BMBF

Verbmobil

Verbundvorhaben

PROSODIC SCORING OF WORD HYPOTHESES GRAPHS

R. Kompe, A. Kießling,
H. Niemann, E. Nöth,
E.G. Schukat–Talamazzini,
A. Zottmann, A. Batliner

F.–A.–Universität Erlangen–Nürnberg

L.M.-Universität München

Vm

Report 90
Juli 1995

Juli 1995

R. Kompe, A. Kießling,
H. Niemann, E. Nöth,
E.G. Schukat–Talamazzini,
A. Zottmann, A. Batliner

Lehrstuhl für Mustererkennung (Inf. 5)
Friedrich–Alexander–Universität Erlangen–Nürnberg
Martensstr. 3
D–91058 Erlangen

Institut für Deutsche Philologie
Ludwig–Maximilian Universität München
Schellingstr. 3
D–80799 München

Tel.: (09131) 85 - 7890
e-mail: `kompe@informatik.uni-erlangen.de`

Gehört zum Antragsabschnitt: 3.11, 3.12, 6.4

Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministers für Bildung, Wissenschaft, Forschung und Technologie unter dem Förderkennzeichen 01 IV 102 H/0 und 01 IV 102 F/4 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den Autoren.

PROSODIC SCORING OF WORD HYPOTHESES GRAPHS

R. Kompe[1] *A. Kießling*[1] *H. Niemann*[1] *E. Nöth*[1] *E.G. Schukat–Talamazzini*[1] *A. Zottmann*[1] *A. Batliner*[2]

[1]Univ. Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Inf. 5), Martensstr. 3, 91058 Erlangen, F.R. of Germany

[2]L.M.-Universität München, Institut für Deutsche Philologie, Schellingstr. 3, 80799 München, F.R. of Germany

E–mail: kompe@informatik.uni-erlangen.de
www: http://www5.informatik.uni-erlangen.de

ABSTRACT

Prosodic boundary detection is important to disambiguate parsing, especially in spontaneous speech, where elliptic sentences occur frequently. Word graphs are an efficient interface between word recognition and parser. Prosodic classification of word chains has been published earlier. The adjustments necessary for applying these classification techniques to word graphs are discussed in this paper. When classifying a word hypothesis a set of context words has to be determined appropriately. A method has been developed to use stochastic language models for prosodic classification. This as well has been adopted for the use on word graphs. We also improved the set of acoustic–prosodic features with which the recognition errors were reduced by about 60% on the read speech we were working on previously, now achieving 10% error rate for 3 boundary classes and 5% for 2 accent classes. Moving to spontaneous speech the recognition error increases significantly (e.g. 16% for a 2–class boundary task). We show that even on word graphs the combination of language models which model a larger context with acoustic–prosodic classifiers reduces the recognition error by up to 50%.

1. INTRODUCTION

In automatic speech understanding systems prosody can be used to disambiguate during syntactic analysis or semantic interpretation [14] or it can be used to guide dialog control [7, 6]. The research presented in this paper has been conducted under the Verbmobil project (henceforth VM, cf. [15]) which aims at automatic speech–to–speech translation in appointment scheduling dialogs. Currently, we concentrate our efforts in prosody on the recognition of clause boundaries and of accentuated words.

The clause boundaries are used for disambiguation during parsing. In general and in the VM corpora as well spontaneous speech contains many elliptic sentences. Thus, it is very important to reduce the search space during parsing by the means of prosodic clause boundaries. The following sentence is a typical example taken from the VM corpora:

ja | zur Not | geht's | auch | am Samstag |

The vertical bars indicate possible positions for clause boundaries. In written language most of these bars can be substituted by either comma, period or question mark. In total there exist at least 36 different syntactically correct alternatives for putting the punctuation marks. The following examples show two of these alternatives together with a translation into English:

1. *Ja? Zur Not geht's? Auch am Samstag?*
   Really? It's possible if necessary? Even on Saturday?
2. *Ja. Zur Not. Geht's auch am Samstag?*
   Yes. If necessary. Would Saturday be possible as well?

The position of the phrasal accent is used for disambiguation as well. Consider the following example:

Dann müssten wir noch einen Termin ausmachen.

If the phrase accent is on the particle *'noch'*, an appropriate translation would be:

Then we need another meeting date.

