gefördert vom
Bundesministerium
für Forschung
und Technologie

Verbmobil
Verbundvorhaben

AUTOMATIC LABELING OF PHRASE ACCENTS IN GERMAN

Andreas Kiessling, Ralf Kompe
Anton Batliner, Heinrich Niemann
Elmar Nöth

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

Vm
Report 35
Oktober 1994

Oktober 1994

Andreas Kiessling, Ralf Kompe
Anton Batliner, Heinrich Niemann
Elmar 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 Forschung und Technologie unter dem Förderkennzeichen 01 IV 102 H/0 und 01 IV 102 C 6 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

AUTOMATIC LABELING OF PHRASE ACCENTS IN GERMAN

Andreas Kießling[1], **Ralf Kompe**[1], **Anton Batliner**[2], **Heinrich Niemann**[1], **Elmar Nöth**[1]
[1]Friedrich-Alexander-Universität Erlangen-Nürnberg,
Lehrstuhl für Mustererkennung (Informatik 5),
Martensstr. 3, D-91058 Erlangen, F.R. of Germany

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

ABSTRACT

In this paper a method for the automatic labeling of phrase accents is described, based on a large text corpus that has been generated automatically and read by 100 speakers. Perception experiments on a subset of 500 utterances show a high agreement between the automatically generated accent labels and the judgment scores obtained. We computed different prosodic feature vectors from the speech signal for each syllable and trained different Gaussian distribution classifiers and artificial neural networks using the automatically generated accent labels. Recognition rates of up to 83% could be achieved for the distinction of accentuated vs. unaccentuated syllables. Similar results could be obtained for the comparison of the listeners judgments with the automatic classification.

I. INTRODUCTION

Information about accents can be used in different fields of automatic speech understanding (for an overview cf. [7]), e.g. for the improvement of semantic interpretation (detection of focal accent), isolated word recognition with large-vocabularies [13], and continuous word recognition (rescoring of the n-best sentence hypotheses computed by a word recognizer). For a proper treatment of phrasing and sentence mood, accent structure has to be taken into account because all these phenomena highly interact with each other.

For the training of statistical classifiers, large databases with reference labels have to be available. In general, for word recognition a transliteration is sufficient for a successful training. The labeling of prosodically marked phrase boundaries, accents, or sentence mood, however, is much more difficult and time consuming, because usually it has to be done manually after the recording by well trained experts. We therefore developed a method for an automatic generation of phrase accents for a large text corpus for which speech data were already recorded. With this material we wanted to train classifiers for the detection of stressed syllables and to reuse these classifiers for other speech material following a bootstrap strategy. In the context of this paper the words 'accent' and 'accentuated' are used to denote the syntactically motivated, automatically generated accent labels for words and syllables.

II. MATERIAL

The material we investigated is the German speech database ERBA, "**Er**langer **B**ahn **A**nfragen" (Erlangen train inquiries) a large speech training database for word recognition in the domain of train table inquiries. A stochastic sentence generator was used based on a context free grammar and 38 sentence templates to create a large text corpus. At four different sites a subset of 10,000 unique sentences was recorded in quiet office environments (100 untrained speakers, 100 utterances each) resulting in a speech database of about 14 hours. The speakers were given the word sequences with punctuation marks; for more details concerning ERBA see [2].

The set of 100 speakers was partitioned into the following three subsets: 69 speakers (44 male, 25 female, 6,900 sentences) for training, 21 speakers (12 male, 9 female, 2,100 sentences) for testing, and the remaining 10 speakers for perception tests and also for testing.

The perception experiments were conducted in order to get reference labels for prosodically marked phrase boundaries and accentuated syllables. This information is used to improve the automatic generation of phrase accents in an iterative process of generation and control. Ten "naive" listeners were given 500 utterances[1] from 10 speakers (5 male, 5 female, 50 utterances each) in orthographic form without any punctuation marks. In a first experiment their task was to mark the space between two words if they felt it separated two different "chunks" of speech. In a second experiment another group of ten "naive" listeners was asked to mark each syllable they perceived as stressed. Thus, each possible accent position (= syllable) and each possible phrase boundary position (= word boundary) got a perception score from 0 (no mark) up to 10 (all 10 subjects in the test perceived an accent or a phrase boundary as marked). The listeners were instructed not to rely upon their knowledge of canonical forms or sentence structure, although influence of these factors can certainly not be ruled out altogether.

