Gefördert vom
bmb+f
Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie

Verbmobil
Verbundvorhaben

Tempo and its Change in Spontaneous Speech.

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

F.–A.–Universität Erlangen–Nürnberg

Report 216
September 1997

September 1997

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

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

Tel.: (09131) 85 - 7883
e-mail: batliner@informatik.uni-erlangen.de

Gehört zum Antragsabschnitt: 1.7.2

Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministers für Bildung, Wissenschaft, Forschung und Technologie unter dem Förderkennzeichen 01 IV 701 K/5 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den Autoren.
ISSN 0949-6084

TEMPO AND ITS CHANGE IN SPONTANEOUS SPEECH

A. Batliner   A. Kießling   R. Kompe+   H. Niemann   E. Nöth*

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

* now with Ericsson Eurolab, Nürnberg

+ now with Sony Stuttgart Technology Center, Fellbach

E–mail: batliner@informatik.uni-erlangen.de

www: http://www5.informatik.uni-erlangen.de

ABSTRACT

In this paper, we give a first account of speech tempo and its change in spontaneous speech in a very large data base (Verbmobil, i.e., human-human appointment dialogs). As features representing speech tempo, we computed mean normalized speech duration (speaking rate) and normalized phone duration in different ways. The importance of these features is evaluated with an automatic classification of boundaries and accents where different sets of prosodic features (including also information about F0, energy, pause, etc.) were used. The best results (83% for accents, 88% for boundaries, two classes each) could be achieved when all features were used. For the 2nd issue change of tempo was labelled manually. We present the characterizing feature values for changes from slow to fast and from fast to slow, as well as the results of an automatic classification of change of tempo (72% for three classes). Finally, we discuss the possible function of change of tempo and its use in automatic speech processing.

1. INTRODUCTION

Speech tempo characterizes first of all an individual speaker who, however, can vary it either in order to signal different emotional states or in order to use it for different rhetoric functions, as, e.g., planning, holding the floor, etc. This means for listeners, that they have to calibrate their perception for the specific speech tempo of a speaker, in particular if this tempo is extraordinary fast. Above that, they can notice the *overall change of tempo* (e.g., as an indication of emotion [12]) or the *local change of tempo* (e.g., as an indication of the structuring of the dialog [10, pp. 390]). In analogy, taking into account tempo and its change in automatic speech processing can be useful for word recognition, for the classification of suprasegmental events (prosodic marking of accents and boundaries), and for semantic and dialog analysis. Tempo can be used as a basis for the normalization of phone duration.

Up to now, almost all phonetic studies on speech tempo were based on controlled, elicited speech [11]. This holds for psycholinguistic studies on planning strategies as well; studies on emotion were sometimes based on 'real life' speech; however, they did not use strict measurement procedures.

2. COMPUTATION OF SPEECH TEMPO

In [5] it is shown that mean and standard deviation of phone duration depend both roughly linearly on speech tempo; this means that we can model duration with gamma distributions [4]. Faster speech tempo thus co-varies with a reduction of mean phone and syllable duration and their standard deviation. [13] introduced the *mean normalized speech duration* τ, that can be computed for a longer stretch of speech, e.g., for a phrase or for the whole turn; cf. Eqn. (1). With this mean normalized speech duration, we can compute the *mean normalized phone duration* DURATION$_{norm}$ of phones, syllable nuclei, words, etc., cf. Eqn. (2).

$$\tau = \frac{1}{k}\sum_{i=1}^{k}\frac{d(i)}{\mu_{phone(i)}} \qquad (1)$$

$$DURATION_{norm} = \frac{1}{l}\sum_{i=1}^{l}\frac{d(i)-\tau\cdot\mu_{phone(i)}}{\tau\cdot\sigma_{phone(i)}} \qquad (2)$$

In these equations the following parameters are used:

$phone(i)$	the phone type of the i-th phone segment
k	number of phones in the stretch of speech, without pauses
$d(i)$	duration of i-th phone segment with phone type $phone(i)$
$\mu_{phone(i)}$	mean duration of $phone(i)$
$\sigma_{phone(i)}$	standard deviation of the duration of $phone(i)$
l	number of phones in the syllable (for syllable nuclei: one phone)

The phone intrinsic values $\mu_{phone(i)}$ und $\sigma_{phone(i)}$ were estimated beforehand with the help of a training data base. Three different contexts of a phone were taken into account for the estimation of μ and σ while in any case Eqn. (2) is used for the computation of the duration.

