From *Proc. of the 14th Int. Congress of Phonetic Sciences, San Francisco, August 1999, (pp. 2315–2318)*

PROSODIC FEATURE EVALUATION: BRUTE FORCE OR WELL DESIGNED?

Anton Batliner, Jan Buckow, Richard Huber, Volker Warnke, Elmar Nöth, Heinrich Niemann
University of Erlangen-Nuremberg, Chair for Pattern Recognition, Erlangen, Germany

ABSTRACT

In this paper we want to bridge the gap between phonetic/phonological theory on the one hand and automatic speech processing on the other hand. As material, we use a subset of the German VERBMOBIL database that is annotated with prosodic boundary and accent information. We computed a large prosodic feature vector: 276 features for a context window of up to five words modelling F_0, duration, energy, tempo, pauses, and linguistic information on the word level. Linear Discriminant Analysis (LDA) was used in order to minimize the number of features without too much loss in classification performance. This number could be reduced drastically from 276 to 11 for boundaries and to 6 for accents; the overall classification rate was only reduced by some two to three percent. We discuss the 'surviving' relevant features as well as limitations of this approach.

1. INTRODUCTION

Which features are most relevant for the marking of prosodic events such as phrase boundaries or phrase accents and should thus be used either for phonetic/phonological theory or for modelling in automatic speech processing? For this task, there are basically two opposite approaches: either a well designed phonetic/phonological experiment with a small number of controlled features and with read speech, or the applied, 'brute force' approach in automatic speech processing with large spontaneous speech corpora, many features, and automatic classification procedures. Both have their disadvantages: the first one cannot evaluate those (many) features that it does not control, the second one normally does not evaluate the contribution of single features up to a considerable extent. In other words: phonetics cannot see the wood for the trees, and vice versa, speech processing cannot see the trees for the wood.

Up to now, we computed in our research a very large set of 276 prosodic features, cf. section 3, and put them all into a neural network (NN); subsets of feature groups yielded always worse recognition rates than all features taken together [1]. Note, however, that this is no definite proof that a feature evaluation prior to the NN could not have worked better. The problem with NN is that they could be used for feature evaluation but not in a convenient way. For this task, it is better to use statistic procedures with 'built–in' evaluation strategies, as, e.g., LDA [4]. Other procedures that could be used for this task as well are, e.g., decision trees [8] which will not be dealt with in this paper.

In theory, to remove features that are highly correlated with other features should not necessarily result in loss of performance; in reality, however, this is quite often the case. The question is thus whether these cons can be counterbalanced by the pros (better modelling, better phonetic explanation, i.e., more knowledge, possibly better results in subsequent statistic analyses). In this paper, we will concentrate the discussion on the interpretation of those few most relevant features that 'survive' our evaluation procedure.

2. MATERIAL AND ANNOTATION

The research presented in this paper has been conducted under the VERBMOBIL project [5], which aims at automatic speech–to–speech translation in appointment scheduling dialogues. The experiments have been performed on subsets of this spontaneous speech database. For the training of classifiers, appropriate reference labels are needed. The perceptually based prosodic labelling of boundaries and accents was performed by our VERBMOBIL partner University of Braunschweig [6]. Four types of word–based boundary labels are distinguished: B3: *full boundary* with strong intonational marking, often with lengthening/pause; 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 which can be mapped onto word–based labels denoting if a word is accentuated 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 'accentuated word' ($\mathsf{A} = \{\mathsf{PA}, \mathsf{SA}, \mathsf{EC}\}$) vs. 'not accentuated word' ($\neg\mathsf{A} = \mathsf{A0}$), summing up the respective classes. Note that another clustering that, e.g., assigns the intermediate labels B2 and/or SA to B and ¬A, resp., would of course be possible as well. 33 VERBMOBIL dialogs (approx. 2 h of speech) have been labelled along these lines.

3. PROSODIC FEATURES

It is still an open question which prosodic features are most relevant for different classification problems and how the different features are interrelated. Generally, we therefore tried to be as exhaustive as possible and leave it to the classifier to find out the relevant features and the optimal weighting of them. Many prosodic features were therefore extracted over a prosodic unit and composed into a huge feature vector which represents the prosodic properties of this and of several surrounding units in a specific context.

