Duration Features in Prosodic Classification: Why Normalization Comes Second, and what they Really Encode.

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

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

Abstract

For the classification of boundaries and accents in German and English spontaneous speech in the VERBMOBIL project (speech to speech translation system), we use a large prosodic feature vector; duration features represent the most important feature class. They are computed in three different ways: (1) The word duration is normalized with respect to the 'expected' word duration: DURNORM; (2) Duration is normalized as for the number of syllables in the word: DURSYLL; (3) The absolute duration value DURABS of a word is taken. Normally, we use all these feature classes simultaneously. In the present paper, we have a look at the impact of each of these duration classes separately. In addition, we use part-of-speech (POS) information as a further knowledge source. It turns out that throughout, the best feature class, if used alone, is DURABS, followed by DURSYLL, and third comes DURNORM. Best results are achieved by using all feature classes together. With POS information, better results can be achieved than without. This effect is larger for accent classification than for boundary classification, and much larger in combination with DURNORM than in combination with DURSYLL or DURABS. These results indicate that especially DURABS does not only encode prosodic but to a large extent syntactic POS information as well: content words are normally more prone to be accentuated than function words, and at the same time, they tend to be longer. This information is of course lost if duration is normalized, as is the case for DURSYLL and DURNORM.

1. Introduction

In [4], we compare 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 (LDA) [9] 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.

Thus it seems that the 'prototypical' prosodic feature (group) F0 is not that important for the prediction of these two 'classic' prosodic events, i.e., the marking of boundaries and accents. Two questions can be asked: Why is F0 not that important, and why is duration that important? In [4], we dealt with the first question, in the present paper, we will concentrate on the second question. Normally, we use all duration features, together with all other features, simultaneously for classification. In this paper, we want to have a closer look at the different duration classes to find out which one contributes most to classification.

2. Material and Procedure

In the end–to–end German speech understanding system VERBMOBIL [13], which aims at automatic speech–to–speech translation in appointment scheduling dialogues, we normally use a large feature vector comprising 95 word–based features for the classification of prosodic events: 17 duration features, 34 energy features, 36 F0 features, and 8 pause features. Reference point is always the end of a word; a context of +/– two words around the the actual word is computed. The features are raw or normalized to utterance–specific mean values. Energy and F0 values model either prominent points of the contour (Maximum, Minimum, Onset, Offset) or regressions. A full account of the features and their evaluation is beyond the scope of this paper; more information and further references are given in [2].

Table 1 shows the 95 prosodic features used and their context. The mean values DurTauLoc, EnTauLoc, 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 [5, 3]. The abbreviations can be explained as follows:

- **duration features 'Dur':** absolute (Abs) and normalized (Norm); the normalization is described in [2] and below, in section 3.1; 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 [2]; the value EnTauLoc 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).

The experiments described in the present paper have been performed on subsets of the VERBMOBIL 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 [11]. 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' (¬B = $\{B0, B2, B9\}$) and 'accentuated word' (A = $\{PA, SA, EC\}$) vs. 'not accentuated word' (¬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 ¬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 [2, 7]. Here, we confine ourselves to *leave–one–out (loo)* analyses using the TRAINING set. For these two subsets, the number of the prosodic events is for German: 2310 B, 10964 ¬B, 5140 A, 8134 ¬A, and for English: 638 B, 4137 ¬B, 1958 A, 2817 ¬A. 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 unseen TEST and *loo* TRAINING; sometimes, they are even better for TEST than for TRAINING. 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 LDA. 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.

3. Features used in classification

3.1. Duration Features

Duration features are computed in three different ways. Two of them are very straightforward: The absolute duration DURABS is given in the word hypotheses graph WHG, syllable–based normalization DURSYLL is computed by dividing DURABS by the number of syllables. The third normalization DURNORM is based on the variations of the speaking-rate which has different effects on individual phonemes. Plosives are for instance much less affected by changes in speaking-rate than vowels. The variablity of the duration of a phoneme in a syllable depends also on the position of that syllable in the word and the position of the word accent. These considerations have led to the normalization that is described in the following.