- For DURATION$_{norm}$, we use all tokens of $phone(i)$ in the whole sample for the estimation of $\mu_{phone(i)}$ und $\sigma_{phone(i)}$.
- For DURATION$_{norm}^{word\ acc}$, we estimate different $\mu_{phone(i)}$, $\sigma_{phone(i)}$ depending on whether the syllable with the pertinent phone is the carrier of the lexical word accent or not. For each phone class, we thus estimate two (μ,σ)–pairs out of the training sample.
- For DURATION$_{norm}^{syll\ pos}$, we again estimate different $\mu_{phone(i)}$, $\sigma_{phone(i)}$; this depends on the fact whether the phone is in a monosyllabic word or in the word initial, word final or word internal syllable of a polysyllabic word. As it is additionally distinguished if the syllable carries the lexical word accent or not, this results in the estimation of in total eight (μ,σ)–pairs for each phone class.

All these normalizations are *explicit* normalizations for the duration of phones, syllables, words, etc. In analogy, by taking into account the same three different contexts, we can compute the mean normalized speech duration of the turn in three different ways by using the corresponding $\mu_{phone(i)}$ in equation 1. These values are estimates for the speaking rate of the phrase/turn etc. and can be directly used in the feature vector for the sake of an *implicit* normalization of the other prosodic features in the vector.

[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 would like to thank Elisabeth Maier, DFKI, Saarbrücken, for analyzing the dialog acts adjacent to tempo change.

feature sets (SET)	number of features	SET alone				ALL \ SET			
		A | ¬A		B3 | B[029]		A | ¬A		B3 | B[029]	
		$\mathcal{ER}$	$\mathcal{ER}_{\overline{K}}$	$\mathcal{ER}$	$\mathcal{ER}_{\overline{K}}$	$\mathcal{ER}$	$\mathcal{ER}_{\overline{K}}$	$\mathcal{ER}$	$\mathcal{ER}_{\overline{K}}$
ALL	276/0	**82.6**	**(82.2)**	**88.3**	**(86.8)**	—		—	
DURATION$_{all}$	60/216	74.9	(74.7)	78.7	(77.7)	81.7	(81.4)	83.9	(85.1)
GLOBAL DURATION	3/273	50.4	(51.3)	48.6	(54.9)	82.0	(81.5)	87.7	(86.2)
		SET alone				ALL \ DURATION$_{all}$ ∪ SET			
DURATION$_{non\ norm}$	15/231	67.0	(67.0)	74.4	(75.0)	82.2	(81.8)	85.6	(84.8)
DURATION$_{norm}$	15/231	69.5	(69.2)	72.3	(74.1)	81.8	(81.4)	87.2	(85.1)
DURATION$_{norm}^{word\ acc}$	15/231	66.7	(66.2)	72.9	(73.6)	82.4	(82.0)	86.4	(85.2)
DURATION$_{norm}^{syll\ pos}$	15/231	68.5	(67.7)	71.9	(73.3)	82.0	(81.6)	85.4	(84.6)

Table 1. Recognition rates for the classification of accents (A | ¬A) and prosodic boundaries (B3 | B[029]) with different feature sets. Best results with all features are depicted with bold face.

3. CLASSIFICATION OF ACCENTS AND BOUNDARIES

Speech tempo influences in particular the duration of phones and words and thus the prosodic realization of accents (lengthening of prominent syllables) and of boundaries (phrase final lengthening). We were therefore interested in the question whether an explicit or implicit normalization of speech tempo improves the automatic classification of accents and boundaries. To our knowledge, this is the first study that uses tempo parameters for the automatic classification of accents and boundaries in spontaneous speech; as for read material, cf. [13].

We used a subsample of the Verbmobil (VM) data base for which perceptual prosodic labels (± accent, ± boundary) are available: 30 dialogs for training and 3 dialogs for testing (in total, 861 turns, 2 hours of speech) [8]. Based on an automatic time alignment of the spoken word chain, a huge feature vector (276 prosodic features) was computed for each syllable and each word that encodes the prosodic properties (F0, duration, energy, pause, etc.) of the actual word and its context (the maximum context was ± two syllables or words). Multi–layer perceptrons (MLPs) with different topologies were trained and tested for the classification of accents and boundaries. 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. The a priori probability is not modelled in the MLP. More detailed presentations of feature set and classification can be found in [7, 8, 9].

