Multiple Cues for Phonetic Phrase Boundaries in German Spontaneous Speech

Benno Peters

Institute of Phonetics and Digital Speech Processing (IPDS)
Christian-Albrechts-University, Kiel, Germany
bp@ipds.uni-kiel.de

ABSTRACT

The degree of cohesion of adjacent linguistic/phonetic elements (e.g. dialogue turns, phrases, words, or syllables) is signalled by bundles of phonetic cues. The empirical investigation of German spontaneous speech described in this paper focuses on phonetic phemomena that occur in variable combinations as delimitative features at prosodic phrase boundaries. The term *phrase boundary* is defined as a point of clearly audible, phonetically cued separation between adjacent units. Four different categories of phrase boundaries are distinguished: turn internal boundaries, phrase boundaries at turn-transitions, boundaries linked to syntactic errors, and boundaries with hesitational lengthening in disfluent speech. A classification of the different types of phrase boundaries is presented according to their phonetic characteristics and linguistic function.

INTRODUCTION

The phonetic characteristics of phrase boundaries are highly variable. The degree of cohesion/separation of adjacent linguistic/phonetic elements (e.g. phrases, words, or syllables) is signalled by bundles of phonetic cues. The paper focuses on the following phemomena that cause phonetic separation and which may occur in variable combinations as delimitative cues at prosodic phrase boundaries:

- Strong tonal breaks that in most cases occur at one of the following intonation contours: high rising movements that reach the highest region of a speaker's pitch range, or contours that fall to the bottom of a speaker's range (often in combination with laryngealization.)
- F0-reset, i.e. the restart of downstep of F0-maxima in sequences of pitch accents. This phenomenon is also reviewed by some as a restart of the declination line.
- Segmental lengthening
- Pauses and breathing
- Glottalization and glottal stop

Phonetic experiments on syntactic disambiguation through prosodic phrasing show the importance of prosodic phrase boundaries for a correct interpretation of linguistic utterances [1]. Experiments with isolated, grammatically correct sentences of read speech, however, take only a small part of the above mentioned phenomena into account. In spontaneous speech, one can distinguish three basic factors that cause separation:

- *Content and syntactic structure of an utterance.* Parts of an utterance are phonetically separated in accordance with text structure. These breaks often correspond to syntactic nodes and facilitate the understanding of the content.
- *Speaker-interaction in spontaneous conversation.* Dialogue partners give cues to each other wheather and when it is time for them to speak.
- *Difficulties in planning and execution of an utterance.* Disfluencies occur at any position in the utterance and provide the hearer with information about the cognitive process of speech planning by reflecting difficulties and errors on a phonetic level.

The phonetic cues of audible separation differ depending on the cause of the break. The type of phonetic signalling is thus important for the listener to interpret phonetic breaks with respect to the function in conversation.

In this paper, an empirical investigation of a large corpus of German spontaneous dialogues is presented. The phonetic correlates of manually labelled phrase boundaries are collected automatically and interpreted according to their function in the dialogues. The results reveal a clear correspondence between the combinations of phonetic cues described above and the different functional types of breaks. The aim of this study is to provide a subclassification of phrase boundaries with regard to their signalling strength and their (para-) linguistic function.

ISBN 1-876346-48-5

DATA AND METHOD

The empirical investigation is based on a large, manually annotated corpus of German spontaneous dialogues (*The Kiel Corpus of Spontaneous Speech*), [2], [3] which contains segmental and prosodic labelling. The dialogues were recorded in an Appointment Making Scenario as part of the VERBMOBIL project [4]. The database includes 11 dialogue sessions with 22 speakers (13 male, 9 female); the total recording time amounts to approximately 2.5 hours (25000 words).

In VERBMOBIL, an unusual recording set-up was used: Whenever the speakers whish to speak to their partner, they have to press a button. Pressing the button blocks the other speaker's recording channel. By preventing overlap at turn transitions this set-up leads to a strict temporal delimitation of turns.