3.1.1. Duration Normalization on the Phoneme Level

In order to model local speaking-rate variations we use measures that are based on the work of [15]. First, we are interested in capturing how much faster or slower an utterance was produced compared to the 'average speaker'. For a large training database, we compute for each phoneme its mean duration $\mu_{duration(u)}$ and standard deviation $\sigma_{duration(u)}$. $\mu_{duration(u)}$ constitutes the duration of unit u spoken by the 'average speaker'. The ratio $\frac{duration(u)}{\mu_{duration(u)}}$ measures how much faster or slower u was produced. The average of this ratio over an interval I is our measure $\tau_{duration}$, which is defined in Equation 1. Note that in the Equations 1 and 2, τ is stated more generally: the feature parameter F can be replaced not only by $duration$ but also, e.g., by $energy$.

The value $\tau_{duration}$ is used to scale the mean duration $\mu_{duration(u)}$ and the standard deviation $\sigma_{duration(u)}$ of a speech unit u. The product $\tau_{duration}(I)\mu_{duration(u)}$ can be interpreted as the mean duration of the speech unit u if uttered with speaking-rate $\tau_{duration}(I)$. This interpretation is justified by the experiments in [15]; there it was demonstrated that the mean and the standard deviation of speech-sound categories depend linearly on the speaking-rate.

The difference $duration(u) - \tau_{duration}(I)\mu_{duration(u)}$ is negative if $duration(u)$ is smaller than the scaled mean duration $\tau_{duration}(I)\mu_{duration(u)}$ of the speech unit u. A negative difference indicates faster speech; a positive difference indicates slower speech. This difference can be used to detect strong deviations from the scaled mean duration; the disadvantage of this measure, however, is that the deviation depends on the speech-sound category. If we divide the difference by the scaled standard deviation of the duration $\tau_{duration}(I)\sigma_{duration}(u)$ we get a measure that is normalized w.r.t. speech-sound dependent variation. In Equation 2, $\zeta_F(J, I)$ is defined as the average of that fraction in an interval J (interval I is used as 'reference'). With this approach it is also possible to distinguish between phonemes in accentuated and not accentuated syllables, and between phonemes that are in word initial, word final, word-internal syllables, or one-syllable words. This can be achieved simply by using such units u in the Equations 1 and 2.

$$\tau_F(I) \quad := \quad \frac{1}{\#I}\sum_{u\in I}\frac{F(u)}{\mu_{F(u)}} \qquad (1)$$

$$\zeta_F(J,I) \quad := \quad \frac{1}{\#J}\sum_{u\in J}\frac{F(u)-\tau_F(I)\mu_{F(u)}}{\tau_F(I)\sigma_{F(u)}} \qquad (2)$$

3.1.2. Duration Normalization on the Word Level

The measures $\tau_{duration}(I)$ and $\zeta_{duration}(J, I)$ (computed with phonemes as speech units u), as defined in Equations 1 and 2 can already be used as prosodic features and, in fact, are often used, e.g., in [15], [1], and [8]. These measures have several

features	context size				
	-2	-1	0	1	2
DurTauLoc; EnTauLoc; 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

disadvantages, though. First, during feature extraction the duration of each phoneme has to be determined in order to compute these measures. To compute a phoneme segmentation of the recognized words, however, is time consuming and requires considerable memory resources. The word recognition modules in the VERBMOBIL system cannot provide this segmentation due to architectural constraints. Second, the phoneme segmentation suffers if the audio quality is degraded. Furthermore, pronunciation variants can cause the phoneme segmentation to be incorrect and thus lead to erroneous features.

The normalization according to the Equations 1 and 2 can be used on the word level as well. The word duration statistics $\mu_{duration(w)}$ and $\sigma_{duration(w)}$ for a word w can either be determined directly if enough tokens of this word have been observed in the training data. Otherwise the word duration statistics can be approximated based on the duration statistics of the phonemes that w consists of; this approach is thus time–consuming only during the training. This word based normalization circumvents the disadvantages mentioned above and is, therefore, currently used in the VERBMOBIL system.

The duration statistics was computed for a large sub-set of the VERBMOBIL database: German VM1: 655 speakers (not always disjunct), 13901 turns; German VM2: 108 speakers (disjunct), 7268 turns; English, VM1: 191 speakers (not always disjunct), 4081 turns; English VM2: 48 speakers (disjunct), 9887 turns.