For the computation of the prosodic features, a fixed reference point has to be chosen. We decided in favour of the end of a word because the word is the genuine domain in word recognition, and because this point can more easily be defined automatically than, e.g., the middle of the syllable nucleus in word accent position. A full account of the feature selection is beyond the scope of this paper. It is described in more detail in [3]. Our feature set is comparable to that used by [8, p. 475f] with the following differences: guided by our experience that raw values yield better recognition results than ratio values, we decided in favour of raw values. We use the same

Prosodic Features	
context	-2, -1, 0, +1, +2
domain	syllable nucleus, syllable, word
parameter	energy, F_0, duration, pause
value	minimum, maximum, onset, offset (in time domain (POS) and in frequency domain)
region	mean, RegCoeff, RegMSE
computation	absolute vs. normalized, local vs. global
Phonological/linguistic Flags	
position	word accent, syllable position in word (binary)
segment	phone/phone class

Table 1: Sketch of used prosodic/linguistic features

feature set for accent and boundary classification and leave it to the classifier to select the appropriate features for the specific task. For the same reason, we use less phonetic knowledge for a selection of different contexts the feature extraction is based on but chose almost always a fixed window size for the units before and after the reference unit. 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 $\pm$ 2 syllables/$\pm$ 2 words. We compute one feature vector per word, performing a word–based $\mathsf{A}|\neg\mathsf{A}$ classification, i.e., the position of an accentuated syllable within a word is given trivially with a lexicon look up. For syllable–based features, we have to determine the position of the phone boundaries which are, however, not given by the output of the word recognizer and thus have to be computed separately; in the future, we will only use word–based features because the position of the word boundaries is a by–product of word recognition.

Table 1 gives an overview of the used prosodic features; note that the last flag, segment, has a pronounced non–Gaussian distribution and is thus always excluded from the LDA. The features are described in the following; in square brackets, the relative values which the features can have for B and A compared with $\neg\mathsf{B}$ and $\neg\mathsf{A}$ are given; these values are used throughout in Tables 3 to 6:

- duration (absolute and normalized as in [7]) for each syllable nucleus/syllable/word *[shorter/longer]*
- for each syllable and word in this context
  - minimum and maximum of F_0 *[higher/lower]* normalized as to the F_0-mean (all F_0 values are interpolated at unvoiced stretches of speech and transformed into semitones) and their position (POS) relative to the reference point *[earlier/later]*
  - maximum energy, also normalized, *[higher/lower]* and its position relative to the reference point *[earlier/later]*, and mean energy, also normalized *[higher/lower]*
- F_0-offset *[higher/lower]* and its position (POS) *[earlier/later]* for the actual and preceding word (the F_0–offset is the last non-zero F_0 value in a segment)
- F_0–onset *[higher/lower]* and its position (POS) *[earlier/later]* for the actual and succeeding word (the F_0–onset is the first non-zero F_0 value in a segment)
- for each syllable in the considered context: flags indicating whether the syllable carries the lexical word accent *[$\pm$]* or whether it is in a word–final position *[$\pm$]*
- length of the pause preceding/succeeding the actual word *[shorter/longer]*
- linear regression coefficients RegCoeff of F_0 contour and energy contour over 11 different windows to the left and to the right of the actual syllable *[rising/falling]* and their mean square error RegMSE *[$\pm$ variation]*
- for normalization, measures for the speaking rate are computed over the whole utterance based on the absolute and the normalized syllable duration as in [7]. It is used to explicitly normalize the duration features and it is added to the feature vector for an implicit normalization of the other features [2].

4. CLASSIFICATION

For classification, we use an LDA provided by the statistic package SPSS8.01 for Windows. Due to memory limitations, in a first run, duration, energy and F_0 features, each with global features and flags, were evaluated separately with a tolerance criterion (wilks) of 0.01. This criterion excludes features from the analysis which are almost a linear combination of other features because they are highly correlated with them; the higher the criterion, the more features are excluded. By that, the 276 features were reduced to 88 for boundaries and 77 for accents. Then, all feature groups were taken together and the number of features was reduced by subsequent sharpening of the tolerance criterion (0.001, 0.01, 0.05,0.1, 0.5, 0.9). Analyses were conducted with and without linguistic flags, and with and without syllable–based features. In addition, we computed a principal component (PC) analysis and put the PC in a subsequent LDA. Table 2 shows the classification results for some of these analyses; we display overall recognition rate $\mathcal{RR}$ and recall for the classes separately. For comparisons with older results obtained with NN, ref.nr. (1) and (2) are given for the same independent test sample. For the NN, words/syllables at the end of turns were not considered for the classification of boundaries; more details can be found in [1]. All other results are given for a jack–knife procedure (leave one out) because here, all speakers are 'known' to the LDA and thus, strong speaker idiosyncrasies do not influence the results. For the LDA, the a priori probability of the two classes was set to 0.5.