The default position of the phrasal accent is, however, on *'Termin'* with the following translation:

Then we still need a meeting date.

In [10] the automatic detection of phrase boundaries has been successfully used to rescore the n–best sentence hypotheses computed by a word recognizer, i.e. the quality of the best sentence hypothesis has been improved by the rescoring. In the VM project the interface between word recognition and parsing is a word graph, which is a compact representation of n–best word chains. The parser is integrated in an A^*–search for the best path in the word graph [12]. In this way, parsing is very efficient, because when using n–best word chains the same partial chains have to be parsed repeatedly. The overhead for the A^*–search is neglectable. During the A^*–search, the partial parses are scored by combining the scores of the acoustic models, a language model, and the prosody module [1]. Because the search space is very large the prosody module cannot compute its scores based on the word chains underlying partial parses, but it has to score the word graph prior to the syntactic analysis.

First promising results concerning the use of prosodic clause boundaries during parsing of word graphs have been presented in [1] for a train time table inquiry task using read speech. In this paper we will focus on the implications of scoring word graphs vs. word chains. Experimental results are presented for the spontaneous VM speech data.

The paper is organized as follows: Firstly, the speech data is specified; secondly, the methods used in the experiments are described including the training of the acoustic–prosodic model, the algorithm for polygram–classification, and scoring word graphs prosodically. Finally, experimental results are given.

2. MATERIAL

The ERBA material has already been described in [5, 2]; as in these studies 6,900 utterances were used for training and 2,100 utterances were used for testing.

For VM there are 25 dialogs labelled prosodically. Out of these we chose 22 for training (592 utterances, 32 different speakers, 71 minutes of speech, 9336 words), and 3 for testing (80 utterances, 4 different speakers). For these 80 utterances word graphs of approx. 19 words per spoken word (not counting non–verbals and pauses) were generated with our word recognizer [13]. 48 of these word graphs contained the spoken word chain[2]. These make up the test set for all VM evaluations described in this paper (cf. Section 4.3)[3]. These 48 utterances consist of 237 seconds of speech, 520 words.

[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.

[2]The word accuracy for the VM data is in the order of 70%. For the generation of these word graphs the bigram language model for the word recognizer has been trained on and thus restricted to the 80 test sentences. The acoustic models were trained on a larger corpus.

[3]The recognition rates on the spoken word chain are of the same order as in Section 5 when using leave–one–out mode for training/testing.

	ERBA	VM
$B9$	*not used*	irregular boundary, mostly hesitation lengthening
$B3$	sentence or clause boundary	main phrase boundary
$B2$	constituent boundary prosodically marked	intermediate phrase boundary
$B1$	constituent boundary not prosodically marked	*not used*
$B0$	every other word boundary	every other word boundary
$A4$	*not used*	emphatic or contrastive accent
$A3$	sentence accent (1 per $B3$ phrase)	*not used*
$A2$	primary phrase accent (1 per $B2$ phrase)	primary phrase accent
$A1$	secondary phrase accent (1 per $B1$ phrase)	secondary phrase accent
$A0$	unaccentuated syllable	unaccentuated syllable

Table 1. Definition of the prosodic labels

For VM the prosodic reference labels are based on perceptive evaluation done by non–naive listeners [11]. For ERBA they were created automatically based on rules using linguistic knowledge and expectations about prosodic marking. Listening experiments showed a high agreement with the automatically created labels [2]. Thus there is a rough correspondence between the reference labels for ERBA and VM. A list of all the labels is given in Table 1. The VM test utterances contain 74 $B3$, 36 $B2$, 13 $B9$, and 349 $B0$ boundaries not counting the end of utterances. They also contain 243 accentuated words.
Note, that the boundary labels are attached to word boundaries and the accent labels to each of the syllables in the spoken words.

3. ACOUSTIC–PROSODIC FEATURES

The computation of the features is based on a time alignment of the words on the phoneme level computed during word recognition. For each syllable to be classified the following prosodic features were computed from the speech signal for the syllable under consideration and for the six syllables in the left and the right context:

- the normalized duration of the syllable nucleus [16]
- the F0 minimum, maximum, onset, and offset and the maximum energy and their positions on the time axis relative to the position of the actual syllable
- the mean energy, and the mean F0
- flags indicating if the syllable carries the lexical word accent or if it is in a word final position

Furthermore the following features were computed only for the syllable under consideration:

- the length of the pause (if any) preceding or succeeding the word containing the syllable
- the linear regression coefficients of the F0–contour and the energy contour computed over different 15 windows to the left and to the right of the syllable

This yields a total of 242 features. The feature set probably contains useless or redundant features, but to our experience this does not hurt the classification performance of the neural networks provided enough training data. In [4] the contribution of different groups of features to the classification results was investigated.

