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

Verbmobil
Verbundvorhaben

Can We Tell apart Intonation from Prosody (if we Look at Accents and Boundaries)?

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

F.–A.–Universität Erlangen–Nürnberg

Vm
Report 215
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

CAN WE TELL APART INTONATION FROM PROSODY (IF WE LOOK AT ACCENTS AND BOUNDARIES)?

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

Studies on prosody/intonation normally look for important (distinctive) features denoting linguistic contrasts in production or perception experiments. The recent development in automatic speech processing and the availability of large speech data bases made it possible to have a fresh look at these topics. In our study, we classify automatically accent and boundary positions in a spontaneous speech corpus with a large feature vector comprising as many relevant prosodic features as possible. The results obtained for different subsets of prosodic features (F_0, duration, energy, etc.) show that each feature class contributes to the marking of accents and boundaries, and that the best results can be achieved by simply using all feature subsets together. Finally, we discuss possible conclusions for prosodic theory and for the application of prosody in speech processing.

1. INTRODUCTION

1.1. Theories and Methods

A time-honored topic in phonology is the distinction between dinstinctive and redundant features, e.g., in intonational research the question, whether tones or movements are the appropriate units, and in prosodic research, whether we can do with intonation (i.e. pitch and F_0, resp.) alone or whether we should model all prosodic parameters, i.e. duration, energy etc. as well, in a unified approach. In phonetics/phonology, most of the models on suprasegmentals nowadays are models of intonation, not of prosody: the tone–sequence approach, the IPO approach, the Fujisaki model, the Lund model, etc.; cf., e.g., [8]. Of course, other prosodic parameters are more or less implicitly included as well in intonation models because after all, speech and thus intonation are time phenomena, and that means that prominent intonational events (F_0 movements, H*/L* tones, etc.) are aligned to certain points on the time axis. In the statistical models used in automatic speech understanding (ASU), the question of distinctive vs. redundant is boiled down to the problem of feature evaluation: it is of minor interest what class a feature belongs to; as long as it is a good predictor, it is included into the feature vector the classifier is based on. Empirically minded phonologists and phoneticians sometimes design perception experiments as an *experimentum crucis* that should help finding *the* distinctive features. In other perception experiments and generally in production experiments, the contribution of various features at stake are investigated using statistical methods as, e.g., regression analysis or linear discriminant analysis. In a way orthogonal to the concept of distinctiveness is the concept of *trading relations* where the smaller value of one parameter can be compensated by a greater value of another parameter, cf. [12].[2]

In all these cases, there is, however, a systematic missing link to real–life, spontaneous speech: in an experiment, one can force subjects to pay more attention to the one or to the other feature, but one never knows, whether their experimental behavior really mirrors their behavior in real–life. The normal speaker/hearer does not compute, let's say, a F_0 analysis, and does certainly not tell apart analytically pitch from duration from energy etc. but he/she unifies every percept into one single concept - e.g.: this word is accented, that word is phrase–final.

The results of such *in vitro* experiments are most of the time discussed on an *as if* basis: as if they would be able to decide between hypotheses or theories. In the long run, it is, however, just a matter of *cumulative evidence*. Recently, advances in ASU and the availability of large spontaneous speech data bases made it possible to have a new look at the relationship of intonation with other prosodic parameters. A cumulative evidence can be obtained as a sort of byproduct from ASU if we look closer at single features, feature groups and their respective relevance for the classification of speech events as, e.g., accents and boundaries.

1.2. The Applied Approach

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 trading relation between prosodic parameters
- 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 use a statistical approach and propose a method where as many relevant prosodic features as possible 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 labelling of prosodic events which has to be synchronous with the stream of feature vectors, statistical classifiers (here: multi-layer perceptrons, MLP) are trained using a training database and are evaluated on a different database.

The material our investigations are based on is to our knowledge so far the largest spontaneous speech data base

[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 Grant 01 IV 701 K5. The responsibility for the contents of this study lies with the authors.

[2] Note that all these concepts we mentioned here are not exactly at the heart of the scientific debate today, but we believe, that they - implicitly - still form a common ground.

used for such a purpose, and the feature vector is to our knowledge the largest one as well. Note that 'largest' does not necessarily mean 'best' but in fact, we will argue for exactly this conclusion.