In an LDA, two different coefficients are computed for each feature; the better the coefficient, the more important the feature. Both coefficients have, however, to be interpreted with care: *Standardized Canonical Discriminant Function Coefficients* [STAND] [...] *give us the variable's contribution to calculating the discriminant* score. [they] *take into consideration the simultaneous contributions of all the other variables. Structure Coefficients* [STRUCT] [...] *are simple bivariate correlations, so they are not affected by relationships with the other variables.* [...] *The perverse tendency of such situations to arise in discriminant analysis implies that the structure coefficients are* [at least: can be] *a better guide to the meaning of the canonical discriminant functions than the standardized coefficients are* [4, p. 33f].

For learn $\neq$ test, the LDA reduces the feature set from 276 to 57 and 46, resp., but yields worse results than the NN as well, especially for the 'marked' classes B and A, cf. ref.nr. (1) and (2). This might be traced back to the fact that in theory, there are 'redundant' features that can be explained fully by other features, but that in reality, this is not the case, cf. section 1. It might be as well that the NN is better at coping with non–Gaussian distribution. The same tendency can be observed if we sharpen the tolerance criterion (from 0.01 to 0.9) and by that, reduce the number of features used, cf. ref.nr. (3)/(5) and (8)/(9), resp. In comparison, ref.nr. (4)

and (6) display recognition rates for analyses with the same number of features, but this time, with the most relevant STRUCT features. Again, classification performance is reduced, but not to a large extent. Even if we only take those very few features that at the same time are STAND and STRUCT features for tol. $= 0.9$, cf. ref.nr. (7), the recognition rates are not too bad. A comparison of ref.nr. (8) and (9) with ref.nr. (3) and (5), resp., shows that leaving aside syllable–based information reduces the performance as well, but not to a large extent. Summing up the interpretation of ref.nr. (3) to (9) we can say that with every step we took, classification performance was reduced, but that the extreme reduction of 276 to three or two features, cf. ref.nr. (7), does not result in a break down of the performance, even if recall for the marked classes B (63.7%) and A (70.9% and 67.8%) is some 5 to 10 percent worse than with all 'relevant' features, cf. ref.nr. (3).

For such highly correlated features, it might be desirable first to conduct a PC analysis, and then use these PC which are not correlated with each other in a subsequent LDA. These results are given for 25 PC in ref.nr. (10) and (11); they are only slightly worse than those given in ref.nr. (8) for 41/40/37 features, and slightly better than those given in ref.nr. (9) for 12/5/8 features. The interpretation of these PC is very interesting but unfortunately beyond the scope of this paper.

5. INTERPRETATION

Tables 3 and 5 display for B and A the not or very low inter–correlated most relevant STAND features, Tables 4 and 6 those with the highly inter–correlated STRUCT features. The common features which are rather independent from the other features, cf. ref.nr. (7) in Table 2, are italicized. The subscript denotes syllable–based *(s)* or word–based *(w)* features. The STAND features cover the whole context from -2 to +2, i.e., *before*, *at*, *after*, the STRUCT features cover most of the time only the context 0, i.e., *at* (*no* context). This fact can be explained in a simple way: the marking takes place there where it belongs to, at the center, and it all points into the same 'direction'. Still, at the 'edges', i.e., at the contexts ± 1 and ± 2, something happens as well. If we do not look at the context structure but at the phonetic substance, then we can say that for the STAND features of B, energy, F_0, and duration contribute; this holds for A without flags as well. As for the STRUCT features, there is no F_0 feature amongst the most relevant ones for B. Note that POS features are actually duration features: if they are earlier, then the voiced part of the relevant domain is longer. We see that all feature groups contribute to the classification of B and A, and that F_0 based features are not the most important feature group; this corroborates our findings from [1]. In more detail, the four Tables can be interpreted as follows:

