spontaneous speech and due to the relatively long sentences occurring in the VERBMOBIL domain.

Consider the following two different syntactic readings for an identical word sequence taken from the VERBMOBIL domain where only the sentence boundaries disambiguate between the two different syntactic structures, their semantic meanings, and their pragmatic interpretations.

(1) *"Vielleicht. Am Montag bei mir. Paßt das?"*
"Maybe. On Monday, at my place. Is that OK?"

(2) *"Vielleicht am Montag. Bei mir paßt das."*
"Maybe on Monday. That's possible for me."

Both VERBMOBIL syntax modules use the clause boundary scores of the prosody module along with the acoustic score of the word hypotheses and n–gram stochastic language models, to preselect among the many combinatorically possible paths through the word graph. These preselected word chains (which contain information about sentence boundaries) are then analyzed using a Trace Unification Grammar (TUG) in the syntax module from Siemens [4, 3] and a Head–driven Phrase Structure Grammar (HPSG) in the syntax module from IBM [21].

The Siemens parser is combined with an A*–search and operates directly on WHGs [24]. The grammar contains a *prosodic syntactic clause boundary* symbol (PSCB). Word chains starting at the first node of the WHG and ending somewhere in the WHG together with the partial syntactic analyses build the hypotheses of the search. At each step of the search the best scored hypothesis is taken from the agenda and partially analyzed. If the parse succeeds, the hypothesis is extended according to the WHG and also by the PSCB symbol. Each of these newly created hypotheses is scored with respect to the acoustic score of the words, a trigram language model for word sequences, and the prosodic score for PSCB or $\neg$PSCB. In this way the prosodic information "rules out" unlikely hypotheses without making hard decisions, i.e., certain hypotheses get a bad score due to the prosodic information so that they are rarely considered any further during the analysis, but they might still be considered in the case the prosody module makes an error. The use of prosody in this parser is described in more detail in [15].

In the IBM module preselection and deep analysis are done sequentially: First the n-best word chains are extracted from a WHG. A pair of such word chains differs in the words and/or in the position of a PSCB symbol and/or in the position of the empty element. In a German main clause the verb is usually in second position, whereas in a subordinate clause it is in final position, where the "final" position does not necessarily coincide with the end of the sentence. The empty element in verb–second sentences takes the position where the verb would be in verb–final sentences.

Determining this position is highly ambiguous and is supported by prosodic boundary information. The use of prosody in this syntax module is described in more detail in [1].

Semantic construction:
The VERBMOBIL semantic module receives a parse tree, the underlying word chain and the prosodic scores for accentuation from the syntax module. Based on these, underspecified *Discourse Representation Structures* (DRS) [10, 6] are created. These yield assertions, representing the direct meaning of a sentence, and presuppositions. If several DRS are plausible due to ambiguities, accent information is used to rule out the wrong DRS. Context information might also be used to disambiguate the interpretation, however, prosodic information can be utilized at much lower cost [5]. This use of prosody can be illustrated by the following examples from the VERBMOBIL corpus where the meaning of both sentences is the same. However, the position of the primary accent changes the scope and thereby the presupposition of the utterances, which results in a different translation of the particle *noch* (*still, another*).

(3) "*Dann müssen wir noch einen <u>Termin</u> ausmachen.*"
"*Then we still have to fix a <u>date</u>.*"

(4) "*Dann müssen wir <u>noch</u> einen Termin ausmachen.*"
"*Then we have to fix <u>another</u> date.*"

Dialog processing:
One of the tasks of the dialog module [22] is to keep track of the state of the dialog in terms of dialog acts. Dialog act recognition is done by statistical classifiers. Dialog acts are, e.g., *greeting, confirmation of a date, suggestion of a place*. In VERBMOBIL, a turn of a user can consist of more than one dialog act. Currently, the processing is done in two steps: First, the best path in the WHG (extracted by a Viterbi search using acoustic and trigram scores) is segmented into dialog act units. Second, these units are classified into dialog acts. For the segmentation into dialog acts we use the same prosodic clause boundary information as used by the syntax modules. Due to less amount of training data the use of a different classifier trained directly on dialog act boundaries did not improve the recognition rate. Further details can be found in [13, 18].

Transfer:
The transfer module of the VERBMOBIL system translates DRS representing the semantic information underlying the utterance into DRS corresponding to English sentences [8]. This task might involve pragmatic analysis and disambiguation which is partly done by the semantic evaluation module. The transfer module uses accent and sentence mood information for a few tasks. The sentence mood information is used to distinguish between questions and non–questions if grammatical indicators are missing; e.g., questions and declaratives with topic elision can have an identical word order. The accent information disambiguates mainly the interpretation of particles. In the following examples, the same word chain has different meanings depending on whether the accent is on *schon* or on *finde*. For further use of prosodic information in the VERBMOBIL transfer module cf. [23].