Based on the best recognition rates obtained up to now for accents and prosodic boundaries [7, 8], we studied the influence of different normalizations (explicit normalization) and the use of the mean normalized speech duration τ as a feature (implicit normalization). The results of these experiments are given in Table 1 for the classification of prosodic boundaries (strong boundary B3 vs. [weak boundary B2, no boundary B0, irregular boundary B9]) and for the classification of accents (accented A vs. not accented ¬A). For each feature set, the best result achieved is displayed in Table 1. In column 'SET alone', the results for the specific feature set are given, in column 'ALL \ SET', the complements are given, i.e. recognition rates for all features *without* those specified in column one. For the results in column 'ALL \ DURATION$_{all}$ ∪ SET', we only used the 216 'non–durational features' together with those 15 durational features specified in column one. Mean recognition rate $\mathcal{ER}$ and, in parentheses, the mean of the class–dependent computed recognition rate $\mathcal{ER}_{\overline{K}}$ are given.

In the experiments, we used three mean normalized speech durations (row 'GLOBAL DURATION') that were computed for the whole turn, cf. section 2: τ, τ taking into account position of accent, and τ taking into account the position of the syllable within the word. Of course, it does not make much sense to use these features for the classification of accents or boundaries without taking into account other prosodic features (row 'GLOBAL DURATION', column 'SET alone'). These figures are only given for completeness. If we compare row 'ALL' with 'ALL \ GLOBAL DURATION' we can see that the implicit normalization of the features with the help of the global durations leads to an improvement of the recognition rates for both accents and boundaries. At first sight this is surprising because the durational features have been normalized already. The reason might be that global duration is important not only for durational but also for F0 and energy features because all of them co–vary with speech tempo up to a certain extent as well.

For the classification of accents (A | ¬A), the normalized durational features (DURATION$_{norm}$) for 'SET alone' are better than the other durational features; in combination with the other 216 'non–durational features', however, the normalization that takes into account position of word accent (DURATION$_{norm}^{word\ acc}$) yields the best result. We can achieve the best result if we use all 60 durational features (row 'DURATION$_{all}$', column 'SET alone').

If we only use durational features for the classification of boundaries (B3 | B[029]), slightly better results could be achieved for the non–normalized durational features (row 'DURATION$_{non\ norm}$', column 'SET alone'); in combination with the other 216 'non–durational features', DURATION$_{norm}$ yields slightly better results. If we use all 60 durational features (row 'DURATION$_{all}$', column 'SET alone') instead of only one type of normalization, the results are always better. For example, compared with the 15 non normalized features (row 'DURATION$_{non\ norm}$'), the error rate could be reduced by about 17 % by using all 60 durational features (row 'DURATION$_{all}$').

Summing up it can be established that an implicit normalization of the acoustic–prosodic features with the speech duration yields better classification results both for accents and for boundaries: ca. 1 % for the classification of accents and ca. 4 % for the classification of boundaries. It is of minor importance which special normalization is used. In any case, using both normalized *and* non normalized durational features yields much better results than using just one special normalization.

4. CHANGE OF TEMPO

4.1. Annotation

For the investigation of change of tempo, we used almost all turns of the first five CD-ROMs of the VM data base for which syntactic–prosodic labels [3] are available. One phonetic expert listened to the turns; a clear change of tempo lasting for more than two words was labelled with TCA (Tempo Change Allegro, i.e., change from slower to faster) or with TCL (Tempo Change Lento, i.e., change from faster to slower). The labels are associated to the word boundaries. In the following, TC is used as cover term for all tempo changes (TCA and TCL), and T0 is used for the complement, i.e., for any other word boundary without a TC. Note that we only took into account TC on the 'macro' level, not on the 'micro' level, as, e.g., final lengthening or hesitation. The latter phenomena are already labelled in the VM basic transliteration and represent a sort of arhythmic 'stumbling', not a real change of tempo. In order to reduce the effort needed for the annotation, the endpoint of a TC is not labelled; normally,