III. AUTOMATIC GENERATION OF PHRASE ACCENTS

3.1. Phrase boundary labels as prerequisite

The automatic generation of phrase accents is based on the automatically generated phrase boundary markers described in [2] and [6]: Syntactic boundaries were marked in the grammar and included in the sentence generation process with some context-sensitive post-processing. The result is the orthographic word chain separated by boundary labels. We distinguish four types of phrase boundaries: Boundary B3 is placed between elliptic clause and clause or between main and subordinate clause, B2 is positioned between constituents or at coordinating particles between constituents, B1 belongs syntactically to the normal constituent boundary B2 but is most certainly not marked

[1]For the perception tests only sufficiently long and semantically meaningful sentences were used: When generating sentences with a context free grammar "nonsense" sentences like *"between ten and ten o'clock"* can not be avoided. The intonation of such sentences might be irregular, even hesitations may occur, which can be the reason for "miss"-classification. Since ERBA initially was intended to train word recognizers such "nonsense" sentences were not discarded.

prosodically because it is close to a B3 boundary or to the beginning/end of the utterance, and B0 is any other word boundary that does not belong to B1, B2, B3. In [1] it is described how the automatic classification of these phrase boundaries can be used by a parser to reduce the possible syntactic derivations and thereby speed up the parsing time.

For the assignment of accents, it has to be decided which words in an utterance are accentuated. In words with more than one syllable, normally one of these syllables bears the word accent; this syllable can be looked up in the lexicon. Factors that might influence whether or not a word is accentuated include the form class of a word (content word: CW vs. function word: FW), its position in a larger prosodic context, and tempo (isolating vs. integrating accentuation). Rhythmic constraints can influence the location of accent within a word; for details, cf. [8]. In order to take into account most of these factors the automatic generation of accent labels was iteratively controlled with and adapted to the results of the perception experiments.

3.2. Assigning the lexical word accent

Before creating the accent labels, we first compared the word accents marked in the lexicon with the results of the perception experiments in order to derive rules for the position of the phrase accent. For the labeling of the accents in the lexicon we decided in favor of a rather broad labeling, i.e. we only distinguish accentuated from unaccentuated syllables. Secondary accentuation is not labeled because in a canonical citation pronunciation, these differences might be produced and perceived systematically but not in a more casual pronunciation as is the case in fluent speech. In the lexicon, the 75 FWs (articles, pronomina, auxiliary verbs, prepositions, conjunctions) were not marked as accentuated. They are normally clitic i.e. without accent and integrate with the following constituent into a greater prosodic phrase. In general, CWs are represented with just one accentuated syllable (word accent). If more than one accentuation is possible without change of the meaning as e.g. in some proper nouns and longer words (*"Erlangen"* and *"zweiundzwanzig"* respectively with accent on the first **or** on the penultimate syllable) both positions are marked in the lexicon.

3.3. Assigning the accent label to a word within a phrase

The next step was to decide which words within a phrase are accentuated. Since for the moment we do not consider emphatic or contrastive accents we assume that in each prosodic phrase (bounded by B1, B2, or B3[2]) one and only one word is more prominent than the others. In German, the phrase accent is normally positioned on the rightmost CW in a NP (*'rightmost principle'*); in a PP and in a VP, by default the argument is the carrier of the phrase accent, i.e. not the preposition or the verb (cf. [11], [3]).

