Classification of Boundaries and Accents in Spontaneous Speech

A. Kießling, R. Kompe,
A. Batliner, H. Niemann,
E. Nöth

F.–A.–Universität Erlangen–Nürnberg
L.–M.–Universität München

Verbundvorhaben

Report 156
August 1996

August 1996

A. Kießling, R. Kompe,
A. Batliner, H. Niemann,
E. Nöth

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 - 7799
e-mail: {kiessl}@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.

To appear in *Proc. of the CRIM/FORWISS Workshop*, Montreal, Oct. 1996.

CLASSIFICATION OF BOUNDARIES AND ACCENTS IN SPONTANEOUS SPEECH

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

[1]Lehrstuhl für Mustererkennung (Informatik 5), Universität Erlangen–Nürnberg,
Martensstr. 3, D-91058 Erlangen, Germany
Phone: +49/9131/85-7799 Fax: +49/9131/303811
e-mail: kiessl@informatik.uni-erlangen.de

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

1 Introduction

The ultimate aim of automatic speech understanding (ASU) systems like dialog or translation systems is the correct recognition of the speakers intention. In order to reach this goal, not only the recognized word chain should be taken into consideration but also the prosodic information which usually is contained in every utterance. In spontaneous speech prosody can be used in many different ways for marking different kinds of prosodic events expressing different types of prosodic functions. Examples for prosodic functions that already have been investigated with regard to their use in ASU systems are prosodically marked boundaries, accents and sentence mood. The marking of spontaneous speech phenomena ,e.g., of agrammatical boundaries like hesitations, self repairs or interruptions are subject of ongoing research. In future systems, the exploitation of paralinguistic prosodic functions as, e.g., the emotional state or attitude of the speaker (doubtful, angry, happy, etc.) and the consideration of indexical (idiosyncratic or sociolinguistic) functions, like sex, age, social origin or language-specific aspects like rhythm seems to be possible.

For several reasons, the extraction of prosodic features and their classification into prosodic classes is not an easy task: Besides the fact that it is not clear at all how many prosodic classes, e.g., two, three or more boundaries, should be distinguished and have thus to be classified, the most important problems are

- the mutual influence of segmental (i.e. word chain) and suprasegmental (i.e. prosodic) information
- the interferences of the different prosodic functions which are realized to a great extent with the same prosodic parameters
- the interaction (trading relation) between prosodic parameters, where the smaller value of one parameter can be compensated by a greater value of another parameter

*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 F/4 and 01 IV 102 H/0. The responsibility for the contents lies with the authors.

- the optionality of prosodic means; a specific function *can* be expressed with prosody but it must not, e.g., when other grammatical means are already sufficient (as in Wh-questions)
- speaker and language specific use of prosodic features

For all these problems, we want to use a statistic approach and propose a method where as many relevant prosodic features as possibly are extracted over a prosodic unit, e.g., a syllable or a word, and composed into a huge feature vector which represents the prosodic properties of this and of several surrounding units in a specific context. Based on reference labeling of prosodic events which has to be synchronous with the stream of feature vectors, statistic classifiers (here: multi-layer perceptrons, MLP) are trained using a training database and evaluated on a different database. In the following we will concentrate on the word–based classification of prosodic boundaries, of syntactic-prosodic boundaries and of prosodically marked accents, but as indicated in the outlook the approach can easily be extended towards the classification of spontaneous speech phenomena.

2 Material

The research presented in this paper has been conducted under the VERBMOBIL project (cf. [17]), which aims at automatic speech–to–speech translation in appointment scheduling dialogs. All experiments reported in the following have been performed on subsets of this spontaneous speech database. For the training of the stochastic classifiers, appropriate reference labels are needed. The perceptually based prosodic labeling of boundaries and accents was performed by our VERBMOBIL partner University of Braunschweig, cf. [15, 14, 13]. Four types of word–based boundary labels are distinguished:

- B3: full boundary with strong intonational marking, often with lengthening
- B2: intermediate phrase boundary with weak intonational marking
- B0: normal word boundary (not labeled explicitly)
- B9: "agrammatical" boundary, e.g., hesitation or repair

and 4 different types of syllable based accent labels which can easily be mapped onto word–based labels denoting if a word is accented or not:

- PA: primary accent
- SA: secondary accent
- EC: emphatic or contrastive accent
- A0: any other syllable (not labeled explicitly)

In the following we are only interested in the two-class problems 'boundary' ($\mathsf{B} = \mathsf{B3}$) vs. 'no boundary' ($\neg\mathsf{B} = \{\mathsf{B0}, \mathsf{B2}, \mathsf{B9}\}$) and 'accented word' ($\mathsf{A} = \{\mathsf{PA}, \mathsf{SA}, \mathsf{EC}\}$) vs. 'not accented word' ($\neg\mathsf{A} = \mathsf{A0}$) summing up the respective classes. So far 33 VERBMOBIL dialogs (approx. 2 h of speech) have been labeled along these lines.