features	context TCA (# = 128)			context TCL (# = 69)		
	before	at TCA	after	before	at TCL	after
duration, norm.	0.30	0.76	-0.74	-0.59	0.34	0.68
duration, abs.	0.24	0.55	-0.52	-0.36	0.13	0.35
energy, mean	-0.07	-0.10	0.37	0.13	-0.16	0.04
energy, reg.coeff.	-0.05	-0.02	0.73	-0.13	-0.19	0.39
F0, mean	-0.15	-0.10	0.30	-0.16	-0.13	0.05
F0, reg.coeff.	-0.06	0.11	0.27	-0.11	0.03	0.00

Table 2. Mean values of relevant features before/at/after TCA/TCL

	# tokens	TCA	TCL	T0
TCA	72	78	7	15
TCL	39	10	75	15
T0	2332	19	16	65

Table 3. Recognition rates for TC in percent

it is at the next strong syntactic boundary, but in some cases, it cannot be localized precisely. The data base comprises 362 speakers, 7286 turns with 149.643 words, and 322 TCs (208 TCA and 114 TCL). 80 speakers (22%) used at least one TC; in 201 turns (2.8%) occurred at least one TC. A TC is thus rather a speaker specific phenomenon because 78% of the speakers did not use it.

4.2. Feature Values

From now on, we only inspect those 201 turns with at least one TC. In Table 2, the mean values of duration, energy, F0, and the mean values of the regression coefficients of energy and F0 are given for the training set (cf. section 4.3) computed for the largest context of six syllables/three words which yields the best recognition rate, cf. Table 4. First, we display the mean value for the three words *before* the word with a TC, then the value for the word *at* the TC, and then the value for the three words *after* the TC. The respective values for T0 (n=3816) cluster around zero and are therefore not given in the table. The values can be interpreted as follows: For a TCA, the tempo slows down towards the TC, esp. for the word immediately at the TC; then, the tempo gets faster. For a TCL, it is the other way round: The last three words are markedly faster than the mean; then the tempo slows down. Both energy and F0 behave similarly for the left context of TCA and TCL: before the TC, the values are equal or slightly lower than for T0; after a TCA, they are higher, and after a TCL, they are lower than for T0. The same holds for the regression coefficients. This correlation of the feature values with each other might partly be automatic (i.e., a co–variation of redundant with distinctive features) and partly controlled by the speaker. We can say that the speakers in a way 'swing back', i.e., they behave antagonistically before the TC. They slow down before a TCA and speed up before a TCL. This does not mean, of course, that this is always the case; for the moment, it only means that the criterion for the annotation of a TC 'clear change of tempo' is met especially by these TCs with an antagonistic leading phase. If such a behavior, however, characterizes a TC in general, it will surely facilitate automatic processing of these phenomena.

4.3. Automatic Classification

For the automatic classification of TCs, we used the same feature set and classifiers as described above. Table 3 shows the result of an automatic classification with 421 features. This feature set yields the best classification (cf. Table 4). Note again, that for training and test only those turns were used that contained at least one TC. The TCs of CD–ROM2 formed the training set (123 turns), all other turns the test set (71 turns). Note that our classes are ordered: T0 should be between TCA and TCL and is thus prone to be more confused with them than TCA with TCL and vice versa. For the best classification of 72%, such a 'bad' confusion of TCA with TCL only occurred in 9 cases (i.e. 0.4% of all word boundaries).
Table 4 shows the class-dependent mean recognition rates with different feature sets. The context considered (#syllables, #words) and thus the number of components in the feature vector increases from left to right. The recognition rate for the three classes T0, TCA, TCL was 57% to 72%, depending on the context for which the features were computed: the larger the context, the better the recognition. (Due to limited resources no larger context than six syllables/three words was considered here; further improvement is expected when using more context.) This result mirrors in a way the strategy of the labeller only to annotate TCs on the macro level.