The examination of the perception scores showed in some cases additional tendencies not to put stress on 'semantically weak' CWs or to put stress on 'strong' FWs. In the following example the syllables to be expected as stressed are typed bold: *ich möchte* B1 *am nächsten* ***Diens****tag* B2 *zwischen* **drei** B2 *und* **sechs** *Uhr* B2 *von* ***Ham****burg* B2 *nach* **Ulm** B1 *fahren* (*I would like* B1 *next Tuesday* B2 *between three* B2 *and six o'clock* B2 *from Hamburg* B2 *to Ulm* B1 *to go*). This example contains the two most important exceptions: The word *"Uhr"* and other CWs like e.g. verbs such as *"fahren"* that are rather predictable in the domain of train table inquiries and therefore semantically weak or clitic, are usually not accentuated and thus got a rather low perception score. Therefore, in the last two phrases not the verb *fahren* but the city name *Ulm* is expected to be stressed. On the other hand, often FWs with a rather high perception score could be observed, e.g. interrogative pronouns such as *"was"*, *"wann"*, *"welche"* that obviously are semantically and pragmatically strong words in this domain. The semantic weakness of the verb coincides with the above mentioned rule that verbs by default are not accentuated. There are, however, exceptions, as, e.g. the so called particle verbs like *"ankommen" (arrive)* and *"abfahren" (leave)* that might be accentuated.

Based on these observations the following rules for our algorithm were formulated: For each phrase bounded on the right by symbol Bx ($x \in \{1,2,3\}$) look successively for the rightmost CW*[3], or (if not found) for the rightmost verb, or for the word 'Uhr', or for an interrogative pronoun, or for an auxiliary verb, or for any other word and mark the first instance by symbol `Ax` (where `x` corresponds to x in Bx). After applying this rule to a sentence, in each phrase one and only one word is marked by an accent label `Ax` ($x \in \{1,2,3\}$).
In order to take into account that there are semantically weak words occurring in short phrases before a B3 boundary, we have to add another rule: If the actual word is not a CW* and the phrase is bounded on the left by B1 and on the right by B3 and there is a CW* on the left hand side of the B1 boundary then exchange the accent labels of these two words. This rule e.g. changes *"...nach* `A1`*Ulm* B1 `A3`*fahren* B3*"* into *"...nach* `A3`*Ulm* B1 `A1`*fahren* B3*"*.
Special treatment is necessary for certain compound words that occur very frequently in our application (e.g. city names, see also paragraph 3.4). In our lexicon these words are characterized by a linking hyphen or dash. Following the rightmost principle, we marked the rightmost word by `Ax`, and all other words of the compound word by `Axi`, denoting that there is an 'implication' from left to right, i.e. if any word of the compound word is stressed, all its right hand neighbors are stressed as well. It has to be noted, that this rule is rather straightforward and does not take into account other possibly relevant factors as, e.g. rhythmic constraints.

3.4. Assigning the accent label to a syllable within a word

After the accent labels are assigned to the words, we have to determine the syllables within the words bearing the accent. In our material, this assignment depends on several factors:

- If the word has only one syllable marked as the (lexical) word accent in the lexicon, this syllable inherits the symbol `Ax` from the word.
- If more than one syllable can be accentuated, all these syllables get the symbol `Axa`, denoting that they are real alternatives, and that it is at discretion of the speaker which of those alternatives actually is stressed.
- If there is no lexical accent at all for this word (which is usually the case for FWs) the first syllable in the word gets the symbol `Axn`, denoting that it is just a default (root) accent[4].

These rules apply in the same way to single words and to the parts of the compound word marked by 'implication' labels. For example, the syllables of the greeting *Grüß_Gott* are labeled with [`Axi Ax`] and the city name *Riebnitz-Damgarten-West* is labeled with [`Axi A0 Axi A0 A0 Ax`].

To take into account that syllables positioned directly before a phrase boundary are usually produced differently from others due to phrase final lengthening, we introduced additional markers. Another reason for labeling these syllables in a special way is that at present we are also investigating the combined recognition of phrase boundaries and phrase accents based on syllables (cf. [5]) as well as the modeling of phrase structures by Hidden Markov Models. Therefore, if one of the already marked accentuated syllables is positioned directly before a phrase boundary marked by Bz ($z \in \{1,2,3\}$) it gets the additional label `+Bz`. All the remaining (unaccentuated) syllables in the sentence are labeled with `Bz` if they are positioned directly before a phrase boundary marked by Bz, otherwise they are marked as `A0`.

[2]Note, that for the generation of the accent labels also the beginning and the end of an utterance is assumed to be a B3 boundary.

[3]CW* denotes in our context any word that is not a FW, verb, auxiliary verb, interrogative pronoun or the word 'Uhr'.

[4]This simple rule can of course not be applied to all German FWs but it works reasonably well within our lexicon.