Alternatively, we developed a syntactic-prosodic labeling scheme [3, 4] which enables us to label large spontaneous speech corpora in a comparably short amount of time by only using the transliterated dialogs. Thus, 7286 VERBMOBIL turns (approx. 17 h of speech) were labeled by one of the authors in about four months. We distinguish 10 different boundary types which can be mapped onto the three cover classes (as for more detail cf. [3, 4]):

- M3: strong syntactic boundary
- M0: weak or no boundary
- MU: 'undefined' boundaries

Note that besides the M3 boundaries, the MU boundaries are of special interest for the syntactic analysis. The syntactic structure of an utterance is usually ambiguous at MU positions, i.e. it cannot be predicted from the text alone; often there are two or more alternative word boundaries, where the syntactic boundary could be placed, and disambiguation can only be performed by prosodic means (i.e. in practice, by a prosodic classifier).

For a better comparison, all recognition results described in the following were obtained on the same test set comprising 3 VERBMOBIL dialogs* (64 turns of 3 male and 3 female speakers, 12 minutes in total). In case of the prosodic B-boundaries and the accents, for the training of the MLPs 30 disjoint dialogs (797 turns of 53 male and 7 female speakers, 100 minutes in total) were used; for the syntactic-prosodic M-boundaries more training data was available (6209 turns of 322 male and 203 female speakers, approx. 13 h of speech). The recognition rates for the accent classification are evaluated on all wordfinal syllables of the test set; the rates for the boundary classification are only determined on the wordfinal syllables without taking into account the utterance final syllables as their detection is more or less trivial.

3 Extraction of prosodic features

We distinguish different categories of prosodic feature levels; an overview is shown in figure 1 (as for more detail cf. [8]). The *acoustic prosodic features* are signal–based features that usually span over speech units that are larger than phonemes (syllables, words, turns, etc.). Normally, they are extracted from the specific speech signal interval that belongs to the prosodic unit, describing its specific prosodic properties, and can be fed directly into a classifier, e.g., into an MLP. Within this group we can further distinguish:

- *basic prosodic features*
  which are extracted from the pure speech signal without any explicit segmentation into prosodic units. Examples are the frame-based extraction of fundamental frequency (F_0) and energy. Usually the basic prosodic features cannot be directly used for a prosodic classification.

- *structured prosodic features*
  are computed over a larger speech unit, e.g., syllable, syllable nucleus, word, turn, partly from the prosodic basic features, e.g., features describing the shape of F_0 or energy contour, partly based on segmental information that can be provided e.g. from the output

*Note, that in comparison to the experiments we described in [10] here a more representative test set was used and a slightly larger training set was available.