3.2. Part of Speech Features

A Part of Speech (POS) flag is assigned to each word in the lexicon, cf. [6]. For German, 15 different POS classes were annotated in the lexicon and mapped onto six cover classes: AUX (auxiliaries), PAJ (particles, articles, and interjections), VERB (verbs), APN (adjectives and participles, not inflected), API (adjectives and participles, inflected), and NOUN (nouns, proper nouns). For English, the POS classes of the Penn treebank [10] were also mapped onto some higher categories; those which are displayed in Table 6 below are: T: the infinitive particle *to*, P: pronoun, C: conjunction or determiner, M: modal verbs, W: Wh-words, V: verbs, R: prepositions or adverbs or particles, L: cardinal numbers etc., J: adjectives, and N: nouns. (The remaining POS classes occur very seldom.) For the context of +/- two words, this sums up to 6x5, i.e., 30 POS features for German and 14 x 5, i.e., 70 POS features for English.

4. Classification and Discussion

All statistics were computed with the LDA procedure provided by the SPSS package. The analyses were done strictly parallel for the four constellations German boundaries, German accents, English boundaries, and English accents. For an 'upper baseline', we display in Table 2 results for an LDA with all 95 prosodic features as predictors. By sharpening the tolerance criterion, we could reduce the number of features. Results are given in the two lines 'German, best' and English, best' for those analyses that yielded the best classification rates; note, however, that using all features is almost as good, cf. [4]. These results represent a sort of upper baseline. Classification rates are always given for the overall recognition rate RR as well as for the class-wise computed recognition rate CL (mean of the recognition rates for the two classes B and ¬B, and A and ¬A, respectively). For the two languages, results are given for analyses without POS features (-POS) and with POS features (+POS), and for four different analyses using only 'normalized' (DURNORM), only 'syllable normalized' (DURSYLL), only absolute duration values (DURABS), and all three duration feature classes taken together (ALL). It turns out that the best feature class, if used alone, is DURABS, followed by DURSYLL, and third comes DURNORM. Best results are achieved by using all feature classes (ALL) together. This holds for both German and English and boundaries and accents. [*]

At first glance, this result is rather puzzling: the 'most primitive, straightforward' feature group absolute duration is markedly better than those features which use more knowledge, and which presumably mirror pre-final lengthening or longer duration in accent position much better than the raw features. The solution can be found by looking at the results with POS features: with POS information, better results can be achieved than without. This effect is larger for accent classification than for boundary classification, and much larger in combination with DURNORM than in combination with DURSYLL or DURABS. This suggests that DURABS encodes POS information that is not entailed in the other two feature groups. Table 3 summarizes the results of Table 2 by displaying the differences in percent of classification rates between analyses with and without POS information.

To check this assumption, we computed, again for the four constellations German and English, boundaries and accents,

[*] Just for comparison, recognition rates using only the F0 features F0Max, F0Min, and F0Mean, are given as well; it can be seen that classification rates based on these features alone are almost always markedly worse than those based on DURABS.

constellation	features	$CL_{bound.}$	$RR_{bound.}$	$CL_{acc.}$	$RR_{acc.}$
German, -POS	DURNORM	68.0	70.0	56.1	58.5
	DURSYLL	70.9	76.3	65.9	68.3
	DURABS	77.5	81.0	74.6	76.6
	ALL	80.4	82.0	75.7	77.4
German, +POS	DURNORM	72.9	74.3	74.8	76.6
	DURSYLL	76.3	76.5	75.7	77.1
	DURABS	79.3	81.3	77.3	78.8
	ALL	81.9	82.5	77.6	78.9
German, POS	all POS	72.5	74.0	75.0	77.0
	only 0,0	71.7	75.2	75.0	76.9
German	best	82.8	88.3	78.3	81.2
	only F0	67.7	75.4	71.3	81.2
English, -POS	DURNORM	69.4	69.9	56.7	58.2
	DURSYLL	81.5	80.5	70.1	71.3
	DURABS	78.3	81.1	74.2	75.4
	ALL	81.5	83.1	77.2	77.4
English, +POS	DURNORM	75.4	78.9	76.4	77.2
	DURSYLL	81.3	83.8	76.7	77.3
	DURABS	81.4	84.6	78.2	78.6
	ALL	83.9	86.2	78.1	78.5
English, POS	all POS	75.3	78.8	78.1	78.5
	only 0,0	72.2	79.3	73.2	75.3
English	best	84.6	92.3	77.5	77.8
	only F0	63.8	78.0	69.5	69.9