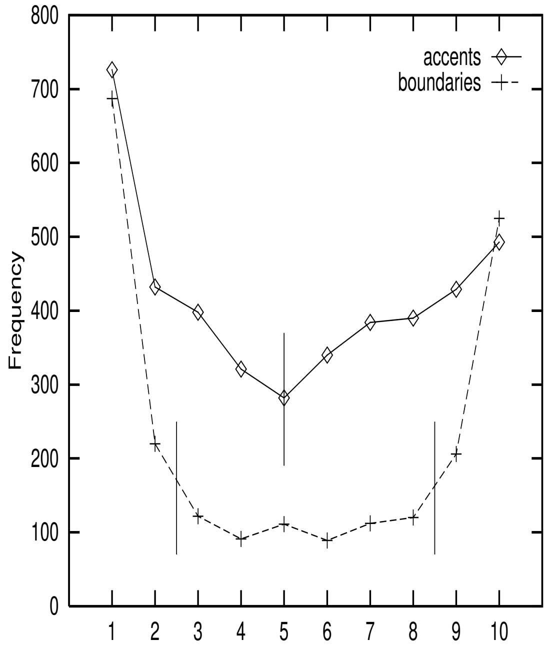

Figure 1. Frequency of accent and boundary scores

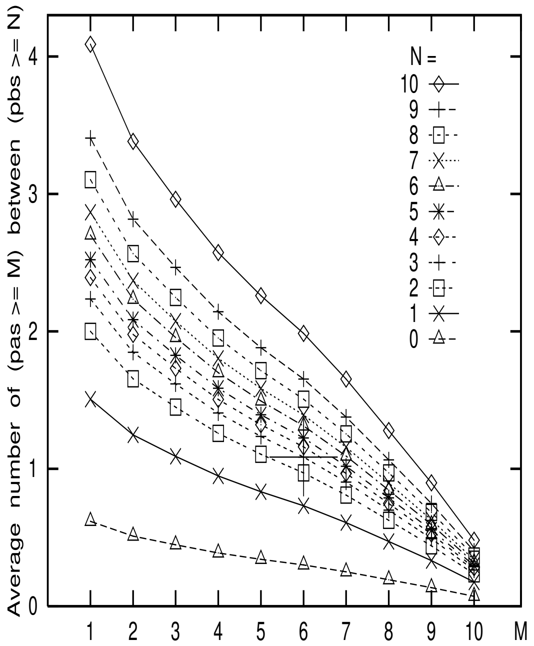

Figure 2. The relation between accent and boundary scores

By applying all these rules to the whole ERBA database of 10,000 sentences in total 199,078 syllables were marked by 30 different symbols.

3.5. Comparison of the accent labels with the listeners judgments

The perception data were compared with the automatically labeled places of phrase boundaries and phrase accents. The 500 utterances contain 71 types of FWs with 3346 tokens and 588 types of CWs with 3396 tokens. FWs got an average score of 1.4 with a minimum of 0 and a maximum of 8; 10% were above 5 and 36% above the mean. CWs got an average score of 7.4 with a minimum of 0 and a maximum of 10; 14% were less than 5 and 47% were less than the mean.

In Figure 1, the frequencies of the perceptual scores for boundaries and accents are plotted. The curve for the accents is V-shaped with a turning point at 5, that is in the middle of the scale. It thus makes sense to define syllables with a score higher than 5 as accentuated. For the phrase boundaries the curve is U-shaped with no clear turning point. We assume that our boundary labels fall not into two but into three distinct classes: B01, B2, B3 (cf. also [2]). It thus makes sense to define two turning points: B01 below 3, B3 above 8, and B2 in between. (The assumed thresholds are marked in Figure 1 by vertical lines.) This last assumption is supported by the relationship between accent and boundary scores illustrated in Figure 2: The abscissa represents a threshold M, partitioning the perceived accent scores (pas) into two classes: if pas$\geq$M, the syllable is defined to be accentuated, otherwise it is not accentuated. Each of the curves (marked with N $\in$ [0;10]) represents a threshold, partitioning the perceived boundary scores (pbs) into two classes: if pbs$\geq$N, the word boundary is defined to be a phrase boundary. The cross plotted indicates M=6, N=3 and an ordinate value of about 1; i.e. the mean value of the number of accent scores higher than 5 within a phrase bounded by a boundary with a perceptual score higher than 2 is about 1.