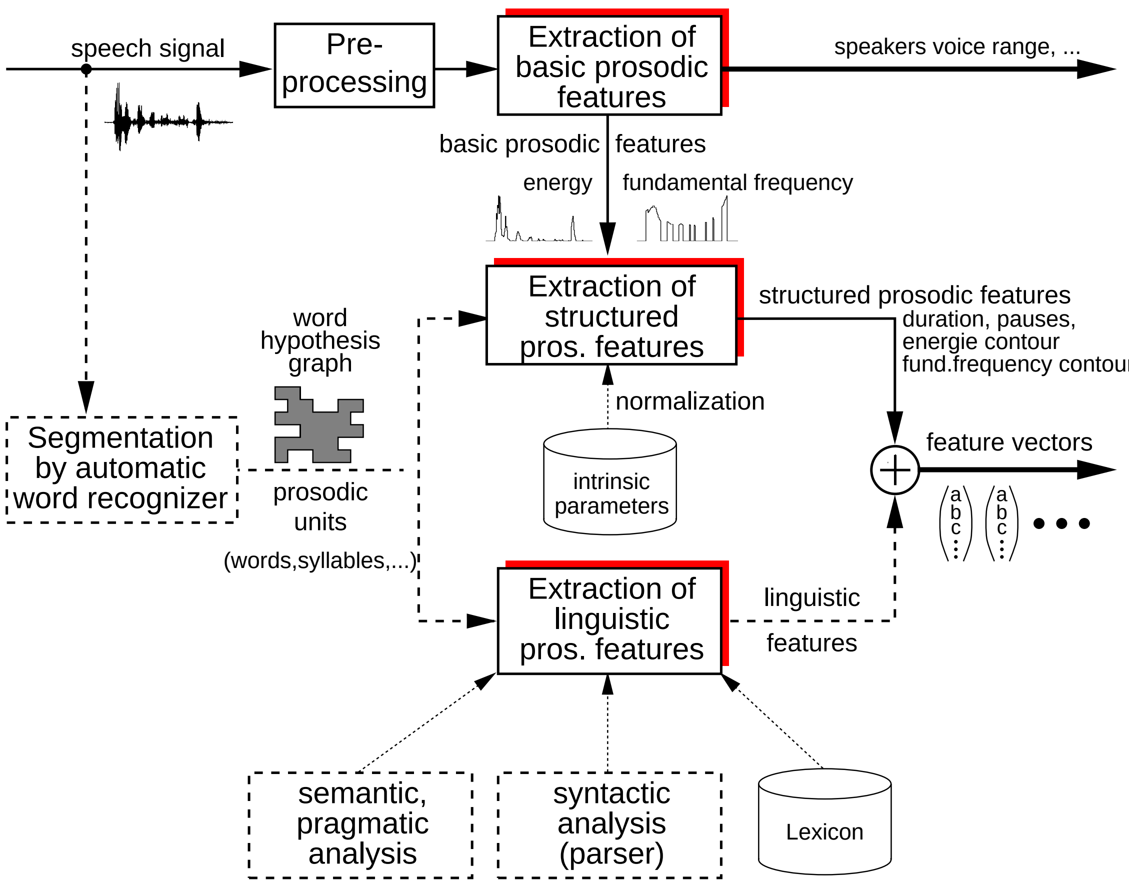

Figure 1: Sketch of the process of prosodic feature extraction.

of a word recognizer, e.g., features describing durational properties of phonemes, syllable nuclei, syllables, pauses.

On the other hand prosodic information is highly interrelated with 'higher' linguistic information, i.e. the underlying linguistic information strongly influences the actual realization and relevance of the measured acoustic prosodic features. In this sense, we speak of *linguistic prosodic features* that can be introduced from other knowledge sources, as lexicon, syntax, or semantics; usually they have either an intensifying or an inhibitory effect on the acoustic prosodic features. The linguistic prosodic features can be further divided into:

- *lexical prosodic features*
  are categorical features that can be extracted from a lexicon that contains syllable boundaries in the phonetic transcription of the words. Examples for these features are flags marking if a syllable is wordfinal or not or denoting which syllable carries the lexical word accent. Other possibilities not considered here might be special flags marking content and function words.

- *syntactic/semantic prosodic features*
  encode the syntactic and/or semantic structure of an utterance. They can be obtained from syntax, e.g., from the syntactic tree as in [7, 6], or they can be based on predictions of possibly important – and thus accented – words from the semantic or the dialog module.

In this paper we do not consider syntactic/semantic prosodic features; in the following, the cover term prosodic features means mostly structured prosodic features and some lexical prosodic features. However, in [10, 1, 12], e.g., we use polygram language models (i.e. N-grams of variable length) and/or Semantic Classification Trees for modeling the higher level information and combine the output of these classifiers with the MLP output which uses more or less only acoustic prosodic features. The rest of this paper only deals with this latter topic.

In this context, the computation of the prosodic features is based on an automatic time alignment of the phoneme sequence corresponding to the spoken or recognized words. In this paper, we only use the aligned spoken words thus simulating 100% word recognition. The time alignment is done by a standard hidden Markov model word recognizer.

Due to the problems discussed in the first section, it is still an open question, which prosodic features are the most relevant for the different classification problems and how the different features are interrelated. We try therefore to be as exhaustive as possible, and leave it to the statistic classifier to find out the relevant features and the optimal weighting of them. As many relevant prosodic features as possible are therefore extracted over a prosodic unit (here: the word final syllable) and composed into a huge feature vector which represents the prosodic properties of this and of several surrounding units in a specific context.