B, STAND, Table 3: Flags are not amongst the 11 relevant features; we therefore display here only the analysis without flags. Rising energy *after* can be seen as a sort of resetting of the energy contour. Prefinal lowering of F_0 might be the reason for the low F_0 min *at* and the overall falling F_0 contour *before*, *at*, and *after*. The resetting of the F_0 baseline might cause less variation *after*. Duration (i.e. F_0 POS earlier as well) is longer *at*, and shorter *before* and *after* (prefinal lengthening). Pauses as a standard feature for B are longer *at* the word boundary in question, and shorter *before*.

B, STRUCT, Table 4: There is more energy variation *at* and no F_0 feature, but most of the features (POS and duration) denote prefinal lengthening *at*; the four different duration features denote different normalized or not normalized computations.

A, STAND, Table 5: For analyses with flags, there is lower energy *after* and more energy variation *at*. F_0 is falling *before* and rising *at*; it might be that the overlapping window (-1,0) makes the feature value more consistent. As for the flags, there are less word–final syllables *before* and more word–final syllables *after*. This mirrors the fact that monosyllabic words are most of the time unaccentuated function words. For analyses without flags, duration so to speak takes over the role of the flags, and co–varies *at* with energy and F_0: there is always a 'more' of the features than *before* or *after*.

A, STRUCT, Table 6: The feature [– word accent] *at* again mirrors the fact that polysyllabic words are more often accentuated than monosyllabic ones. Again, duration takes over the role of the flags if these were not considered, and again, there is a 'more' of the features *at*: more energy variation, F_0 max higher, longer duration, etc.

6. CONCLUSION

We have shown that we can find a small set of most relevant, so to speak 'discriminating', features. Most important are of course those features that model the center of the context. All feature groups contribute, and so does left and right context, albeit to a lesser extent; F_0 features are not more important than energy or duration features. Our most relevant features are in accordance with phonetic theory and literature. They should, however, not exactly be taken as a 'final' set because all our features are extracted automatically. We therefore cannot be sure that a manually corrected feature set would not end up with some other relevant features which are (highly) correlated with our present features, also because of the mechanism of feature selection in the LDA, and because of the intrinsic nature of language: it is redundant, not just 'discriminating'.

ACKNOWLEDGMENTS

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, 01 IV 102 F/4, and 01 IV 701 K5. The responsibility for the contents of this study lies with the authors.

REFERENCES

[1] Batliner, A., Kießling, A., Kompe, R., Niemann, H., and Nöth, E. 1997. Can We Tell apart Intonation from Prosody (if we Look at Accents and Boundaries)? In G. Kouroupetroglou, editor, *Proc. of an ESCA Workshop on Intonation*, vol. 2, p. 39–42, Athens.

[2] Batliner, A., Kießling, A., Kompe, R., Niemann, H., and Nöth, E. 1997. Tempo and its Change in Spontaneous Speech. In *Proc. Eurospeech*, p. 763–766, Rhodes.

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

[4] Klecka, W.R. 1988. *Discriminant Analysis*. SAGE PUBLICATIONS Inc., Beverly Hills, 9 edition.

[5] Niemann, H., Nöth, E., Kießling, A., Kompe, R., and Batliner, A. 1997. Prosodic Processing and its use in Verbmobil. In *Proc. ICASSP*, vol. 2, p. 75–78, München.

[6] Reyelt, M. 1995. Consistency of Prosodic Transcriptions. Labelling Experiments with Trained and Untrained Transcribers. In *Proc. ICPhS*, vol. 4, p. 212–215, Stockholm.