Usually it is assumed that in each phrase there is one prominent syllable, represented, e.g., in the tone sequence approach (cf. [10]) by one starred tone. As illustrated here, by setting the M threshold for the accent scores to 6, the N threshold for the boundary scores to 3, this assumption is supported pretty well by our empirically obtained perception data: The mean value of the number of accented scores is roughly 1, i.e. for each phrase defined in that way there is on the average one prominent syllable that can be defined as the carrier of the phrase accent. As can be easily seen these phrases correspond to the constituents that are marked by B2 boundaries (cf. [2]). (This is of course no "prove" but rather a sort of cumulative evidence.)

For the comparison of the generated accent labels with the listeners judgments the critical cases (i.e. the 'alternative' and the 'implicated' accents) are not taken into consideration and the original 30 accent symbols are mapped onto five accent types: `A1`, `A2`, and `A3` denote phrase accents corresponding to the phrases of type B1, B2, and B3; `B` denotes unaccentuated syllables immediately preceding a phrase boundary, `A0` any other (unaccentuated) syllable. In Figure 3, these five accent types are cross-classified with the listeners judgments. The scores for `A0` and `B`, i.e. the unaccentuated syllables meet our expectation; 90% of the `A0` and more than 91% of the `B` syllables were perceived as stressed by less than 2 listeners. The accent types `A2` and `A3` clearly cluster at the right end although the tendency is not as distinct as for the corresponding phrase boundaries B2 and B3 (cf. [2]). The accent type `A1` (word accent syllable in a prosodically 'weak' constituent) is obviously marked more often than `A0` (unaccentuated syllable). Note that the `A3` scores are not markedly higher than the `A2` scores. It is often assumed that the sentence accent in German is by default the rightmost phrase accent in an utterance (`A3` accent in our material) and more prominent than any other phrase accents (`A2` accents in our material). Our result might be taken as an argument against a phonetic manifestation of sentence accent in German.

IV. AUTOMATIC CLASSIFICATION OF PHRASE ACCENTS

For each syllable of an utterance different features were computed from the speech signal, describing prosodic properties like timing, intonation, and intensity. The time alignment of words and syllables was computed with our hidden Markov model word recognizer [9]. In [5], the features and a first evaluation of different subsets of the features are discussed in more detail. Here, only an short description of the used features is given:

- length of the pause following the syllable obtained from the time alignment of the word chain

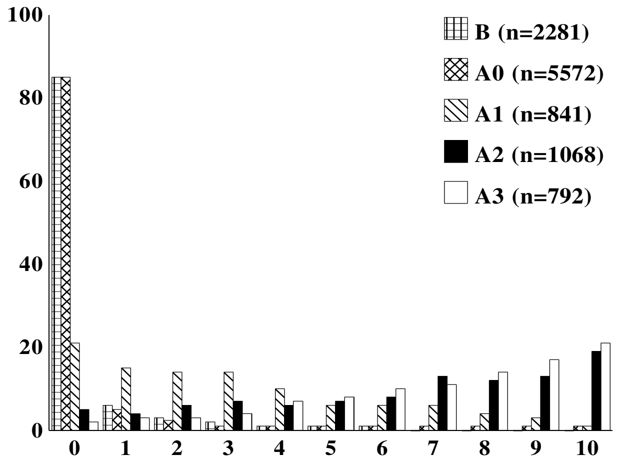

Figure 3. Frequency (in %) of scores for accent types

- duration of the syllable and of the syllable nucleus, normalized in different ways (comparable to [14])
- mean and maximum of the intensity normalized over different contexts
- linear regression coefficients, as well as the minimum and the maximum of the F0-contour[5] computed over different regions around the syllable

For these first, preliminary classification experiments the special distinction into 'alternative', 'implicated', and 'no lexical accent' and therefore the additional markings (`a`, `i`, and `n`) were ignored. (This distinction is being investigated in ongoing work.) Additionally, the remaining labels were mapped onto six super-classes representing `accentuated` syllables before a strong phrase boundary (`A23+B2`[6], `A23+B3`) or before no boundary (`A23+B01`) and `unaccentuated` syllables before a strong phrase boundary (`A01+B2`, `A01+B3`) or before no boundary (`A01+B01`).