4. METHODS

4.1. Training of the acoustic–prosodic model

Multi–layer perceptrons (MLP) were trained using Quickpropagation to classify the features described in Section 3. Training is based on the time alignment of the spoken word chain, which was computed with our hidden Markov model (HMM) word recognizer [13]. For the experiments on the VM data, MLPs with 40/20 nodes in the first/second hidden layer were used. For ERBA, where more training data is available, a MLP with 60/30 nodes in the first/second hidden layer was used. The MLPs have one output node per class.
For ERBA one MLP was trained to distinguish between the six classes $A0B01$, $A0B2$, $A0B3$, $A123B01$, $A123B2$, $A123B3$. All the 242 features described in Section 3 were used as input. This MLP was used separately for boundary and accent classification. In both cases the MLP outputs were added appropriately.
Since for VM much less training data was available, we used different subsets of the prosodic features of Section 3, and we trained separate MLPs for boundary and accent classification: one MLP distinguishes between $A0$ and $A124$, another between $B0$, $B2$, $B3$, and $B9$, and a third one between $B029$ and $B3$.
In the following, we assume that the MLP computes a posteriori probabilities. However, in order to balance for the a priori probabilities of the different classes, during training the MLP was presented with an equal number of feature vectors from each class. Furthermore, the sum of the MLP outputs was normalized to be equal to one, though we observed that in most cases the sum is close to one.

4.2. Polygram–classification

In [5] we already showed that a combination of an acoustic–prosodic classifier for phrase boundaries with a stochastic language model improves the recognition rate. At that time we worked on the spoken word chain. In the following a modification of the language models is proposed, so that they can be used for classification on the basis of word graphs.
Let w_i be a word out of a vocabulary where i denotes the position in the utterance; v_i denotes a symbol out of a predefined set V of prosodic symbols. These can be for example $\{B01, B2, B3\}$, $\{A01, A23\}$ or a combination of both $\{B01A01, B01A23, \ldots, B3A23\}$ depending on the specific classification task (cf. Section 2). For example $v_i = B01$ means that the i^{th} word in an utterance is succeeded by the prosodic label $B01$ (i.e., no prosodic boundary), and $v_i = A23$ means that the i^{th} word is accentuated.
Ideally one would like to model the following a priori probability

$$P(w_1 v_1 w_2 v_2 \ldots w_m v_m)$$

which is the probability for strings, where words and prosodic labels alternate (m is the number of words in the utterance).
In [5] we used a language model similar to this one to score chains containing words and prosodic labels. In the following, we are interested in the recognition of prosodic classes given a (partial) word chain (which in the case of word graphs is obtained from the best path through the word hypothesis to be classified). When determining the appropriate label to substitute v_i the labels at positions v_{i-k} and v_{i+k} are not known ($k = 1, 2, \ldots$). Thus, we used the following probabilities:

$$P(w_1 \ldots w_i v_i w_{i+1} \ldots w_m) = P_l P_v P_r \qquad (1)$$

where P_l, P_v, and P_r are defined as follows:

$$\begin{aligned} P_l &= P(w_1)P(w_2|w_1) \cdot \ldots \cdot P(w_i|w_1 \ldots w_{i-1}) && (2)\\ P_v &= P(v_i|w_1 \ldots w_i) && (3)\\ P_r &= P(w_{i+1}|w_1 \ldots w_i v_i) \\ &\quad \cdot \ldots \cdot P(w_m|w_1 \ldots w_i v_i w_{i+1} \ldots w_{m-1}) && (4) \end{aligned}$$

