Boiling down Prosody for the Classification of Boundaries and Accents in German and English

Anton Batliner, Jan Buckow, Richard Huber, Volker Warnke, Elmar Nöth, Heinrich Niemann

Chair for Pattern Recognition
University of Erlangen-Nuremberg, Germany
batliner@informatik.uni-erlangen.de

Abstract

In the focus of this paper is a comparison of the most relevant prosodic features/feature classes for the classification of boundaries and accents in German and in English. Principal components were computed based on a large prosodic feature vector; these principal components were used as predictor variables in a Linear Discriminant analysis as well as in a Classification and Regression Tree. The number of the most relevant principal components was between three and five; for both languages and for boundary and accent classification alike, most important were principal components modelling duration, in combination with energy, followed by pauses and F0.

1. Introduction

In earlier studies [3, 2], we had a look at those prosodic features that were most relevant for the classification of boundaries and accents in German. The present paper differs from them in several ways: First, we use no longer syllable–based and word-based features but only word–based features because for the computation of the latter ones, phone segment boundaries have not to be computed during recognition which means considerably less computational overhead. This use of word–based features does not result in a loss of performance, cf. [6, 1]. Second, we do not use the prosodic features themselves as predictors, but principal components (PCs) based on these features. By that, we can reduce the number of predictors even more, and these predictors are orthogonal to each other; this makes interpretation easier. Third, there is now an (America) English corpus available which is designed very much alike the German corpus. It is thus possible to analyze the two languages in exactly the same way and to find out, whether they differ with respect to the prosodic marking of boundaries and accents. By that, we certainly cannot solve but at least shed some light on the question whether English is really a 'pitch accent' language, cf. [4, p. 21f]) or whether is is a more pronounced pitch accent language than German.

It is our experience throughout that best classification can be obtained with features that are rather 'raw', i.e., normalized, if necessary, but not explicitely set into relationship with other features, as F0-range (Max minus Min) or the integral of energy. To use all features yields sometimes less classification performance than selecting those features that are at the same time relevant and not too much correlated with other features. This loss is, however, not too severe. If we reduce the number of predictors to a considerable extent, classification performance goes down. However, by that it is possible to boil down a large number of predictors to a small one which can be interpreted more easily. There is thus always a certain trade–off: the clarity of interpretation is negatively correlated with classification performance. In this paper, we are not interested in optimizing classification performance; this has been described in [1, 6]. We confine ourselves to the reduction of the number of predictors and their interpretation, and to the comparison of German with English.

2. Material and Procedure

The research presented in this paper has been conducted under the VERBMOBIL project [9], 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 [8]. 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' ($B = B3$) vs. 'no boundary' ($\neg B = \{B0, B2, B9\}$) and 'accentuated word' ($A = \{PA, SA, EC\}$) vs. 'not accentuated word' ($\neg A = A0$), summing up the respective classes. Note that another clustering that, e.g., assigns the intermediate labels B2 and/or SA to B and $\neg A$, resp., would of course be possible as well.

For the analyses described in the following, we use subsets of the German and English VERBMOBIL database; the data are each divided into a TRAINING and a TEST set (German TRAINING: 30 dialogues, 45 speakers, German TEST: 3 dialogues, 6 speakers; English TRAINING: 33 dialogues, 12 speakers, English TEST: 4 dialogues, 6 speakers). For the TEST sets, classification results obtained with Neural Networks (NNs) are described in [1, 6]. Here, we confine ourselves to *leave–one–out* (loo) analyses using the TRAINING set. By that, we only have seen speakers in our database; this means that results do not diverge to a large extent because some unseen speakers might be modelled badly based on the TRAINING data. Note, however, that results do not differ considerably between TEST and TRAINING; sometimes, they are even better for the unseen TEST sample than for the TRAINING sample. Due to lack of space, these figures will not be given in more detail. Generally,

it turned out that NNs are a bit better at classifying prosodic events than Linear Discriminant Analysis (LDA) [7] or Classification and Regression Trees (CRTs) [5]; NNs are used in the VERBMOBIL system. They are, however, suboptimal if one wants to reduce the number of predictors because of the processing time needed for the training of the NN. For that, the other statistical procedures are better.