We investigated different contexts of up to ± 6 syllables (± 3 words, resp.) to the left and to the right of the actual wordfinal syllable. For every classification problem investigated many different subsets of these features were analysed. The best results so far for the $\neg$B|B and the $\neg$A|A problem were achieved by using 276 features computed for each word considering a context of ± 2 syllables (± 2 words, resp.). Note, that in contrast to the accent classification experiments in [10], here we chose a different strategy: In [10] we computed one feature vector for each syllable performing a syllable-based $\neg$A|A classification. As for the semantic analysis it is more important to know which *words* are accented and of less interest which syllables are accented, the syllable-based classification results had to be mapped onto a judgement for the word. Here, we compute one feature vector per word, performing a word-based $\neg$A|A classification. Experiments on the same data bases as in [10] showed a reduction of the error rate of about 20 % compared to the syllable-based approach. In more detail the features used here are:

- duration (absolute and normalized as in [18]) for each syllable/syllable nucleus/word
- for each syllable and word in this context
  - minimum and maximum of fundamental frequency (F_0) and their positions on the time axis relative to the position of the actual syllable as well as the F_0-mean
  - maximum energy (also normalized) + positions and mean energy (also normalized)
- F_0-offset + position for actual and preceding word
- F_0-onset + position for actual and succeeding word
- for each syllable: flags indicating whether the syllable carries the lexical word accent or whether it is in a word final position
- length of the pause preceding/succeeding actual word
- linear regression coefficients of F_0–contour and energy contour over 11 different windows to the left and to the right of the actual syllable
- for an implicit normalization of the other features, measures for the speaking rate are computed over the whole utterance based on the absolute and the normalized syllable durations (as in [18])

reference	#	classified as B	classified as ¬B
B	165	84.8	15.2
¬B	1284	11.2	88.8

Table 1: Confusion matrix for the classification of prosodic boundaries (¬B|B).

feature set (SET)	number of features	SET alone ¬A | A $\mathcal{RR}$	SET alone ¬A | A $\mathcal{RR}_{\overline{C}}$	SET alone ¬B | B $\mathcal{RR}$	SET alone ¬B | B $\mathcal{RR}_{\overline{C}}$	ALL \ SET ¬A | A $\mathcal{RR}$	ALL \ SET ¬A | A $\mathcal{RR}_{\overline{C}}$	ALL \ SET ¬B | B $\mathcal{RR}$	ALL \ SET ¬B | B $\mathcal{RR}_{\overline{C}}$
ALL	276	**82.6**	**(82.2)**	**88.3**	**(86.8)**	—		—	
DURATION	60	74.9	(74.7)	78.7	(77.7)	81.7	(81.4)	83.9	(85.1)
F_0	80	79.4	(79.1)	81.3	(82.6)	81.7	(81.5)	84.2	(85.5)
ENERGY	112	77.3	(77.0)	81.8	(81.8)	82.2	(81.8)	85.6	(85.3)
PAUSE	6	57.4	(55.4)	88.4	(72.1)	82.3	(82.0)	87.4	(85.3)
SPEAKING RATE	3	50.4	(51.3)	48.6	(54.9)	82.0	(81.5)	87.7	(86.2)
FLAGS	15	79.2	(79.4)	69.6	(74.9)	81.6	(81.2)	86.6	(85.6)
(F_0 without POS)	56	76.2	(75.3)	78.6	(75.8)	82.4	(82.0)	84.5	(85.5)

Table 2: Recognition rates for the classification of accents (¬A | A) and prosodic boundaries (¬B | B) for different feature sets. It is distinguished between the classification with different feature sets (column SET alone) and the classification with all features but the ones corresponding to the actual row (column ALL \ SET). Besides the overall recognition rate ($\mathcal{RR}$) in parenthesis also the averages of the class-wise recognition rates ($\mathcal{RR}_{\overline{C}}$) are given.

4 Experiments and results

In this paper, we will only report results obtained with multi-layer perceptrons (MLP) that turned out to be superior compared to Gaussian distribution classifiers in similar investigations [9]. Different MLP topologies were analysed for the various classification problems. As training procedure the Quickpropagation algorithm [5] with sigmoid activation function was used. Experiments were performed with different feature sets. In any case the MLPs had as many input nodes as the dimension of the specific feature vector and one output node for each of the classes to be recognized. During training the desired output for each of the feature vectors is set to one for the node corresponding to the reference label; the other one is set to zero. With this method in theory the MLP estimates a posteriori probabilities for the classes under consideration. 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.