2. MATERIAL

The research presented in this paper has been conducted under the VERBMOBIL project, cf. [10], which aims at automatic speech–to–speech translation in appointment scheduling dialogs. The 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 labelling of boundaries and accents was performed by our VERBMOBIL partner University of Braunschweig, cf. [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 labelled explicitly; B9: *'agrammatical' boundary*, e.g., hesitation or repair. Four different types of syllable based accent labels are distinguished as well 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*, and A0: *any other syllable*, not labelled explicitly. Here, 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. 33 VERBMOBIL dialogs (approx. 2 h of speech) have been labelled along these lines.
All recognition results described in the following were obtained on the same test set comprising three VERBMOBIL dialogs (64 turns of 3 male and 3 female speakers, 12 minutes in total). For the training of the MLPs 30 disjoint dialogs (797 turns of 53 male and 7 female speakers, 100 minutes in total) were used. The recognition rates for the accent classification are evaluated for all words of the test set; the rates for the boundary classification are only determined for the turn–internal words without taking into account the turn–final words because a detection of a boundary at the end of a turn is rather trivial.

3. EXTRACTION OF PROSODIC FEATURES

We distinguish different categories of prosodic feature levels. 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. Within this group we can further distinguish:

- *basic prosodic features*
  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 (syllable, syllable nucleus, word, turn, etc.) 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 taken from the output 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 two categories:

- *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 word–final 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, or they can be based on predictions of possibly important – and thus accented – words from the semantic or the dialog module.

All these categories are dealt with in more detail in [5]. Here, 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.[3] We only use the aligned spoken words thus simulating 100% word recognition – and by that, simulating the capability of a human listener. The time alignment is done by a standard hidden Markov model word recognizer.
Due to the problems discussed so far, 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. Generally, we therefore try to be as exhaustive as possible, and leave it to the statistical 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. In prior studies, we investigated different contexts of up to ± 6 syllables (± 3 words, resp.) to the left and to the right of the actual word–final syllable. For every classification problem investigated many different subsets of these features were analyzed. The best results so far for the $\mathsf{B}|\neg\mathsf{B}$ and the $\mathsf{A}|\neg\mathsf{A}$ problem were achieved by using 276 features computed for each word considering a context of ± 2 syllables (± 2 words, resp.).[4] Note, that in contrast to the accent classification experiments in [6], here we chose a different strategy: in [6] we computed one feature vector for each syllable performing a syllable-based $\mathsf{A}|\neg\mathsf{A}$ classification. As it is more important for semantic analysis 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 $\mathsf{A}|\neg\mathsf{A}$ classification. Experiments on the same data bases as in [6] showed a reduction of the error rate of about 20 % compared to the syllable-based approach. In more detail the features used here are:

- 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
- F_0-offset + position for actual and preceding word
- F_0-onset + position for actual and succeeding 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

[3] In [1], we use polygram language models (i.e. N-grams of variable length) for modelling the higher level information and combine the output of these classifiers with the MLP output which uses more or less only acoustic prosodic features.

[4] A full list of these features can be found in [5], pp. 257-258.

feature set (SET)	number of features	SET alone				ALL $\setminus$ SET			
		A | ¬A		B | ¬B		A | ¬A		B | ¬B	
		$\mathcal{ER}$	$\mathcal{ER}_{\overline{K}}$	$\mathcal{ER}$	$\mathcal{ER}_{\overline{K}}$	$\mathcal{ER}$	$\mathcal{ER}_{\overline{K}}$	$\mathcal{ER}$	$\mathcal{ER}_{\overline{K}}$
ALL	276	**82.6**	**82.2**	**88.3**	**86.8**	—		—	
F_0, with POS	80	79.5	79.2	81.4	82.1	81.7	81.5	85.0	84.9
F_0, without POS	56	76.2	75.3	78.7	75.9	82.4	82.0	86.3	85.9
F_0–MAX/MIN/ON/OFF, only POS	24	79.4	79.2	77.7	79.8	82.5	82.2	85.4	86.2
F_0–MAX/MIN/ON/OFF, without POS	24	73.9	73.1	78.6	73.1	81.9	81.6	86.0	85.0
F_0–REGRESSION	22	74.9	74.0	78.8	75.4	82.7	82.3	88.0	86.4
ENERGY, with POS	112	77.3	77.0	82.9	81.9	82.4	81.9	86.6	85.6
ENERGY, without POS	102	77.5	77.3	80.7	80.9	82.0	81.7	85.0	85.4
ENERGY, only POS	10	70.5	70.5	77.9	79.4	82.2	81.8	86.1	86.1
DURATION	60	75.4	75.2	78.7	77.7	81.7	81.4	85.8	85.4
PAUSE	6	*57.4*	*55.4*	*88.4*	*72.1*	82.3	82.0	86.6	85.1
SPEAKING RATE	3	*50.5*	*51.4*	*48.6*	*54.9*	82.0	81.6	87.7	86.2
FLAGS	15	78.6	78.8	74.3	74.9	81.7	81.3	86.6	85.6
maximum–minimum (without ALL)		9.0	8.7	8.6	9.0	1.0	1.0	3.0	1.5
ALL–*maximum*		3.1	3.0	5.4	4.7	-0.1	-0.1	0.3	0.4
reduction of error rate w.r.t. *maximum–minimum*		30.5	29.5	33.4	33.5	5.5	5.3	20.0	9.9
reduction of error rate w.r.t. ALL–*maximum*		15.1	14.4	31.6	26.3	-0.6	-0.6	2.5	2.9

Table 1. Recognition rates for the classification of accents (A | ¬A) and prosodic boundaries (B | ¬B) for different feature sets. All values are given in percent. Further explanations are given in the text.