2.1. Prosodic Features

For spontaneous speech it is still an open question which prosodic features are relevant for the different classification problems, and how the different features are interrelated. We try therefore to be as exhaustive as possible, and we use a highly redundant feature set leaving it to the statistic classifier to find out the relevant features and the optimal weighting of them. For the computation of the prosodic features, a fixed reference point has to be chosen. We decided in favor of the end of a word because the word is a well–defined unit in word recognition, and because this point can be more easily defined than, for example, the middle of the syllable nucleus in word accent position. Many relevant prosodic features are extracted from different context windows with the size of two words before, that is, contexts -2 and -1, and two words after, i.e. contexts 1 and 2 in Table 1, around the final syllable of a word or a word hypothesis, namely context 0 in Table 1; by that, we use so to speak a 'prosodic five–gram'. A full account of the strategy for the feature selection is beyond the scope of this paper; details are given in [1]. Table 1 shows the 95 prosodic features used and their context. The mean values DurTauLoc, EnTauEnLoc, and F0MeanGlob are computed for the whole utterance; thus they are identical for each word in the utterance, and only context 0 is necessary. Note that these features do not necessarily represent *the* optimal feature set; this could only be obtained by reducing a much larger set to those features which prove to be relevant for the actual task, but in our experience, the effort needed to find the optimal set does normally not pay off in terms of classification performance [3, 2]. The abbreviations can be explained as follows:

duration features 'Dur':

absolute (Abs) and normalized (Norm); the normalization is described in [1]; the value DurTauLoc is used to scale the mean duration values, absolute duration divided by number of syllables AbsSyl represents another sort of normalization;

energy features 'En':

regression coefficient (RegCoeff) with its mean square error (MseReg); mean (Mean), maximum (Max) with its position on the time axis (MaxPos), absolute (Abs) and normalized (Norm) values; the normalization is described in [1]; the value EnTauEnLoc is used to scale the mean energy values, absolute energy divided by number of syllables AbsSyl represents another sort of normalization;

F0 features 'F0':

regression coefficient (RegCoeff) with its mean square error (MseReg); mean (Mean), maximum (Max), minimum (Min), onset (On), and offset (Off) values as well as the position of Max (MaxPos), Min (MinPos), On (OnPos), and Off (OffPos) on the time axis; all F0 features are logarithmized and normalized as to the mean value F0MeanGlob;

length of pauses 'Pause':

the silent pause before (Pause-before) and after (Pause-after), and the filled pause before (PauseFill-before) and after (PauseFill-after).

3. Classification

All statistics was computed with the SPSS package and, for CRT, with the package 'Answer Tree' provided by SPSS as well. The general strategy was the following: All analyses were done strictly parallel for the four constellations German boundaries, German accents, English boundaries, and English accents. The number of cases is given in Table 2. For an 'upper baseline', we computed an LDA with all 95 prosodic features as predictors. By sharpening the tolerance criterion, we could reduce the number of features. In Table 2, results are given in the two lines 'LDA/features' for those analyses that yielded the best classification rates; note, however, that using all features is almost as good. In column '#', the number of the predictors used is given. Classification results are given for the recall of the four classes B, $\neg B$, A, $\neg A$, as well as for all cases taken together (column 'all'), i.e., overall classification rate. Note that by using different weights, recall for one class can be optimized at the cost of the other class. We simply took an a priori probability of 0.5 for the two classes and did not try to optimize, for instance, performance for the marked classes B and A.

In a second step, PC analyses were computed yielding 22 PCs for both languages with an Eigenvalue greater than 1.0 which were again used as predictors in an LDA (line 'LDA/all PCs') and in a CRT (line 'CRT/all PCs'). Again, the standard constellation was chosen for both LDA and CRT; the number of levels for CRT was five. As we did not put any effort in optimization, we cannot tell whether results for CRTs would always be worse than for LDA; this is at least the case for the constellations used. We define as most relevant those PCs, who meet one of the following criteria: first, for LDA, their standardized canonical discriminant function coefficient is > 0.20, or second, for CRT, their improvement is > 0.01. The values chosen are of course arbitrary but meaningful: the number of predictors is reduced considerably, but not too much. Following these criteria, the most relevant PCs for the two languages German (G) and English (E) and for the classification of boundaries and accents are the following; note that (1) these are not necessarily the PCs with low numbers - which have the highest Eigenvalues - and (2) that the numbering in German and English differs slightly - this is not relevant - but that the PCs themselves are very similar. We only display those features, whose factor loading is above 0.5:

German:

G2:DUR/EN/POS: DurAbs0,0, DurAbsSyl0,0, EnEneAbs0,0, EnEneNorm0,0, EnMaxPos0,0 F0MaxPos0,0, F0MinPos0,0, F0OffPos-1,-1, F0OnPos0,0

G6:F0: F0Max0,0, F0Mean0,0, F0Min0,0, F0Off0,0, F0On0,0

10:DUR/(Pause/En): Pause-fill-after1,1, DurAbsSyl1,2, DurNorm1,2, DurAbGs1,2, EnEneNorm1,2

