Prosodic and Gestural Features of Phrase-internal Disfluencies in Polish Spontaneous Utterances

Maciej Karpiński and Ewa Jarmołowicz-Nowikow

Center for Speech and Language Processing and Institute of Linguistics
Adam Mickiewicz University, Poland
maciej.karpinski@amu.edu.pl, ewa.jarmolowicz@amu.edu.pl

Abstract

The aim of this pilot study is a preliminary description of intra-phrasal disfluencies in Polish task-oriented dialogues. Some cues to their identification and discrimination are tentatively proposed. The data come from eight paper folding task dialogue sessions. Disfluencies are categorized on the basis of their acoustic-phonetic content. Selected prosodic properties of their neighborhood are analyzed, including intonational realizations and durations of the surrounding syllables. Accompanying hand movements of the speakers are described for the alignment of their phases with respective disfluencies in speech.

Index Terms: disfluency, prosody, gesture, dialogue, phrase

1. Background and motivation

Disfluencies in speech may result from a variety of sources. Among them, problems arising at various stages of utterance production and the cognitive processing load are frequently mentioned [11, 10, 23]. Most disfluencies are produced unconsciously and they are frequently, but not always, overlooked or ignored by both the speaker and the hearer in everyday communication [19, 20, 21]. Human speech perception system seems to be equipped with disfluency compensation mechanism [4]. Disfluencies may be also produced consciously as an feature of a peculiar speaking style or for some pragmatic or rhetorical purposes [18]. Clark and Fox Tree [6] suggest that English *uh* and *um are* purposefully produced by speakers discovering delays in speech production. It may be also hypothesized that unconsciously evolving disfluencies may come under control of the speaker as he or she may adjust them in real time. Disfluencies may prove listener-friendly when marking units of speech and when functioning as discourse markers or focusing listeners' attention (e.g., [25]). However, when they occur within the realizations of utterance units (bound to respective intonation units, cf. [9]), they may evoke processing and understanding problems, both for human listeners and automatic speech recognition systems. It has been shown that conversational disfluencies are shaped by a number of factors, including the age and gender of the speakers as well as the topic of the conversation [3].

Phrase-internal disfluencies seem to be worth special attention. Some of them deeply disrupt the phrase structure and lead to incomplete, unfinished units which cannot be regarded as "intonational units" anymore as they do not meet basic criteria for well-formedness. However, in a number of cases, speakers "recover" from disfluencies and maintain the continuity of their utterances. If a speech production problem occurs on any level (e.g., lexicon or syntax), other levels may be engaged in the continuity management.

Some researchers argue [7] that the cues to the coherence of the intonational unit (namely, intonational phrase, IP) come not only from the prosodic level. As long as the IP is regarded as a linguistic (not merely a phonetic) unit, its coherence may in fact be dependent on a number of factors (including semantic, syntactic and even pragmatic ones). Accordingly, there are reasons to expect that disfluencies occurring on one level of an utterance (e.g., lexical) may be somehow compensated or alleviated on some others (e.g., prosodic). On the basis of a number of the studies of gesture-speech relationship [8, 5], one may expect that gestures can also play a role in the mechanism of phrase continuity management.

The number of phonetic studies confessed to speech disfluencies and phrase continuity in Polish is very limited. The aim of the present pilot study is to analyze basic prosodic and gestural properties of the disfluency and its closest surroundings, and to determine directions for more detailed analyses within the current project on multimodal communication.

2. Recording and labelling

The data under study come from DiaGest2 project pilot recordings and consist of eight video- and audio-recorded paper folding task dialogue sessions (cf. [12]). The participants are young educated female and male native speakers of Polish. The total length of the sessions is ca. 40 minutes (5 minutes per session). Only the utterances produced by instruction givers (four females and four males) were analyzed in this study. The recordings were segmented into syllables, transcribed in phonemic transcription (a SAMPA-based system), and divided into intonational units and gestural units.

In the DiaGest2 project, two levels of intonational phrasing were applied: Minor Intonational Phrases (MinIPs) and Major Intonational Phrases (MajIP) [27]. Here, analyses are limited to the MajIP level for MajIPs normally correspond to meaningful dialogue units (like dialogue acts). The following categories of units were distinguished on the MajIP level:

- *IP* – a well-formed intonational phrase;
- *Disfluent IP* (*DisIP*) – an IP containing some disfluencies but still forming a unit, closed with a phrase final melody (boundary tone) and, in most cases, corresponding to a single dialogue act;
- *Unfinished IP* (*UnfIP*) – a stretch of speech which starts as a regular IP but which is discontinued (e.g., abandoned) and never intonationally (usually also syntactically and semantically) closed.

A limited number of units was tagged as both disfluent and unfinished. Some units suffering from major sound quality problems (overlapping speech, distorted sound, etc.) were marked as "sound quality problem units" (SQP).

The syllables surrounding pauses were labeled for their pitch and duration. The durations of pauses (and their components) as well as the pitch of filled pauses was also measured (whenever it was possible). The mean value of pitch frequency was calculated for each syllable. The initial and the final values of f_0 were measured for the surrounding syllables. Pitch movement within syllables and filled pauses was descriptively tagged on the basis of perception, using the following categories: rise, fall, flat (medium f_0 level for a given speaker and stretch of speech), high (high f_0 value), low (low f_0 value). The last three labels imply no perceptually significant pitch movement.