We trained Gaussian distribution classifiers (GDC) and artificial neural networks (ANN) on the 6,900 sentences (137,183 training patterns) to distinguish between the six super-classes. The test was performed on the 2,100 sentences as well as on the 500 sentences of the perception test, where the M threshold for the accent scores was set to 6. For both GDC and ANN, many different feature sets were investigated. After classification, the recognition results (i.e. the six super-classes) are mapped onto the two classes `accentuated` vs. `unaccentuated`. The best recognition results for GDC and ANN distinguishing the six classes, the two classes `accentuated` and `unaccentuated`, and the comparison of the two classes with the listeners judgments are shown in Table 1. Note, that these best recognition results are obtained with different feature sets. The best ANN has 40 input nodes, 2 hidden layers, 40 nodes in the first, 20 nodes in the second hidden layer and 6 output nodes. For training, the Quickpropagation algorithm with sigmoid activation function was used.

V. CONCLUDING REMARKS

Our accent assignment rules are rather straightforward but seem to work reasonably well with this corpus. However, further effort is necessary towards their improvement, because other relevant factors as, e.g. rhythmic or other syntactic constraints are not taken into account yet. Furthermore, other, especially spontaneous, speech data bases might require different rules.

At present we are investigating the integrated recognition of accents and phrase boundaries using large syllable based feature vectors (cf. [5]), as well as the adaptation of the classification to the domain of appointment scheduling in the VERBMOBIL project [12]. In ongoing work we compare different feature sets and different classifiers as well as the employment of language models for modeling the succession of different syllable types. Moreover, we want to investigate a syllable based modeling and recognition of different types of prosodic phrases by Hidden Markov Models.

	6 classes	2 classes	listeners
GDC	61.2%	77.4%	74.6%
ANN	70.4%	83.1%	79.1%

Table 1. Average recognition rates for GDC and ANN

[5] The F0-contour was computed using an iteratively self-improving version of the algorithm described in [4]

[6] In this notation, `A23+B2` for example means that the corresponding syllable carries an `A2` or an `A3` accent and is immediately preceding a B2 boundary.

ACKNOWLEDGEMENTS

This work was supported by the German Ministry for Research and Technology (*BMFT*) in the joint research projects ASL and VERBMOBIL. Only the authors are responsible for the contents.

REFERENCES

[1] G. Bakenecker, U. Block, A. Batliner, R. Kompe, E. Nöth, and P. Regel-Brietzmann. Improving Parsing by Incorporating Prosodic 'Sentence Breaks' into a Grammar. In *Int. Conf. on Spoken Language Processing*, in this volume, 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, (to appear). Springer–Verlag, Berlin, Heidelberg, New York, 1994.

[3] C. Féry. German Intonational Patterns. Niemeyer, Tübingen, 1993.

[4] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. DP-Based Determination of *F0* contours from speech signals. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages II–17–II–20, San Francisco, 1992.

[5] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. Detection of Phrase Boundaries and Accents. In *Proc. CRIM/FORWISS Workshop "Progress and Prospects of Speech Research and Technology"*, (to appear), München, 1994.

[6] 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.

[7] 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.

[8] K. Ross, M. Ostendorf, and S. Shattuck-Hufnagel. Factors Affecting Pitch Accent Placement. In *Int. Conf. on Spoken Language Processing*, volume 1, pages 365–368, Banff, 1992.

[9] E.G. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, and S. Rieck. Automatic Speech Recognition without Phonemes. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 111–114, Berlin, September 1993.

[10] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. TOBI: A standard for labeling English prosody. In *Int. Conf. on Spoken Language Processing*, volume 2, pages 867–870, Banff, 1992.

[11] S. Uhmann. Fokusphonologie. Eine Analyse deutscher Intonationskonturen im Rahmen der nicht-linearen Phonologie. Niemeyer, Tübingen, 1991.

[12] 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, September 1993.

[13] A. Waibel. Prosody and Speech Recognition. Morgan Kaufmann Publishers Inc., San Mateo, California, 1988.

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