- for each syllable and word in this context maximum energy (normalized as in [14]) + positions and mean energy (also normalized)
- duration (absolute and normalized) for each syllable/syllable nucleus/word
- length of the pause preceding/succeeding actual word
- 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 [14])
- for each syllable: flags indicating whether the syllable carries the lexical word accent or whether it is in a word final position.

4. EXPERIMENTS AND RESULTS

In this paper, we will only report results obtained with MLPs that turned out to be superior compared with Gaussian distribution or polynomial classifiers in similar investigations [4]. Different MLP topologies were analyzed for the two classification problems. As training procedure the Quick–propagation algorithm [2] with the 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 fed with an equal number of feature vectors from each class.

In Table 1 the results for experiments with different feature subsets of the best feature set (276 features, cf. above) is shown for the recognition of prosodic boundaries (column B|¬B) as well as for the classification of accents (column A|¬A). 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 $\setminus$ SET). Besides the overall recognition rate ($\mathcal{ER}$), the averages of the class-dependent recognition rates ($\mathcal{ER}_{\overline{K}}$) are given as well. Actually, $\mathcal{ER}_{\overline{K}}$ is more relevant than $\mathcal{ER}$ because the classifier was trained with an equally distributed a priori probability. Values in italics and those for ALL are not taken into account for the computation of the *maximum* and the *minimum* of the columns. All values are given in percent.

We distinguish three 'classic' main groups of features: F_0, ENERGY, and DURATION. Three further groups are PAUSE, SPEAKING RATE, and FLAGS. For F_0 and ENERGY, there are further subgroups: with/without/only position (POS). For F_0–MAX/MIN/ON/OFF, results are given for with/only POS. In order to make the results easier to interpret, we display the range *(maximum–minimum)* of each column without ALL as well as the range for ALL–*maximum*. The values of PAUSE and of SPEAKING RATE are not taken into account for the computation of *maximum* and *minimum* because this would make no sense: it is trivial that SPEAKING RATE alone is randomly distributed for accents and for boundaries, and that PAUSE is irrelevant for accents. For boundaries, PAUSE yields the best result for $\mathcal{ER}$, but the worst for $\mathcal{ER}_{\overline{K}}$. Note that the a priori distribution is not taken into account. This feature can thus 'model' the 'normal' word boundary, i.e., ¬B that occurs much more often than B, but not the distinction between ¬B and B. The decisive figures are the reductions of error rate for these two constellations *maximum–minimum* and ALL–*maximum*. We see that each single feature set yields results better than chance; there is, however, a marked difference between single feature sets: for the 'best' feature set, a reduction of the error rate between 29.5% and 33.5% can be achieved in comparison with the 'worst' feature set. The best single feature set is still markedly worse than ALL feature sets taken together (between 14.4% and 31.6% reduction of error rate for ALL in comparison with the best single feature sets). Each single feature set contributes to the overall recognition rate; this can be seen on the right side of the table (ALL $\setminus$ SET). The only exception might be F_0–REGRESSION: if we exclude only this feature set, we get slightly better results for A|¬A (0.6%, i.e., practically no difference). This feature set is highly correlated with other F_0 features; this and the fact that only a limited amount of training data was available might be responsible for this exception of the overall trend.

To speak about A|¬A and B|¬B makes only sense in a syntagmatic context because all features have to be related to this context (higher/lower, longer/shorter, etc., than the context?). We therefore modelled not only the respective words, but the words and syllables before and after as well. This means, however, that the succession of words with their respective feature values on the time axis is encoded automatically in one single feature vector. This fact is most certainly the reason why *all* single feature sets are markedly better than chance (70.5% and higher). It might as well be the reason why the positions of the prominent F_0 values alone 'F_0–MAX/MIN/ON/OFF only POS' are for A|¬A practically as good as all F_0 features with POS (79.4% vs. 79.5%).

For the B|¬B problem the most important features are F_0, ENERGY (and PAUSE, but cf. above). 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 with DURATION might be the fact that durational information is already modelled in the position features of F_0 and ENERGY, cf. the discussion in the last paragraph. This shows also the distinct drop of the recognition rate if only the 'pure' F_0 features without their positions (rows 'F_0 without POS' or 'F_0–MAX/MIN/ON/OFF, 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.