19:PAUSE/POS: Pause-fill-after0,0, F0OnPos1,1

G20:PAUSE: Pause-after1,1

English:

E3:DUR/EN/POS: DurAbs0,0, DurAbsSyl0,0, EnEneAbs0,0, EnEneNorm0,0, EnMaxPos0,0, F0MaxPos0,0, F0MinPos0,0, F0OnPos0,0

E4:F0: F0Max0,0, F0Max-1,-1, F0Max1,1, F0Mean0,0, F0Min0,0, F0Off0,0, F0On0,0, F0Mean-1,-1, F0Off-1,-1, F0Mean1,1, F0On1,1

E5:DUR/EN/POS: DurAbs-1,-1, DurAbsSyl-1,-1, EnEneAbs-1,-1, EnEneNorm-1,-1, EnMaxPos-1,-1, F0MaxPos-1,-1, F0MinPos-1,-1

E12:DUR: DurNorm0,0, DurNorm1,1, DurNorm1,2

E18:PAUSE: Pause-after1,1, Pause-fill-after0,0

E19:POS: F0OnPos1,1

Tables 3 to 6 display these PCs, ordered by relevance (first those with the higher criterion values, then those with the lower ones), and attributed to the contexts -1, 0, or 1, which they model, for the four different constellations.

In a third step, LDAs with these most relevant PCs were computed; results are given in the lines 'LDA/most rel. PCs' of Table 2. By that, we really can estimate the contribution of these few PCs to the classification. We see that recall is of course worse but still fairly good. If we take the complement sets of these most relevant PCs for classification, i.e., the 'less relevant' PCs, recall goes down drastically, between 15 % and 20 %. So we really can say that these very few most relevant PCs model boundaries and accents pretty well.

4. Discussion

The 'center of information' is, of course, the actual word 0, words -1 and +1 are less important, and words -2 and 2 even less because this context is not modelled by the most important PCs; this meets the expectations. As for the features classes, PCs modelling duration and energy at the same time are most important. This holds for English and German, and for boundaries and accents, as well. For boundaries, pauses are more important than for accents, PCs modelling F0 are not irrelevant, but even less important. What about position on the time axis POS for F0, those features that describe the timing of the prominent F0 features? Of course, for a complete intonational modelling, they are necessary as well. It turns out, however, that most of the time, they go together with duration features – as energy POS features do, cf. PCs G2, E3, E5, and sometimes, with pause features, cf. G19, but never with F0 features. Of course, they convey another kind of information than F0 features. They are thus, strictly speaking, duration features as well.

Why are F0 features not more important? This might be traced back either to the mere fact that they simply are not - in spite of the prevalence of intonation models, or to difficulties in the extraction of these features. Automatic feature extraction is never perfect, and this holds for all feature groups. Gross errors, for instance, octave errors for F0 extraction, erroneous voiced/unvoiced decisions, or wrong segmentation and thus computation of duration do occur. We cannot give any figures as for possible differences in the feature extraction, but at least, for prosodic question classification, which heavily relies on a correct computation of final rises and falls, our algorithms work very well for the same databases, cf. [1]. Another possibility might be that speakers use different tonal structures (LH* vs. L*H etc.) indicating different semantics for the marking of accents, and that our algorithms are not able to tell these structures apart: as H* is much more common than L*, the latter ones might simply be discarded by our classifier or, and that is more likely, the classifiers 'learns' that it cannot rely too much on F0 and instead, it relies on the other, more stable feature groups. We do not believe, however, this being the case because a common trait of all accents is the F0 range (Max - Min) which is higher for accentuated than for unaccentuated words. The range can implicitly be modelled by the classifier.

Maybe we can, to conclude the discussion, speculate a bit on these two dimensions: time and frequency. Pitch needs both, because F0 values without any positioning on the time axis do not make any sense. Duration, on the other hand, only needs the time dimension. Moreover, we know that there is some correlation between F0 excursion and duration: for instance, the higher the F0Max, the more time is needed for going up and down. Thus duration features comprise, in a way, information from both dimensions, whereas F0 features only model one dimension. It might be that pitch production 'comes first', and duration is partly triggered automatically. But that could mean, in turn, that duration information is richer and, at the same time, at least at the word level, more robust than pitch information.

5. Concluding remarks

We have seen that (American) English and German use almost the same prosodic information for modelling boundaries and accents, i.e., in order of relevance, duration, energy, pauses, F0. We have found no indication that English relies heavily on F0 information, or that English is to a larger extent a pitch accent language than German. This is of course no final proof but a reasonable working hypothesis if one wants to have a look at other spontaneous German and/or English speech corpora.