The best result for the classification of prosodic boundaries (¬B|B) is illustrated in table 1 in a confusion matrix. It was obtained using an MLPs with 40/20 nodes in the first/second hidden layer. The overall recognition rate ($\mathcal{RR}$) here is 88.3 %, the average of the class-wise recognition rates ($\mathcal{RR}_{\overline{C}}$) is 86.8 %. In table 2 the results for experiments with different feature subsets of this best feature set is shown for the recognition of prosodic boundaries (column ¬B|B) as well as for the classification of accents (column ¬A|A).

It can be noticed that although the sole use of some feature subsets shows already respectable results whereas some (row SPEAKING RATE) seem to be almost neglectable, the best

reference	#	classified as M3	M0
M3	177	87.6	12.4
M0	1169	14.2	85.8
MU	103	61.2	38.8

Table 3: Confusion matrix for the classification of syntactic-prosodic boundaries.

recognition rate can only be achieved, if *all* features are used in combination (row ALL). For the ¬B|B problem the most important features are F_0, ENERGY and PAUSE. Concerning the ¬A|A classification, F_0 is also the most relevant important group and in contrast to the ¬B|B problem more relevant than ENERGY. An explanation for the superiority of F_0 and ENERGY compared to DURATION might be the fact that durational information is already modeled in the position features of F_0 and ENERGY. This shows also the distinct drop of the recognition rate if only the 'pure' F_0 features without their positions (row 'F_0 without POS') are used. The lexical prosodic features (row FLAGS) seem to be much more relevant for the ¬A|A classification than for the ¬B|B classification.

In Table 3, the confusion matrix for the classification of syntactic-prosodic boundaries is given. This best result was obtained using a feature set different from the one described above, comprising 121 prosodic features computed in a context of ± 2 syllables (no word context was considered here). An MLP with 100/50 nodes in the first/second hidden layer was trained only with the M3 and the M0 labels, not with the 'undefined' MU boundaries. The average of the class–wise recognition rate ($\mathcal{RR}_{\overline{C}}$) for M3 vs. M0 is 86.7 %, i.e. almost the same as the recognition rate for the prosodic labels. This is at the same time interesting and encouraging: We only use – besides some flags – acoustic–prosodic features for the training. The M labels, however, were obtained only with the written, not with the spoken word chain. It could therefore not automatically be expected that they are marked prosodically to the same extent as the prosodic–perceptual B labels. Note that the M labels meet the demands of the syntactic analysis to a greater extent than the B labels. A further advantage is that they are more easily obtained. In [4], where we additionally use polygram language models which need large training data bases, we show that probably because of that, the discrimination between M0 and M3 based on a large training data base yields better results than the results given in table 3. Note however, that the MUs cannot be predicted from the text (i.e. by the means of language models) alone; for their correct classification always prosodic information has to be considered.

5 Discussion and future work

In this paper we showed that prosodic as well as syntactic-prosodic boundaries and accents can be classified in spontaneous speech to a fairly high extend by only considering acoustic prosodic and some lexical prosodic features. In [10, 1, 12] we could achieve further improvements by combining the MLP output with polygrams which model the probabilities of word and boundary label subsequences. In the near future, we will further optimize the feature set and the classifiers. The boundary information achieved with our classifiers is already used in the VERBMOBIL project by the higher modules syntax [2], semantics, transfer, and dialog. The feedback based on results obtained with these modules and a parallel detailed error analysis will hopefully result in a further improvement of our labeling system and, in turn, an even more

reference	#	classified as IR3	IZEB	WB
IR3	566	61.0	20.1	18.9
IZEB	1035	14.9	78.2	7.0
WB	34951	12.3	10.5	77.2

Table 4: Confusion matrix for the classification of agrammatical boundaries.

adequate use of prosodic information in the VERBMOBIL system.

The approach described here can easily be extended to different but related problems where prosodic information seems to be helpful. Thus, in a very similar way as for the classification of boundaries, first experiments have been already conducted for the detection of agrammatical boundaries. As agrammatical boundaries (IR3) we understand word boundaries where a word fragment, a new construction or the end of a reparandum occurs (cf., e.g., [11, 16]). In ASU systems the detection of these phenomena can play a very important role, e.g., for the improvement of the syntactic analysis. Although in VERBMOBIL different types of agrammatical boundaries are annotated in the transliteration of the dialogs, in the following we do not distinguish between them. Because of this annotation, agrammatical boundaries can be easily extracted and – in analogy to the prosodic boundaries – assigned to the corresponding position in the speech signal. In these first experiments we tried to distinguish IR3 from irregular word hesitations (IZEB) and from all other regular word boundaries (WB). Hesitations are of minor importance but usually they are realized with similar prosodic means as the IR3; they can therefore easily be confused with them. The IZEB are also annotated in the transliteration.