The prosodic labelling was carried out with PROLAB, [5] a system which allows to label perceptually and functionally distinctive melodic patterns of German spontaneous speech. PROLAB is based on the *Kiel Intonation Model* (KIM) [6]. In the process of prosodic labelling audible phonetic breaks are symbolized with the label `pg` (phrasing). The labelling also provides information on syntactic truncations and hesitational lengthening. Thus it is possible to extract the following phonetic information on different types of boundaries from the labelling:

- turn-medial vs. turn-final phrase boundaries
- phrase boundaries with hesitational lengthening
- phrase boundaries that occur in combination with syntactic irregularities (false starts or truncations)

On the basis of this threefold classification (position in the utterance, (dis-)fluency, and syntactic structure) the phonetic cueing of different phrase boundaries is investigated. The delimitative cues of more than 4000 phrase boundaries in the corpus are automatically extracted from the label files and compared for the different types of phrase boundaries.

Tonal breaks, f0-reset, pauses and breathing are directly marked in the segmental/prosodic labelling, whereas final lengthening has to be estimated by a comparison of phrase-internal and phrase-final segment duration. As phrase-final lengthening usually appears in the phrase-final rhyme [7], a value for each rhyme in phrase-final position is calculated that specifies the degree of lengthening at each phrase boundary with the following formula, used by Price et al. [8] and others :

$$\tilde{d}(i) = (d(i) - \mu_\alpha)/\sigma_\alpha$$

where $d(i)$ is the duration of segment i in the phrase-final rhyme, μ_α and σ_α are the mean and standard deviation, respectively, of the duration of the sound category/phoneme α in phrase-internal position.

The value $\tilde{d}$ expresses the relative duration of a segment as the number of standard deviations above or below the average segment duration. If the segment duration d equals the mean value μ_α, the normalized duration is $\tilde{d} = 0$. If the segment duration is larger (smaller) than the mean, the normalized duration is positive (negative). For segment durations within the standard deviation the values range from -1 to +1. To obtain a representative value for the normalized duration of the whole rhyme in each case, the normalized durations of the rhyme segments are averaged. These numerical indicees are divided into three classes:

$\mathbf{N0} < -1$ → shortening
$-1 \leq \mathbf{N1} \leq 1$ → neither shortening nor lenghtening
$\mathbf{N2} > 1$ → lenghtening

When all cues are extracted for each labelled phrase boundary, a new label symbolizing all phonetic cues is attached to the label file at the position of the phrase boundary.

Finally the extracted bundles were evaluated for four separate samples of labelled phrase boundaries of different types: regular turninternal boundaries (n = 2470), regular turnfinal boundaries (n = 1092), boundaries occurring with hesitational lengthening (n = 871), boundaries linked to syntactic irregularities (n = 150).

In addition to the automatic extraction of phonetic cues a fine auditory analysis is done on a large number of phrase boundaries of different type, leading to an interpretation concerning the functions of a break in the dialogue. This interpretation is necessary to establish category boundaries as well as to find prototypical examples for the different types of breaks.

RESULTS

Turn-internal phrase boundaries without disfluency

Turn-internal phrase boundaries without hesitational lengthening constitute the biggest sample (n=2470). In the literature, this type of boundary is mostly referred to as prosodic phrase boundary.

Segmental lengthening is the most frequent feature, occurring in 66.2% of all cases. Table 1 shows the distribution of averaged segment durations in phrase-final rhymes.

As shown in table 2, automatic extraction found at least one of the cues in 93.1% of all turn-internal boundaries. A wide dispersion over the 16 possible categories leads to frequencies of occurrence around 10% in 8 categories. 55.1% contain more than one phonetic cue. Pauses and/or breathing appear in 38.3% of turn-internal phrase boundaries, in 84% of these cases together with final lengthening. Terminal falls to the bottom of a speaker's range occur in 36% of turn-internal

ISBN 1-876346-48-5

Table 1: Distribution of averaged, normalized segment durations over the classes N0, N1, N2 at turn-internal phrase boundaries (absolute and relative frequencies of occurrence)

N0	N1	N2	ges.
59	777	1634	2470
2.4	31.5	66.2	100

Table 2: 16 bundles of phonetic cues at turn-internal phrase boundaries (R Reset, L final lengthening, P pause/breathing, T tonal break, absolute and relative frequencies of occurrence)

	-	L	P	T	LP	PK	LT	LPT	ges.
+R	255	286	44	202	235	60	210	265	1557
%	10.3	11.6	1.8	8.2	9.5	2.4	8.5	10.7	63.0
-R	171	232	31	57	208	16	112	86	913
%	6.9	9.4	1.3	2.3	8.4	0.6	4.5	3.5	36.9

phrase boundaries, high rising contours in only 1.8%. Of the terminal falls, 29% end with creaky voice.