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

[8] Wightman, C.W. and Ostendorf, M. 1994. Automatic Labeling of Prosodic Patterns. *IEEE Trans. on Speech and Audio Processing*, 2(3):469–481.

ref.nr.			boundaries without flags				accents with flags				accents without flags			
	features	tol.	# f.	$\mathcal{RR}$	B	¬B	# f.	$\mathcal{RR}$	A	¬A	# f.	$\mathcal{RR}$	A	¬A
	'base–line': recognition rates with NN, with flags, learn ≠ test													
1	all	–	276	88.3	84.8	88.8	276	82.6	78.3	86.6				
	for comparison with old 'base–line': recognition rates with LDA, with flags, learn ≠ test													
	number of cases			1547	203	1344		1547	697	850				
2	stand.	0.001	57	88.5	75.4	90.5	46	78.8	70.9	85.3				
	recognition rates for leave one out													
	number of cases			13274	1999	11275		13274	5140	8134		13274	5140	8134
	analyses with syllable–based features													
3	stand.	**0.01**	**51**	**88.4**	**74.3**	**90.9**	**46**	**80.9**	**76.1**	**83.9**	**51**	**81.1**	**75.6**	**84.6**
4	*struct.*	–	51	87.7	73.3	90.3	46	80.2	75.4	83.2	51	80.6	74.7	84.3
5	stand.	**0.9**	**11**	**86.6**	**68.5**	**89.8**	**6**	**76.4**	**71.8**	**79.4**	**9**	**77.8**	**71.3**	**82.0**
6	*struct.*	–	11	86.0	67.6	89.2	6	77.6	71.6	81.4	9	77.7	69.8	82.7
7	common	–	3	85.7	63.7	89.6	2	75.7	70.9	78.7	2	75.6	67.8	80.5
	analyses without syllable–based features													
8	stand.	**0.01**	**41**	**87.7**	**72.3**	**90.4**	**40**	**80.2**	**75.4**	**83.3**	**37**	**80.6**	**74.3**	**84.5**
9	stand.	**0.9**	**12**	**86.5**	**68.0**	**89.8**	**5**	**76.3**	**72.1**	**79.0**	**8**	**77.7**	**71.2**	**81.9**
10	PC	–	**25**	**85.4**	**72.2**	**87.7**					**25**	**79.6**	**71.2**	**84.9**
11	PC + Flags	–	41	86.4	70.4	89.2	41	79.1	73.0	83.0				

Table 2: Recognition rates for different constellations

Standardized Coeff. tol. = 0.9, with/without flags

parameter	**context**				
	-2	**-1**	**0**	**+1**	**+2**
energy				rising_w	
energy				rising_w	
F_0			min lower_s		
F_0		falling_w			
F_0				– variation_w	
F_0 POS			*on earlier*$_w$		
duration	shorter_w		*longer*$_s$		shorter_s
pause		no/shorter_s	*longer*$_w$		

Table 3: Boundaries: 11 most relevant features, ref.nr. (5) in Table 2

Structure Coeff., with/without flags

parameter	**context**	
	-1	**0**
energy		+ variation_w
energy POS		$\text{max earlier}_{s,w}$
F_0 POS	Off earlier_w	*on*/min earlier_w
duration		*longer*$_{s,s,s,s}$
pause		*longer*$_w$

Table 4: Boundaries: 11 most relevant features, ref.nr. (6) in Table 2

Standardized Coeff. tol. = 0.9, with flags

parameter	**context**			
	-2	**-1**	**0**	**+1**
energy				mean lower_w
energy			+ *variation*$_w$	
F_0	falling_w			
F_0		rising_w		
flags		– *word–final*$_s$		+ word–final_s

Standardized Coeff. tol. = 0.9, without flags

parameter	**context**				
	-2	**-1**	**0**	**+1**	**+2**
energy	mean lower_w				
energy		rising_w			
energy		– variation_w	+ *variation*$_w$	– variation_w	
F_0			*max higher*$_w$		
F_0			rising_w		
duration			longer_s		longer_s

Table 5: Accents: 6/9 most relevant features, ref.nr. (5) in Table 2

Structure Coeff., with flags

parameter	context	
	-1	0
energy		+ *variation*$_w$
F_0		+ variation_w
energy POS		max earlier_w
F_0 POS		On earlier_w
flags	– *word–final*$_s$	– word accent_s

Structure Coeff., without flags

parameter	context
	0
energy	$\text{max higher}_{s,w}$
energy	+ *variation*$_w$
F_0	*max higher*$_w$
F_0	+ variation_w
energy POS	max earlier_w
F_0 POS	On earlier_w
duration	$\text{longer}_{s,w}$

Table 6: Accents: 6/9 most relevant features, ref.nr. (6) in Table 2