Terms like $w_1 \dots w_i$ in $P(v_i|w_1 \dots w_i)$ are called *history*. As usual in stochastic language modelling the history has to be restricted to a certain length [8]. The stochastic language model approach we used is the so called *polygram* [13], where the histories have variable length depending on the available training data. A maximum history length can be defined.
For each word boundary in the training corpora of ERBA and VM a sufficient number of context words (according to the maximum history length) and the corresponding prosodic reference label are extracted from the text corpora and used to estimate the probabilities of the equations above by counting the frequencies (maximum likelihood estimation) as usually done when training stochastic language models. In fact, not the above probabilities are used, but the words are put into categories. In the case of ERBA only the names of train stations, days of the week, month names, ordinal numbers, and cardinal numbers are put into categories. All other 392 words are not grouped into categories. In the case of VM all the 1186 words were put into 150 categories.
We used the so trained polygrams for the classification of prosodic labels. Given a word chain $w_1 \dots w_i \dots w_m$, the appropriate prosodic class v_i^* is determined by maximizing the probability of equation 1:

$$v_i^* = \underset{v_i \in V}{\operatorname{argmax}} \; P(w_1 \dots w_i v_i w_{i+1} \dots w_m)$$

Note, that the probability P_l is independent of v_i (equation 2). Thus this maximization (and v_i^*) are independent from P_l. Note also, that v_i^* does not only depend on the left context (probability P_v, equation 3) but also on the words succeeding the word w_i (probability P_r, equation 4). In practice, the context is restricted to the maximum history length H_L used during training of the polygram:

$$v_i^* = \underset{v_i \in V}{\operatorname{argmax}} \; P(w_{i-H_L} \dots w_i v_i w_{i+1} \dots w_{i+H_L}) \qquad (5)$$

4.3. Prosodic scoring of word graphs

A word graph is a directed acyclic graph [9]. Each edge corresponds to a word hypothesis which is attached with the acoustic probabilities, the first and the last time frame, and a time alignment of the underlying phoneme sequence. The graph has a single start node (corresponding to time frame 1) and a single end node (the last time frame in the signal). Each path through the graph from the start to the end node forms a sentence hypothesis. Each edge in the graph lies on at least one such path. In the following the term *neighbors* of a word hypothesis in a graph refers to all its predecessor and successor edges.
With *prosodic scoring of word graphs* we mean in fact the annotation of the word hypotheses in the graph with the probabilities for the different prosodic classes. These probabilities are used by the other modules (e.g. parsing) of a speech understanding system. Note, that also in the case of phrase boundaries we do not compute the probability for a prosodic boundary located at a certain node in the word graph, but for each of the word hypotheses in the graph the probability for a boundary being after this word is computed. This is important, since the acoustic–prosodic features also include the duration of syllable nuclei; these are most robustly obtained from the time alignment of the phoneme sequence underlying a word hypothesis computed with the word recognizer, and these durations have to be normalized with respect to the intrinsic phoneme duration.
The following steps have to be conducted for each word hypothesis w_i:

1. determine recursively appropriate neighbors of the word hypothesis until a word chain $w_{i-k} \dots w_{i+l}$ is built which contains enough syllables to compute the acoustic–prosodic feature vector and where $k \geq H_L$, $l \geq H_L$.
2. for each $v_i \in V$ and for each syllable s in the word w_i compute the probabilities

$$P_{sv_i} = \frac{Q_{sv_i}}{\sum_{v_i \in V} Q_{sv_i}} \qquad \text{where}$$

$$Q_{sv_i} = P(v_i|c_{is}) P^{\xi}(w_{i-H_L} \dots w_i v_i w_{i+1} \dots w_{i+H_L})$$

Note, that in the case of boundaries only the word final syllable is considered.

c_{is} denotes the acoustic–prosodic feature vector, ξ is a weight for the combination of the acoustic–prosodic model probability ($P(v_i|c_{is})$ computed by the MLP) and the language model probability; its value has been determined empirically. Different values were used for the different classification experiments described in Section 5.
In the current implementation we just select the hypothesis which is most probable according to the acoustic model as the "appropriate" neighbor of w_i. Note, that this is suboptimal, because the context words may differ from the spoken words. An exact solution would be a weighted sum of all probabilities P_{sv_i} computed on the basis of all the possible contexts. However, this does not seem to be feasible under real–time constraints. As a trade–off the neighbors could be determined on the basis of the best of the paths through the graph which contain the hypothesis w_i. The best path could be determined efficiently with dynamic programming using acoustic and language model scores.
The duration of a syllable nucleus should be normalized with respect to the average speaking rate, which is the reciprocal of the average of the intrinsically normalized phoneme durations [16]. The speaking rate has to be determined on the basis of the word graph:

- the *local speaking rate* is determined on the neighbors of word w_i
- the *global speaking rate* is the average of the speaking rates of all hypotheses in the graph weighted by their acoustic scores.