As these phenomena do usually not occur as often as, e.g., 'normal' syntactic boundaries, even in spontaneous speech, for training and testing our classifiers we chose much larger subset than for the experiments described above (training: 6332 turns, approx. 14 h of speech; test: 1823 turns, approx. 4 h of speech). A first preliminary result with respect to the classification of IR3, IZEB and WB is presented in table 4. Here, the same 121 features and the same classifier as for the classification of the syntactic-prosodic boundaries was used. The overall recognition rate ($\mathcal{RR}$) here is 77.0 %, the average of the class-wise recognition rates ($\mathcal{RR}_{\overline{C}}$) is 72.1 %. Although these rates are not that high they are nevertheless encouraging: Agrammatical boundaries, pure hesitations and regular word boundaries can be distinguished with a probability significantly above chance level by only using prosodic features. It should be noted that neither the features nor the classifiers have been optimized in any respect; further improvements are very likely. A very promising approach that we want to investigate in the near future is a combined classification of agrammatical boundaries and syntactic-prosodic boundaries. We expect further a distinct improvement of the results by considering polygram language models for classification, in an analogous way as for the classification of prosodic boundaries.

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*, Kopenhagen, 1996. (to appear).

[2] A. Batliner, A. Feldhaus, S. Geißler, T. Kiss, R. Kompe, and E. Nöth. Prosody, Empty Categories and Parsing — A Success Story. In *Int. Conf. on Spoken Language Processing*, Philadelphia, 1996. (to appear).

[3] A. Batliner, R. Kompe, A. Kießling, M. Mast, and E. Nöth. All about Ms and Is, not to forget As, and a comparison with Bs and Ss and Ds. Towards a syntactic–prosodic labeling system for large spontaneous speech data bases. Verbmobil Memo 102, 1996.

[4] 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*, Philadelphia, 1996. (to appear).

[5] S.E. Fahlman. An Empirical Study of Learning Speed in Back–Propagation Networks. Technical Report CMU-CS-88–62, Carnegie Mellon University, Pittsburgh, 1988.

[6] A. Hunt. *Models of Prosody and Syntax and their Application to Automatic Speech Recognition.* PhD thesis, University of Sydney, 1995.

[7] A. Hunt. Syntactic Influence on Prosodic Phrasing in the Framework of the Link Grammar. In *Proc. European Conf. on Speech Communication and Technology*, volume 2, pages 997–1000, Madrid, Spain, 1995.

[8] A. Kießling. Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung. Dissertation. Technische Fakultät der Universität Erlangen–Nürnberg, 1996. (to appear).

[9] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. Detection of Phrase Boundaries and Accents. In H. Niemann, R. De Mori, and G. Hanrieder, editors, *Progress and Prospects of Speech Research and Technology: Proc. of the CRIM / FORWISS Workshop*, PAI 1, pages 266–269. Infix, 1994.

[10] 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.

[11] R.J. Lickley, R.C. Shillcock, and E.G. Bard. Processing Disfluent Speech: How and When are Disfluencies Found? In *Proc. European Conf. on Speech Communication and Technology*, volume 3, pages 1499–1502, Genova, 1991.

[12] M. Mast, R. Kompe, St. Harbeck, A. Kießling, H. Niemann, and E. Nöth. Dialog Act Classification with the Help of Prosody. In *Int. Conf. on Spoken Language Processing*, Philadelphia, 1996. (to appear).

[13] M. Reyelt. Consistency of Prosodic Transcriptions Labelling Experiments with Trained and Untrained Transcribers. In *Proc. of the 13th Int. Congress of Phonetic Sciences*, volume 4, pages 212–215, Stockholm, 1995.

[14] M. Reyelt. Ein System zur prosodischen Etikettierung von Spontansprache. In R. Hoffmann and R. Ose, editors, *Elektronische Sprachsignalverarbeitung*, volume 12 of *Studientexte zur Sprachkommunikation*, pages 167–174. TU Dresden, Wolfenbüttel, 1995.

[15] M. Reyelt and A. Batliner. Ein Inventar prosodischer Etiketten für Verbmobil. Verbmobil Memo 33, 1994.

[16] E.E. Shriberg. *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis, University of California at Berkeley, 1994.

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

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