5. CONCLUDING REMARKS

If the distinction between distinctive and redundant features should make sense at all (at least for an application in ASU), distinctive features should be 'good predictors', and redundant features 'bad' ones. Is thus ENERGY distinctive for B|¬B, and F_0 not, and is it the other way round for A|¬A? In our opinion, such a conclusion does not make sense because we have seen that *all* features contribute to the distinction. Our results rather favor a sort of prototype model where no feature is distinctive in the traditional meaning and where all features being member of the relevant feature bundle can take over the role of each other up to a very great extent.[5]

Coming back to the title of this paper 'Can we tell apart intonation from prosody?', this question could be put in at least three different ways: (1) Which feature subset is a good (or the best) predictor for our classification? (2) *Can* we tell apart the contribution of intonational features from that of the other prosodic features? (3) *Should* we try at all to tell apart the one from the other? Question (1) cannot be answered unequivocally because we cannot really tell apart the contribution of single feature sets because of the intrinsic nature of speech (answer to question (2): no). Of course, question (3) can still be answered differently, depending on theoretical assumptions. As a matter of fact, in an application, we are not forced to tell apart intonation from prosody; in linguistic/phonetic theory, it is in our opinion an attractive alternative to the struggle for 'the best' model if we do not have to debate different ontologies but only different notational devices, e.g. labelling systems, that are more or less appropriate to fulfill different tasks. This means that we do not necessarily have to have the same approaches and units for production, generation and synthesis on the one side and for perception, (automatic) recognition and understanding on the other side. (In fact, this mirrors the state of affairs in prosodic research nowadays quite well.) It does, however, of course not mean that both approaches cannot profit from each other, cf. [9], [11].

Of course, some caveats have to be made: First, we only had a look at German prosody; things might be different for other languages where intonation plays a greater role, and especially for tone languages. Second, for the computation of our feature vector, we cannot take the prosodic phrase as a domain but 'only' a context that is supposed to be large enough, simply because we cannot use a unit as input that we want to detect. In practice, this disadvantage can be neglected. Third, even if we included many prosodic features into our feature vector we can of course not be sure whether we did not exclude the one or the other that could contribute to the prosodic marking and by that, to recognition as well. Note, however, that our recognition rates are in the range of the interlabeller consistency for similar tasks, cf., e.g., [3] and [13]. As such manual labels serve as reference labels for our experiments it is rather not likely that automatic recognition could be much better than human labellers. Fourth, the level of description/analysis could be challenged. In our experience, however, it is in any case the best strategy really to use 'raw' feature values and not features obtained from an intermediate level, as, e.g., from the phonological level.

[5] A rather pleasant consequence out of that is that we most certainly can take feature subsets for special applications without loosing too much information, if, e.g., in an incremental computation and classification, some of the other features cannot be computed.

A sort of 'external' validation of our approach is the fact that parsing in VERBMOBIL improved drastically with the use of prosodic boundary information, cf. [10, 7].

REFERENCES

[1] 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, October 1996.

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

[3] M. Grice, M. Reyelt, R. Benzmüller, J. Mayer, and A. Batliner. Consistency in Transcription and Labelling of German Intonation with GToBI. In *Proc. ICSLP*, volume 3, pages 1716–1719, Philadelphia, October 1996.

[4] 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, Sankt Augustin, September 1994. Infix.

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

[6] 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, September 1995.

[7] Ralf Kompe. *Prosody in Speech Understanding Systems.* Lecture Notes in Artificial Intelligence. Springer–Verlag, Berlin, 1997.

[8] R.D. Ladd. "Linear" and "Overlay" Descriptions: An Autosegmental–Metrical Middle Way. In *Proc. ICPhS*, volume 2, pages 116–123, Stockholm, August 1995.

[9] R. Moore. Twenty Things we Still Don't Know about Speech. 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 9–17. Infix, Sankt Augustin, September 1994.

[10] H. Niemann, E. Nöth, A. Kießling, R. Kompe, and A. Batliner. Prosodic Processing and its use in Verbmobil. In *Proc. ICASSP*, volume 1, pages 75–78, München, 1997.

[11] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel. Combining Statistical and Linguistic Methods for Modeling Prosody. In D. House and P. Touati, editors, *Proc. of an ESCA Workshop on Prosody*, pages 272–275. Lund University, Department of Linguistics, Lund, September 1993.

[12] B.H. Repp. Phonetic Trading Relations and Context Effects: New Experimental Evidence for a Speech Mode of Perception, 1981. Haskins Laborotories: Status Report on Speech Research SR-67/68.

[13] M. Reyelt. Consistency of Prosodic Transcriptions. Labelling Experiments with Trained and Untrained Transcribers. In *Proc. ICPhS*, volume 4, pages 212–215, Stockholm, August 1995.

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