All the experiments described below were obtained by using the global speaking rate, for which - compared to the local speaking rate - slightly better results could be observed for experiments based on the word chain; whereas on word graphs the local speaking rate performs slightly better.
The evaluation of the prosodic scores only makes sense on the word graphs containing the spoken word chain:

1. score the word graph prosodically with the probabilities P_{sv_i}. Note, that this is based on the best paths through the hypotheses which may be different from the spoken word chain
2. for each word contained in the (best) path corresponding to the spoken word chain and in the case of accent classification for each syllable in the word determine the prosodic class with the largest probability P_{sv_i} (i.e. the recognized class)
3. compare the recognized classes with the reference labels and determine the recognition error

5. EXPERIMENTAL RESULTS

In Tables 2–5 the recognition rates for different experiments on ERBA and VM are presented. LM$_h$ denotes the polygram–classification as described in Section 4.2, where h specifies the maximum context allowed during training of the polygram. LM$_{bigram}$ denotes the probabilities $P(v_i|w_i)$, i.e. no (suboptimal) context has been used. The columns 'word chain' refer to experiments conducted on the time alignment of the spoken word chain.
On ERBA we could improve the recognition rate for accentuated vs. non–accentuated syllables with respect to [4] from 88.7% to 94.9%. The results for the boundary recognition are given in Table 2: using the MLP the recognition rate could be improved from 75.7% to 90.3% in comparison to [4]; the main improvement results from the additional use of syllable nucleus durations of the context. With the polygram classifier alone a recognition rate of 99.3 could be achieved. This surprisingly good result

	average	$B01$	$B2$	$B3$
MLP	90.3	89.8	92.2	90.6
LM_1	97.7	98.4	95.9	93.9
LM_3	99.3	99.6	98.4	99.4

Table 2. Recognition rates in percent for $B01$ vs. $B2$ vs. $B3$ on ERBA (word chains)

	average	$B0$	$B2$	$B3$	$B9$
MLP	60.6	59.1	48.3	71.9	68.5
LM_3	82.1	95.9	11.4	59.6	28.1

Table 3. Recognition rates in percent for $B01$ vs. $B2$ vs. $B3$ vs. $B9$ on VM (word chains)

is at least partly caused by the rather restricted syntax of the ERBA material.

Table 3 shows the recognition rates for the VM 4–class boundary problem. The results seem not to be very good. Probably, the main reason for this is the small amount of available training data.

In Table 4 the recognition rates for accentuated vs. non–accentuated syllables for VM are given. The performance on word graphs is only slightly worse than on the word chain.

Currently the syntactic analysis in VM is mostly interested in probabilities for $B3$ boundaries. Thus, we performed a series of experiments the results of which are presented in Table 5. Due to the suboptimal context determined for each word hypothesis, the recognition rate drops when switching from the word chain to word graphs. Increasing the history size modeled by the polygram improves the recognition rate, even in the case of word graphs. Due to the sparse training data a history of more than 2 symbols does not change the recognition rate. A combination of both, acoustic–prosodic model (MLP) and stochastic language model (polygram LM_2) yields the best recognition rate (91.7%).

We also tested on VM the MLP trained on ERBA. The recognition rate for accents dropped by only 4%, while the one for the $B3$ boundaries dropped by 17%. We believe that this increase in error is mostly due to the differences in intonation between read and spontaneous speech.

6. CONCLUSION

We showed that the prosodic scoring of word graphs is feasible without a great reduction in recognition rate. Furthermore, a method for combining acoustic–prosodic and stochastic language model scores for prosodic classification purposes has been successfully applied on the basis of word graphs.

In preliminary parsing experiments performed by our colleagues at Siemens München [3] on word graphs computed on VM speech data, the parse time and the number of parse trees could be decreased even more than reported for the ERBA data [1] when using probabilities for $B3$ boundaries computed with the MLP described in this paper.

The prosodic scores for all of the alternative classes of interest are attached to the word hypotheses in the graph and passed to the other modules (e.g. syntactic analysis, semantic interpretation). They are supposed to use these scores rather than a single class symbol.