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 Grant 01 IV 701 K5 and in the framework of the SmartKom project under Grant 01 IL 905 K7. The responsibility for the contents of this study lies with the authors.

6. References

[1] A. Batliner, A. Buckow, H. Niemann, E. Nöth, and V. Warnke. The Prosody Module. In Wahlster [9], pages 106–121.

[2] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann. Prosodic Feature Evaluation: Brute Force or Well Designed? In *Proc. 14th Int. Congress of Phonetic Sciences*, volume 3, pages 2315–2318, San Francisco, August 1999.

[3] A. Batliner, A. Kießling, R. Kompe, H. Niemann, and E. Nöth. 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*, pages 39–42, Athens, September 1997. University of Athens, Department of Informatics.

[4] D. Bolinger. *Intonation and its Parts: Melody in Spoken English*. Edward Arnold, London, 1985.

[5] L. Breiman. *Classification and Regression Trees*. Wadsworth, Belmont CA, 1984.

[6] J. Buckow, V. Warnke, R. Huber, A. Batliner, E. Nöth, and H. Niemann. Fast and Robust Features for Prosodic Classification. In V. Matoušek, P. Mautner, J. Ocelíková, and P. Sojka, editors, *Proc. Workshop on TEXT, SPEECH and DIALOG (TSD'99)*, volume 1692 of *Lecture Notes for Artificial Intelligence*, pages 193–198, Berlin, September 1999. Springer–Verlag.

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

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

[9] W. Wahlster, editor. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, New York, Berlin, 2000.

features	context size				
	-2	-1	0	1	2
DurTauLoc; EnTauEnLoc; F0MeanGlob			•		
Dur: Norm,Abs,AbsSyl; En: RegCoeff,MseReg,Norm,Abs,Mean,Max,MaxPos; F0: RegCoeff,MseReg,Mean,Max,MaxPos,Min,MinPos		• • •	• • •	• • •	
Pause-before, PauseFill-before; F0: Off,Offpos		•	•		
Pause-after, PauseFill-after; F0: On,Onpos			•	•	
Dur: Norm,Abs,AbsSyl En: RegCoeff,MseReg,Norm,Abs,Mean F0: RegCoeff,MseReg	• • •			• • •	
F0: RegCoeff,MseReg; En: RegCoeff,MseReg; Dur: Norm		•			

Table 1: 95 prosodic features and their context

German: 2310 B, 10964 ¬B, 5140 A, 8134 ¬A								
predictors	#	B	¬B	all	#	A	¬A	all
LDA/features	48	74.3	91.3	88.3	40	68.4	88.1	81.2
LDA/all PCs	22	75.8	84.8	83.5	22	65.0	87.3	78.5
LDA/most rel. PCs	4	73.9	85.5	83.5	2	68.0	82.1	76.6
CRT/all PCs	22	52.4	96.1	83.0	22	65.2	88.7	77.2
English: 638 B, 4137 ¬B, 1958 A, 2817 ¬A								
predictors	#	B	¬B	all	#	A	¬A	all
LDA/features	29	73.8	95.3	92.5	22	75.6	79.4	77.8
LDA/all PCs	22	76.8	92.7	90.6	22	72.5	78.1	75.8
LDA/most rel. PCs	3	76.0	90.8	88.8	3	68.9	79.3	75.0
CRT/all PCs	22	53.2	97.9	84.9	22	68.8	84.8	73.7

Table 2: Recognition rates for different constellations; leave–one–out

0	1	2
2:DUR/EN/POS LDA: .87 CRT: .14	10:DUR/(Pause/En) LDA: -.24 CRT: .01	
	20:PAUSE LDA: .46 CRT: .05	
19:PAUSE/POS LDA: .34 CRT: .04		
6:F0 CRT: .02		

Table 3: Most relevant PCs: German, boundaries

-1	0	1	2
	3:DUR/EN/POS LDA: .82 CRT: .17	19:POS — CRT: .05	
	18:PAUSE LDA: .74 CRT: .11		
4:F0 LDA: -22			
	12:DUR CRT: .02		

Table 4: Most relevant PCs: English, boundaries

0	1
2:DUR/EN/POS LDA: .94 CRT: .14	
6:F0 LDA: .25 CRT: .01	
19:PAUSE/POS CRT: .02	

Table 5: Most relevant PCs: German, accents

-1	0	1
5:DUR/EN/POS LDA: -.22 —	3:DUR/EN/POS LDA: .95 CRT: .13	19:POS LDA: -.20 CRT: .01
	18:PAUSE CRT: .01	

Table 6: Most relevant PCs: English, accents