(5) "*Finde ich <u>schon</u>.*" "*I really believe that.*"
(6) "*<u>Finde</u> ich schon.*" "*I'll find it certainly.*"

Speech synthesis:
For a better user acceptance, the synthesized output of a translation system should be adapted to the voice of the original speaker (especially in a multi–party scenario). With respect to prosody this means that parameters like the pitch level and the speaking rate should be adapted. So far, the speech synthesis of the VERBMOBIL system is only switched to a male or a female voice according to the F0 contour of the original user utterance.

task	% recognized
clause boundary vs. no–boundary	94%
accented vs. not–accented word	83%

Table 1. Results of the prosody module.

	without prosody	with prosody	improvement
# readings	137.7	5.6	96%
parse time (secs)	38.6	3.1	92%

Table 2. Results of the Siemens word graph parser.

5. EXPERIMENTS AND RESULTS

Table 1 shows the most important results of the prosody module in isolation. These are obtained by classifying and evaluating the spoken word chains, i.e., simulating 100% correct word recognition. When moving to WHGs the recognition rate for boundaries drops by about 2 percent points. Note that the recognition rate for boundaries in the table refers to the combination of MLP and n-gram. The MLP alone, i.e. pure acoustic–prosodic classification, yields a recognition rate of 86%. The recognition rates were obtained on real VERBMOBIL spontaneous speech data.

The usefulness of prosodic information in the different modules of VERBMOBIL could be demonstrated at the press conference in October 1996 in Munich [27]. So far systematic evaluations of the improvement of a linguistic module by using prosodic information has only been done with the two syntax modules. Table 2 shows the improvement of the Siemens WHG parser by using the prosodic clause boundary probabilities. It can be seen that the number of readings as well as the parse time are drastically reduced. These results were obtained on 594 real spontaneous speech utterances. These utterances are independent from any training material used for the prosody module as well as from testing material used for the improvement of grammar and parser. For the IBM parser results are only available for speech recorded during tests with the VERBMOBIL system by non–naive users. With this material a speed–up of 46% was achieved by using the prosodic clause boundary information.

6. CONCLUSION AND FUTURE WORK

Apart from the still missing systematic evaluation in many cases, a drawback of the realization of the prosody module results from the strictly sequential bottom–up processing. The syntax module, e.g., uses prosodic scores in the analysis of a path in a WHG. However, the computation of these scores might be based on context words not included in the path under investigation by the syntax. Therefore, our current computation makes assumptions leading to errors. In the future we plan a higher integration of the modules. The first step will be the integration of the n-gram directly in the A^*–search of the parser. This can be done without any extra computational costs. Next, the acoustic–prosodic classification might also be integrated in the A^*–search as a procedure call. In this case the trade–off between higher computation time and reduction in errors has to be carefully investigated.

With respect to accent recognition we currently work on a scheme for generating accent reference labels based on transliterations. We want to use them for n-gram training and expect an improvement similar to the boundary classification task.

In the dialog module the prosodic information might also be used for dialog act classification, and classification and segmentation of dialog acts will be integrated within a search procedure.

Additionally, within the framework of VERBMOBIL, for the next years it is planned to extend the prosodic processing to different signal qualities (mobile telephone) as well as to other languages like English and Japanese.

REFERENCES

[1] A. Batliner, A. Feldhaus, S. Geissler, A. Kießling, T. Kiss, R. Kompe, and E. Nöth. Integrating Syntactic and Prosodic Information for the Efficient Detection of Empty Categories. In *Proc. of the Int. Conf. on Computational Linguistics*, volume 1, pages 71–76, Kopenhagen, 1996.

[2] A. Batliner, R. Kompe, A. Kießling, H. Niemann, and E. Nöth. Syntactic–prosodic Labelling of Large Spontaneous Speech Data–bases. In *Int. Conf. on Spoken Language Processing*, volume 3, pages 1720–1723, Philadelphia, 1996.

[3] H.U. Block. The Language Components in Verbmobil. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, München, 1997.

[4] H.U. Block and S. Schachtl. Trace & Unification Grammar. In *Proc. of the Int. Conf. on Computational Linguistics*, volume 1, pages 87–93, Nantes, 1992.

[5] J. Bos. Personal communication, July 1996.

[6] J. Bos, B. Gambäck, Ch. Lieske, Y. Mori, M. Pinkal, and K. Worm. Compositional Semantics in Verbmobil. In *Proc. of the Int. Conf. on Computational Linguistics*, volume 1, pages 131–136, Kopenhagen, 1996.

[7] T. Bub and J. Schwinn. Verbmobil: The Evolution of a Complex Large Speech-to-Speech Translation System. In *Int. Conf. on Spoken Language Processing*, volume 4, pages 1026–1029, Philadelphia, 1996.

[8] K. Eberle. Disambiguation by Information Structure in DRT. In *Proc. of the Int. Conf. on Computational Linguistics*, volume 1, pages 334–339, Kopenhagen, 1996.