4.4. Tempo Change at Syntactic Boundaries

For this data base, there exist syntactic–prosodic labels that denote boundaries with different strength; 'syntactic–prosodic' means, that the criteria are mainly syntactic but that prosody is taken into account up to a certain content; as for details, cf. [3]. TCs are triggered by planning processes, not by the different strength of syntactic boundaries. A similar phenomenon is filled pauses: in [1], we report that 9/10th of filled pauses can be found at syntactic boundaries. A systematic evaluation of our data shows that TCs behave in a similar way: they can mostly be found at different types of strong syntactic boundaries and do not tell apart different types of syntactic boundaries. They occur mostly at boundaries that mark sentences, parentheses, and discourse particles (as *'well', 'ok'* with following pauses) that are prototypical candidates for positions where some planning goes on. It turns out that TCs are seventeen times more frequent at syntactic boundaries than at 'normal' word boundaries. A more detailed account can be found in [2].

4.5. The Function of Tempo Changes

In some cases, a TC is the only means to disambiguate between two syntactic readings. The 'classic' case is certainly parentheses which, however, are rather seldom in our material. TCs obviously co-occur very often with syntactic boundaries, but this is most certainly not the primary function intended by the speaker. The primary function could either be dialog–specific, i.e., control of turn taking, or semantic, i.e., indication of salient information, or rhetoric, e.g., variation of the speech tempo (in order to hold the attention of the speaker). It can as well be that there is no real function at all but that it is just as speaker–specific as, e.g., nasalisation, slurring, or the continuous use of allegro or lento speech can be. A null hypothesis might be that TCs are mainly be caused by planning processes in the same way as filled pauses, repairs, fresh starts a.s.o. are.
In Table 5, we cross–classify TCAs and TCLs with respect to two conditions: First, whether they occur 'solo' in a turn [–mixed], without any other TC, or together (alternating) with at least one other TC [+mixed]. Second, whether they can be found in a [+final] position. Such a position is either close to the end of the turn (EOT), i.e., at the last syntactic boundary before the EOT, or it *could* be close to the EOT because essential information is already given. TCs at a [–final] position are either inside a constituent *(syntactic criterion)*, or an EOT close to them would make no sense because essential information is still missing *(dialog specific criterion)*. [+final] TCs are thus followed by an addendum that is not strictly necessary in order to produce a felicitous turn – a turn that the dialog partner is content with and can respond to. In about ten instances, this addendum is after a TCA and before the EOT and fulfills rather a phatic function, as "*...so let's meet on Saturday* TCA *will that suit you* EOT" because in a co–operative and symmetric communication, such a suggestion must always be confirmed by the partner. Such instances might be the reason why for TCAs in the [–mixed] condition, there are almost twice as many in the [+final] condition compared with the [–final] condition. The TCLs in the [–mixed] condition, however,

# syllables	0	1	2	1	2	3	3	4	5	5	6
# words	0	0	0	1	1	1	2	2	2	3	3
# features	45	83	121	127	165	203	251	289	327	383	421
rec.rate (%)	57	61	66	62	67	69	69	66	68	70	72

Table 4. Class-dependent mean recognition rate in percent

	TCA		TCL	
	-mixed	+mixed	–mixed	+mixed
+final	63	51	8	51
–final	36	58	7	48

Table 5. Frequency of TCs in [±mixed]/[±final] condition

are equally distributed across [±final] but only a few compared with TCAs in the same conditions and with TCLs in the [+mixed] condition. This might be caused by the fact that the overall setting of the scenario generally favors a 'planning' behavior and thereby a general slow speaking style (compared with the individual speaking rate of the speaker, of course). TCAs might therefore be more pronounced than TCLs — and thus more prone to be perceived as such and annotated at all. In the [+mixed] condition, the variation between TCA and TCL might be a rhetoric means that gives rise to the one and the other ('what goes up must come down'). We can see, that the [+mixed] cells are almost equally distributed for TCA and TCL, and for [+final] and [–final].

4.6. Tempo Change and Dialog Acts

In VM, the dialogs are annotated with dialog acts (DAs) that denote illocutionary force and can be subcategorised as for their functional role or their propositional content; for details, cf. [6]. We compared the distribution of those DAs that followed a TCA in the [–mixed] condition with the overall distribution of the DAs in a reference sample, namely CD–ROM7 (6621 DAs).[2] In Table 6, the respective frequencies of occurrence in percent are given for some important DAs whose frequencies differ considerably. (Note that the columns do not sum up to 100% because infrequent DAs are not given in the table.) TCAs do not occur at all at DAs that establish the *social setting.*[3] In the crucial *decision phase*, not many TCAs can be observed. In contrast, they are frequent in the *negotiation phase*, when DAs *elicit* information but do not give *salient new* information.