In this paper we did not consider the classification of sentence mood depending on the intonation contour. However, in the current version of our prosody module the classifier described in [6] is used to compute the probability for three different classes of sentence mood which is as well attached to the word hypotheses in the graph.

	word chain	word graph
MLP	82.5 (78.9)	82.0 (81.1)

Table 4. Average recognition rates in percent for $A0$ vs. $A124$ ($A124$ in parentheses) on VM

	word chain	word graph
MLP	83.5 (74.3)	78.4 (66.2)
LM_{bigram}	83.1 (14.9)	83.1 (14.9)
LM_1	88.3 (36.5)	87.3 (29.7)
LM_2	88.8 (48.6)	88.1 (43.2)
MLP + LM_2	91.7 (64.9)	90.5 (54.1)

Table 5. Average recognition rates in percent for $B029$ vs. $B3$ ($B3$ in parentheses) on VM

REFERENCES

[1] G. Bakenecker, U. Block, A. Batliner, R. Kompe, E. Nöth, and P. Regel-Brietzmann. Improving Parsing by Incorporating 'Prosodic Clause Boundaries' into a Grammar. In *Proc. ICSLP*, Vol. 3, pp. 1115–1118, Yokohama, September 1994.

[2] A. Batliner, R. Kompe, A. Kießling, E. Nöth, H. Niemann, and U. Kilian. The prosodic marking of phrase boundaries: Expectations and Results. In A. Rubio, editor, *New Advances and Trends in Speech Recognition and Coding*, NATO ASI Series F. Springer–Verlag, 1995. (to appear).

[3] U. Block. Personal communication . Siemens, Munich, 1995.

[4] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. Detection of Phrase Boundaries and Accents. In Niemann, de Mori, and Hanrieder, editors, *Progress and Prospects of Speech Research and Technology: Proc. of the CRIM/FORWISS Workshop*, infix, pp. 266–269, Sankt Augustin, 1994.

[5] 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. ICASSP*, Vol. 2, pp. 173–176, Adelaide, 1994.

[6] R. Kompe, E. Nöth, A. Kießling, T. Kuhn, M. Mast, H. Niemann, K. Ott, and A. Batliner. Prosody takes over: Towards a prosodically guided dialog system. *Speech Communication*, 15(1-2), pp 155–167, Oktober 1994.

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

[8] H. Ney, U. Essen, and R. Kneser. On Structuring Probabilistic Dependences on Stochastic Language Modelling . *Computer Speech & Language*, 8(1), pp 1–38, 1994.

[9] M. Oerder and H. Ney. Word Graphs: An Efficient Interface between Continuous–speech Recognition and Language Understanding . In *Proc. ICASSP*, Vol. 2, pp. 119–122, Minneapolis, April 1993.

[10] M. Ostendorf, C.W. Wightman, and N.M. Veilleux. Parse Scoring with Prosodic Information: an Analysis/Synthesis approach. *Computer Speech & Language*, 7(3), pp 193–210, 1993.

[11] M. Reyelt and A. Batliner. Ein Inventar prosodischer Etiketten für VERBMOBIL, Verbmobil–Memo–33–94, Juli 1994.

[12] L.A. Schmid. Parsing Word Graphs Using a Linguistic Grammar and a Statistical Language Model. In *Proc. ICASSP*, Vol. 2, pp. 41–44, Adelaide, 1994.

[13] E.G. Schukat-Talamazzini, T. Kuhn, and H. Niemann. Speech Recognition for Spoken Dialogue Systems. In Niemann, de Mori, and Hanrieder, editors, *Progress and Prospects of Speech Research and Technology: Proc. of the CRIM/FORWISS Workshop*, infix, pp. 110–120, Sankt Augustin, 1994.

[14] 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*, Vol. 46 of *NATO ASI Series F*, pp. 71–99. Springer–Verlag, 1988.

[15] W. Wahlster. Verbmobil — Translation of Face–To–Face Dialogs. In *Proc. European Conf. on Speech Communication and Technology*, Vol. "Opening and Plenary Sessions", pp. 29–38, Berlin, September 1993.

[16] C.W. Wightman. *Automatic Detection of Prosodic Constituents*. PhD thesis, Boston University, 1992.