Table 2: Recognition rates: duration without/with POS features; for comparison, all prosodic features and only F0 features (F0Max, F0Min, F0Mean) as well

constellation	features	$CL_{bound.}$	$RR_{bound.}$	$CL_{acc.}$	$RR_{acc.}$
German	DURNORM	4.9	4.3	18.7	18.3
	DURSYLL	5.4	0.2	9.8	8.8
	DURABS	1.8	0.3	2.7	2.2
	ALL	1.5	0.5	1.9	1.5
English	DURNORM	6.0	9.0	19.7	19.0
	DURSYLL	-0.2	3.3	6.6	6.0
	DURABS	3.1	3.5	4.0	3.2
	ALL	2.4	3.1	0.9	1.1

Table 3: Differences in percent: classification rates obtained with POS features minus classification rates obtained without POS features

classifications with the four different duration classes on the one hand, and on the other hand, as fifth analysis, with only POS features, and saved case-wise the predicted group membership. Table 4 shows the correspondence in percent between cases attributed to one of the two classes. Obviously, the classifier with POS features corresponds most with the classifier with DURABS, less with DURSYLL, and least with DURNORM.

To complete this interpretation, we show in Tables 5 and 6 in the last column mean absolute duration values for the six German and the ten English POS classes described above, together with occurrences of the POS classes in percent for boundaries and accents. It can be seen that for German, the function words AUX and PAJ are shorter and most of the time not accentuated, whereas it is the other way round for the content word classes APN, API, and NOUN. Verbs are somewhat in between. As for boundaries, the distribution is marked as well, but mirrors of course the syntactic structure of German: inflected adjectives and participles cannot be found as often before B as in accent position.

The situation is very much alike in English: verbs again are in between, function words are less, and content words are more accentuated. At boundaries, it can be seen that for instance verbs do very seldom occur in pre–boundary position; this is of course due to the English word order which is different from the German one.

5. Concluding remarks

We did not expect two outcomes: first, that the normalized features, esp. DURNORM, are that bad, and second, that DURABS is that good at classifying boundaries and accents. From a theoretical point of view, DURNORM should really be a good measure of duration. The only reason we can think of at the moment is that it is possibly too coarse because of errors in the automatic time alignment. If this turns out to be correct, it could explain why the relatively straightforward normalization for DURSYLL yields better classification rates than DURNORM. Thus it might be that basically, a normalization like the one computed for DURNORM is a good measure but only in theory, because in practice, this does not help very much

duration	POS			
	German		English	
	B	A	B	A
DURNORM	55.1	52.3	54.7	54.8
DURSYLL	62.2	62.4	66.2	69.4
DURABS	71.4	75.8	73.7	74.1
ALL	71.4	75.7	72.7	73.7

Table 4: “class-wise” computed correspondence in percent between cases attributed to one of the two classes

if automatic time alignment cannot be improved by a considerable extent.

The difference between DURSYLL and DURABS can be traced back to the close correlation of overall duration and POS information. Isolated modelling of prosody seems thus not to be adequate. It is not only that other, syntactic means are available, as, e.g., word order, but that prosodic and other means are closely interwoven. Actually, this phenomenon shows up in other studies on the use of prosody in the automatic classification of dialogue acts as well, cf. e.g., [12]. In these studies, duration, esp. overall duration of turns, is shown to be the most relevant feature as well. However, it is not ‘simply’ the prosodic feature duration as such, but the fact that a large percentage of dialogue acts consists of back-channelling, i.e., of very short phrases as *yes*, *uhm* etc. With other words, duration encodes syntactic/semantic complexity, and that means in turn, simply number of words. This can be illustrated nicely by the differences between dialogues recorded in the two phases of the VERBMOBIL project: VM1 and VM2: The main difference between these two phases, as far the the setting is concerned, is, that in VM1, people had to push a button if they wanted to talk. Turn taking and stalling are thus ruled by this technical device. In VM2, there were no longer push–to–talk buttons but the conversation followed the ‘normal’ rules. In [14], it is shown that for absolute duration of turns and number of words alike, the most striking difference is that there are much more very short turns in VM2 than in VM1, i.e., back–channellings like *mhm*, *yes*.