[9] W. Eckert, E. Nöth, H. Niemann, and E. Schukat-Talamazzini. Real Users Behave Weird — Experiences made collecting large Human–Machine–Dialog Corpora. In P. Dalsgaard, L.B. Larsen, L. Boves, and I. Thomsen, editors, *Proc. of the ESCA Tutorial and Research Workshop on Spoken Dialogue Systems*, pages 193–196, Vigsø, Denmark, 1995. ESCA.

[10] H. Kamp and U. Reyle. *From Discourse to Logic and DRT; An Intorduction to Modeltheoretic Semantics of Natural Language*. Kluwer, Dordrecht, 1993.

[11] A. Kießling. Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung. Dissertation. Technische Fakultät der Universität Erlangen–Nürnberg, 1996.

[12] A. Kießling, R. Kompe, A. Batliner, H. Niemann, and E. Nöth. Classification of Boundaries and Accents in Spontaneous Speech. In R. Kuhn, editor, *Proc. of the CRIM / FORWISS Workshop*, pages 104–113, Montreal, 1996.

[13] R. Kompe. Prosody in Speech Understanding Systems. Dissertation. Technische Fakultät der Universität Erlangen–Nürnberg, 1996.

[14] R. Kompe, A. Batliner, A. Kießling, U. Kilian, H. Niemann, E. Nöth, and P. Regel-Brietzmann. Automatic Classification of Prosodically Marked Phrase Boundaries in German. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 173–176, Adelaide, 1994.

[15] R. Kompe, A. Kießling, H. Niemann, E. Nöth, A. Batliner, S. Schachtl, T. Ruland, and H.U. Block. Improving Parsing of Spontaneous Speech with the Help of Prosodic Boundaries. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, München, 1997.

[16] R. Kompe, A. Kießling, H. Niemann, E. Nöth, E.G. Schukat-Talamazzini, A. Zottmann, and A. Batliner. Prosodic Scoring of Word Hypotheses Graphs. In *Proc. European Conf. on Speech Communication and Technology*, volume 2, pages 1333–1336, Madrid, 1995.

[17] W. Lea. Prosodic Aids to Speech Recognition. In W. Lea, editor, *Trends in Speech Recognition*, pages 166–205. Prentice–Hall Inc., Englewood Cliffs, New Jersey, 1980.

[18] M. Mast, R. Kompe, S. Harbeck, A. Kießling, H. Niemann, and E. Nöth. Dialog Act Classification with the Help of Prosody. In *Int. Conf. on Spoken Language Processing*, volume 3, pages 1728–1731, Philadelphia, 1996.

[19] H. Niemann. *Pattern Analysis and Understanding*, volume 4 of *Series in Information Sciences*. Springer–Verlag, Heidelberg, 1990.

[20] E. Nöth, R. De Mori, J. Fischer, A. Gebhard, S. Harbeck, R. Kompe, R. Kuhn, H. Niemann, and M. Mast. An Integrated Model of Acoustics and Language Using Semantic Classification Trees. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 419–422, Atlanta, 1996.

[21] C. Pollard and I. Sag. *Information–based Syntax and Semantics, Vol. 1*, volume 13 of *CSLI Lecture Notes*. CSLI, Stanford, CA, 1987.

[22] N. Reithinger, E. Maier, and J. Alexandersson. Treatment of Incomplete Dialogues in a Speech–to–speech Translation System. In P. Dalsgaard, L.B. Larsen, L. Boves, and I. Thomsen, editors, *Proc. of the ESCA Tutorial and Research Workshop on Spoken Dialogue Systems*, pages 33–36. ESCA, Vigsø, Denmark, 1995.

[23] B. Ripplinger and J. Alexandersson. Disambiguation and Translation of German Particles in Verbmobil, Verbmobil Memo 70, 1996.

[24] L.A. Schmid. Parsing Word Graphs Using a Linguistic Grammar and a Statistical Language Model. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 41–44, Adelaide, 1994.

[25] H. Tropf. Spontansprachliche syntaktische Phänomene: Analyse eines Korpus aus der Domäne "Terminabsprache". Technical report, Siemens AG, ZFE ST SN 54, München, 1994.

[26] J. Vaissière. The Use of Prosodic Parameters in Automatic Speech Recognition. In H. Niemann, M. Lang, and G. Sagerer, editors, *Recent Advances in Speech Understanding and Dialog Systems*, volume 46 of *NATO ASI Series F*, pages 71–99. Springer–Verlag, Berlin, 1988.

[27] W. Wahlster. Presseerklärung zum Verbmobil-Forschungsprototypen am 25.10.1996 in München, 1996. http://www.dfki.uni-sb.de/verbmobil.

[28] W. Wahlster, T. Bub, and A. Waibel. Verbmobil: The Combination of Deep and Shallow Processing for Spontaneous Speech Translation. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, München, 1997.

[29] M.Q. Wang and J. Hirschberg. Automatic Classification of Intonational Phrase Boundaries. *Computer Speech & Language*, 6(2):175–196, 1992.

[30] C.W. Wightman and M. Ostendorf. Automatic Labeling of Prosodic Patterns. *IEEE Trans. on Speech and Audio Processing*, 2(3):469–481, 1994.