An auditory analysis of many examples from this class of boundaries has shown that most phrase-internal boundaries cooccur with textual and/or syntactic boundaries. In many cases the strength of phonetically cued separation corresponds to the degree of textual and/or syntactic separation.

Turn-final phrase boundaries without disfluency

Phrase boundaries before turn-transitions cannot be classified with regard to a following f0-reset because the turn ends at this position. Nor can pauses and breathing be taken into account here, as one effect of the recording set-up with control buttons is a disturbance of the natural temporal coordination between turn-holder and turn-taker. No turn-transitions are possible unless the turn-holder yields the turn to the dialogue partner by releasing the button. Neither competitive incomings nor overlapping speech is possible. With respect to the turn-holder's behavior at turn-transitions, the set-up thus facilitates the following interpretation: Since the incomer has no possibility to force turn-transitions, every turn-transition that takes place is intended by the speaker.

The results show a distribution of phrase-final segment durations comparable to that in turninternal phrase-final rhymes (74.3% vs. 66.2% N2) see Table 3, but great differences in phrase-final melodic contours. Terminal falls to the bottom of a speaker's range occur at 63.8% of turninternal phrase boundaries, high rising contours at 10.3%, which means that three quarters of all turn-transitions are cued by extreme pitch movements.

Table 3: Distribution of averaged, normalized segment durations over the classes N0, N1, N2 at turn-final phrase boundaries (absolute and relative frequencies of occurrence)

N0	N1	N2	ges.
12	268	812	1092
1.1	24.5	74.3	100

A low falling intonation is in many languages often terminated by creaky voice [9]. Termination of this sort is sometimes used for a regulative function, when speakers use creak as a signal of yielding the floor to another speaker. In the data, 57% of low falling contours at turn-transitions end in creaky voice, whereas the use of creak is speaker specific. Among the 22 speakers, 4 have relative frequencies below 10% (one 0%), 7 between 10 and 30%, 9 between 40 and 60%, and 2 have 70 and 72% respectively. The group data, combined with these individual distributions, suggest that speakers use laryngealization predominantly at the end of turns, and have diverging preferences for the use of this additional phonetic marker of terminal phrase finality.

Phrase boundaries in disfluent speech

Hesitational lengthening is labelled manually in the process of segmentation. For this phenomenon, no quantitative analysis of segment duration can be given in this paper. The method that was used to determine phrase-final lengthening in fluent speech is not appropriate here, because the region in which hesitational lengthening manifests itself varies a great deal. Not only does it occur at single segments in every position in a word, it may also spread over words or groups of words. As auditory and phonetic analyses of many examples from the corpus show, the lengthened segments often reach much extremer durations than in phrase-final lengthening without disfluency.

At phrase boundaries with hesitational lengthening the relative frequency of f0-reset after the phrase boundary is 68.9%, pauses and/or breathing occur in 46.2% of all cases. In 17.8% of the boundaries low falling pitch movements appear at the end of the phrase, whereas high rising contours are present in only 2.3%. Most cases show a flat f0-contour. There are two typical melodic patterns overlaying lenghtened stretches: A completely flat contour which often occurs when the phrase is built up of one or two words only, like *and then...*, and is throughout lengthened, or a pattern having a pitch accent with a rising contour that levels out at the top. The temporal structure of hesitational lenghtening might also be a prominent cue for a correct interpretation by the hearer; this has to be quantified in further investigations.

ISBN 1-876346-48-5

Table 4: Melodic contours at the four bundary types (relative frequencies of occurrence)

	Internal	Final	Hesitation	Error
Low falling	36.0	63.8	17.8	15.3
High rising	1.8	10.3	2.3	0.0
Other	62.2	25.9	79.9	84.7

At interruption points alongside syntactic errors (false-starts and truncations) a strong pitch movement occurs in only 12% of all cases. Pausing and/or breathing is present in 35.4%, hesitational lengthening in 4.6% of the interruption points. The cueing of syntactic errors has glottalization phenomena (glottal stop and/or glottalization) as frequent phonetic exponents. An earlier analysis of the same database showed that 27% of the interruption points were marked by glottal stop and/or glottalization [10].