So there seems to be, at least at two linguistic levels, a close correlation between ‘duration’ on the one hand and ‘linguistic complexity’ on the other hand: a first order correlation at the word level: semantically heavy words tend to be more complex, i.e., have a more complex morphological structure resulting in more syllables per word and thus longer words, than pure syntactic function words. At the – second order – dialogue level, ‘pure’ illocutionary utterances as, e.g., back–channelling, tend to be very short, in comparison to utterances that combine illocutionary force with propositional content, e.g., if a user asks for information. This is, of course, a statistical statement: there are polysyllabic function words, and there can be very short questions or very long back–channellings – by the way, this is of course no news but rather well-known facts. Actual duration might, however, not only be a result of these factors: we still believe that well–known phenomena as pre-final lengthening and prominence via lengthening play a role as well, at least in languages as German and English; this is backed up by the fact that our normalized duration measure DURNORM alone is at least not irrelevant (between 58.2% and 70.0% overall recognition rates in Table 2); DURSYLL alone is most of the time better than the F0 features alone, cf. Table 2.

To disentangle the distribution of these different factors and to combine them in a unified approach might be an interesting area for basic research and a promising task for automatic speech processing.

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] Paul C. Bagshaw. *Automatic prosodic analysis for computer aided pronunciation teaching*. PhD thesis, University of Edinburgh, 1994.

[2] A. Batliner, A. Buckow, H. Niemann, E. Nöth, and V. Warnke. The Prosody Module. In Wahlster [13], pages 106–121.

[3] 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.

[4] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann. Boiling down Prosody for the Classification of Boundaries and Accents in German and English. In *Proc. European Conf. on Speech Communication and Technology*, Aalborg, September 2001.

[5] 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.

[6] A. Batliner, M. Nutt, V. Warnke, E. Nöth, J. Buckow, R. Huber, and H. Niemann. Automatic Annotation and Classification of Phrase Accents in Spontaneous Speech. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 519–522, Budapest, Hungary, September 1999.

[7] 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.

POS classes	#	B	¬B	A	¬A	sec.
AUX: auxiliaries	1032	10.6	89.4	24.6	75.4	0.21
PAJ: part., art., and interj.	7498	7.1	92.9	19.7	80.3	0.21
VERB: verb	997	41.1	58.9	55.7	44.3	0.38
APN: adj./part., not infl.	1103	35.7	64.3	74.3	25.7	0.39
NOUN: (proper) nouns	1932	37.5	62.5	78.0	22.3	0.44
API: adj./part., infl.	712	19.5	80.5	74.4	25.6	0.47

Table 5: German: Occurrences of POS classes in percent for boundaries and accents, ordered by mean absolute duration values; column #: frequency of POS class

POS classes	#	B	¬B	A	¬A	sec.
T: “to”	142	1.4	98.6	5.6	94.4	0.8
P: pronoun	549	13.7	86.3	18.4	81.6	0.10
C: conj., determiner	550	.7	99.3	8.4	91.6	0.12
M: modal	150	.0	100.0	12.7	87.3	0.13
W: Wh-words	204	4.9	95.1	50.0	50.0	0.17
V: Verbs	789	3.2	96.8	38.8	61.1	0.18
R: prep., adv., particle	947	9.9	90.1	28.3	71.7	0.23
L: card. numbers, etc.	273	10.6	89.4	76.6	23.4	0.30
J: adjectives	455	32.5	67.5	81.3	18.7	0.40
N: nouns	691	36.2	63.8	74.4	25.6	0.41

Table 6: English: Occurrences of POS classes in percent for boundaries and accents, ordered by mean absolute duration values; column #: frequency of POS class

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

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

[10] P. M. Mitchell, B. Santorini, and M. A. Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, June 1993.

[11] 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.

[12] E. Shriberg, R. Bates, P. Taylor, A. Stolcke, D. Jurafsky, K. Ries, N. Cocarro, R. Martin, M. Meteer, and C. Van Ess-Dykema. Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech? *Language and Speech 41*, pages 439–487, 1998.

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

[14] K. Weilhammer, D. Oppermann, and S. Burger. The influence of scenario constraints on the spontaneity of speech. A comparison of dialogue corpora. In *Proc. Second International Conference On Language Resources And Evaluation*, pages 969–973, Athens, Greece, 2000.

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