dialog phase	dialog acts	TCA	ref.
establishment of social settings	greet	0	4
	introduce_name	0	4
	bye	0	4
	thank	0	1
negotiation phase	suggest_date	34	23
	request_comment_date	22	3
	request_suggest_date	8	2
	clarify_query	14	1
decision phase	accept	4	10
	reject	1	4

Table 6. Frequency of dialog acts in percent

4.7. General Discussion

Our results with respect to the function of TCs are consistent with the 'classic' interpretation that TCLs mainly indicate planning processes. TCAs are, so to speak, the *results* of planning processes. Their primary function might really simply be to hold the floor. Their secondary function might be a rhetoric one, namely the variation of speech tempo that mainly takes place in the *negotiation* phase of a dialog. At least in our data base, TCs do not occur very often. If this holds across other data bases as well, it might not be worth while to model them for automatic recognition and understanding. They can, however, be used in generation and synthesis in order to produce more natural speech. For such an application, our results can give clues for an adequate use of TCs.

5. CONCLUSION

We have shown that speech tempo should be modelled in automatic speech understanding systems because by that, we could improve the automatic classification of boundaries and accents to a considerable extent. If the feature vectors are continuously calibrated w.r.t. the individual speakers, we cannot only take into account the individual speech tempo of each speaker but *overall* change of speech tempo within one speaker as well.

Local change of speech tempo is an interesting topic by its own. At least for the data base used, there are, however, only a few tokens, so it might not be worth while to model it explicitely in the recognition/understanding phase but only in the generation/synthesis phase. Matters might change with other domains or with other languages. The reason for this distribution could of course be the strategy of our labeller only to label very pronounced TCs. Without the investigation of other data bases we can, however, not decide whether it is not simply the case that TCs are just one – out of many – means to fulfill some rhetoric functions.

[2] TCLs in the [–mixed] condition are only a few, and in the [+mixed] condition, the rhetoric function established in the last section could obscure the distribution of the DAs. We therefore confine our analysis to TCAs in the [–mixed] condition.

[3] Note that greet is normally at the very beginning of a turn; this fact prevents of course a leading TCA because to its left, there is no word sequence whose speech tempo could be compared with the sequence to its right.

REFERENCES

[1] A. Batliner, A. Kießling, S. Burger, and E. Nöth. Filled Pauses in Spontaneous Speech. In *Proc. ICPhS*, volume 3, pages 472–475, Stockholm, 1995.

[2] A. Batliner, A. Kießling, and R. Kompe. Tempo und Tempowechsel in Verbmobil–Dialogen. Verbmobil Memo 110, 1996.

[3] A. Batliner, R. Kompe, A. Kießling, H. Niemann, and E. Nöth. Syntactic–prosodic Labelling of Large Spontaneous Speech Data–bases. In *Proc. ICSLP*, volume 3, pages 1720–1723, Philadelphia, 1996.

[4] T.H. Crystal and A.S. House. Segmental durations in connected–speech signal: Preliminary results. *JASA*, 72:705–716, 1982.

[5] T.H. Crystal and A.S. House. Segmental durations in connected–speech signal: Current results. *JASA*, 83:1553–1573, 1988.

[6] S. Jekat, A. Klein, E. Maier, I. Maleck, M. Mast, and J. Quantz. Dialogue Acts in Verbmobil. Verbmobil Report 65, 1995.

[7] 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 3rd CRIM / FORWISS Workshop*, pages 104–113, Montreal, 1996.

[8] A. Kießling. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*. Berichte aus der Informatik. Shaker Verlag, Aachen, 1997.

[9] 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. EUROSPEECH*, volume 2, pages 1333–1336, Madrid, 1995.

[10] W. Levelt. *Speaking: From Intention to Articulation.* MIT Press, Cambridge, MA, 1989.

[11] S.G. Noteboom. Some observations on the temporal organization and rhythm of speech. In *Proc. ICPhS*, volume 1, pages 228–237, Aix–en–Provence, 1991.

[12] K. Scherer. How Emotion is Expressed in Speech and Singing. In *Proc. ICPHS*, volume 3, pages 90–96, Stockholm, 1995.

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