CONCLUSION

The results show a correlation between certain combinations of delimitative phonetic cues and the different types of phrase boundaries. Turn-final phrase boundaries have the strongest signals of separation: Final lengthening appears in more than 74% of all cases; large-range pitch contours and phonatory changes are frequent. Turn-medial phrase boundaries are often cued by a reset in pitch (63%), but also by lengthening of the final rhyme (66.2%). Phrase boundaries in combination with syntactic irregularities often exhibit glottalization phenomena and glottal stops. Strong pitch movements are rare in both types of disfluent boundaries. Table 4 compares the distribution of low falling and high rising pitch movements in the four boundary types. The dispersion mirrors the relationship between phrase-final intonation contours and their pragmatic function in dialogue. The auditory analysis shows, that especially high rising contours will in most cases be assigned to turn-transitions, as cue for questions. Terminal falls usually mark the end of a major textual unit and are more often found at turn-transitions than in any other position.

OUTLOOK

A new volume of *Kiel Corpus* has just been completed. The speech data are recorded within the *Videotask-Scenario* [11] and the recording set-up allows overlapping talk. With these data it is now possible to analyze the temporal coordination at turn-transitions. The findings of corpus analysis and auditory interpretation should be backed up by perception experiments that investigate the degree of perceptual separation for different bundles of delimitative phonetic features. Perceptual thresholds for the complex interaction of phonetic parameters are necessary to formalize the relationship between phonetic form and (para-)linguistic function of the different phrase boundary types. Finally, the resulting model of prosodic phrasing should be integrated in an over-all model of intonation.

ACKNOWLEDGEMENTS

Part of the work reported here was funded by German Research Council Grant Ko 331/22-2 ("Sound patterns of German spontaneous speech") under the supervision of Klaus Kohler. I would also like to thank Michel Scheffers for writing programmes that helped me with the automatic processing of the corpus data, and Klaus Kohler and Thomas Wesener for many helpful discussions and suggestions.

REFERENCES

[1] L. A. Streeter, "Acoustic determinants of phrase boundary perception," *Journal of the Acoustical Society of America*, vol. 64, pp. 1582–1592, 1978.

[2] IPDS, *The Kiel Corpus of Spontaneous Speech*, vol. 1, CD-ROM#2, IPDS, Kiel, 1995.

[3] IPDS, *The Kiel Corpus of Spontaneous Speech*, vol. 2, CD-ROM#3, IPDS, Kiel, 1996.

[4] R. Karger and W. Wahlster, *VERBMOBIL Handbuch*, Verbmobil Technical Report Nr. 17. DFKI, Saarbrücken, 1994.

[5] K. J. Kohler, "Modelling prosody in spontaneous speech," in *Computing Prosody*, Y. Sagisaka, N. Campbell, and N. Higuchi, Eds., pp. 187–210. Springer, 1997.

[6] K. J. Kohler, "A model of German intonation," in *AIPUK*, vol. 25, pp. 295–368. IPDS, 1991.

[7] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *Journal of the Acoustical Society of America*, vol. 91, pp. 1707–1717, 1992.

[8] P. J. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, "The use of prosody in syntactic disambiguation," *Journal of the Acoustical Society of America*, vol. 90, pp. 2956–2970, 1991.

[9] J. Laver, *Principles of Phonetics*, Cambridge Textbooks in Linguistics. Cambridge, 1994.

[10] K. J. Kohler, B. Peters, and T. Wesener, "Interruption glottalization in german spontaneous speech," in *Disfluency in Spontaneous Speech, DISS 01*, pp. 45–48. Edinburgh, 2001.

[11] B. Peters, "'Video Task' oder 'Daily Soap Szenario': Ein neues Verfahren zur kontrollierten Elizitation von Spontansprache," URL: http://www.ipds.uni-kiel.de/pub_exx/bp2001_1/Linda21.html, 2000.