The recordings were labeled for their gesture phrase content limited to hand gestures as well as hand movements not related to the gesture phrase. The model of gesture phrase proposed by Kendon [15] was employed. According to it, the gesture phrase consists of the following phases: (a) stroke (obligatory phase where the meaning of the gesture is expressed), (b) preparation (movement of the hand that leads to the beginning of the stroke) and (c) retraction (the hand returns to a starting point or the rest position). Pre-stroke holds and post-stroke holds which occurred optionally, respectively, before or after the stroke phase were also distinguished in the gesture phrase structure [16]. A raised but motionless hand that may be observed between gesture phrases and is not a part of their structures, here was also taken into consideration. As the starting point for our analysis were the disfluencies in speech we did not make a separate analysis for the disfluencies in gesture phrases. Hand movements occurring during speech disfluencies were analyzed. To this end, four categories were distinguished which were determined by the gesture phases. The raised but motionless hand observed between gestures was regarded as belonging to a separate, fifth category. Category VI contains movement units where hands are in the rest position. Respective categories are defined as follows:

- Category I: corresponds to the pre-stroke hold or post-stroke hold phase;
- Category II: corresponds to the stroke phase;
- Category III: corresponds to the preparation phase;
- Category IV: corresponds to the retraction phase;
- Category V: hand is raised (placed e.g. at waist, chest, arm height) and motionless; this category does not belong to the gesture phrase structure;
- Category VI: rest position of the hand.

Due to numerous technical and methodological problems related to the analysis of disrupted intonational phrases, all the measurements and labeling had to be done manually. The same applied to the hand movement component of utterances. Therefore, the number of analyzed phrases is limited.

3. Disfluent phrases in the paper folding task dialogues

3.1. Overview

The material under study is transcribed phonemically and segmented into working units [13]. In Table 1, the proportions of most frequent types of units are listed.

The total proportion of disrupted phrases reaches almost 23%. However, only 10.6% of them are categorized as DisIP. For further analyses, this subset is further narrowed down by the elimination the units containing extremely long and complex disfluencies.

Table 1. *The proportions of well-formed, disfluent , unfinished IPs and other ill-formed units in the analyzed material*

Unit category	Proportion
IPs	72.6%
Disfluent IPs (DisIP)	10.6%
Unfinished IPs (UnfIP)	11.0%
Other ill-formed units	1.0%
Sound Quality Problem	4.8%
Total	100% (1004 units)

3.2. Prosodic component

For the analyses presented below, two types of DisIPs were selected: (a) DisIPs with a single silent (SP) or filled (FP) pause or a single sequence of such pauses (in which they were adjacent); (b) DisIP containing a few such sequences but with a clearly distinguishable, separable first sequence of pauses. For the latter category, only the first sequence was analyzed. Sequences of silent and filled pauses are, henceforth, referred to as disfluency sequences (DS). Nevertheless, it should be kept in mind that the range of actual disfluency is obviously wider and includes, e.g., the pre-pausal syllable. While most disfluencies were realized as silent (SP) or filled pauses (FP), or their combinations, a number of repetitions and repairs were also found. However, they were not included in further analysis. The types of DS are listed, with respective proportions, in Table 2. The subset under analysis consisted of 85 DSs.

Table 2. *The proportions of disfluencies realized as filled pauses, silent pauses, or their combinations*

Content of the DS	Proportion
SP	49.4%
FP	22.5%
SP FP	8.2%
FP SP	8.2%
SP FP SP	8.2%
Others	3.5%
Sum	100% (85 DisIPs)

For all the tokens involving SPs and FPs, the following parameters were measured using Praat [2]: (a) durations of all the syllables in the phrase, (b) mean f_0 for the syllable directly preceding and following each DS, (c) the initial and final value of f_0 for all the syllables directly preceding and following the DS, as well as for the DS itself (where it comprised a voiced FP). As the f_0 values are often unstable in the phrase boundary regions, the measuremet was based on stylized ("smoothed") contours and extreme final values were rejected. On the basis of these measurements, further parameters were obtained. Additionally, pitch changes in the syllables surrounding every disfluency were labeled on the basis of auditory inspection.

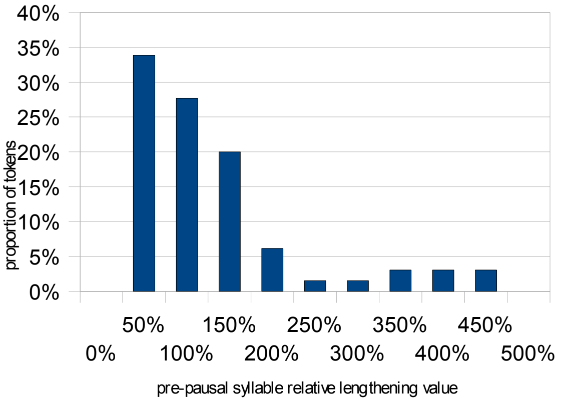

Figure 1. *The distribution of the relative pre-pausal syllable lengthening values*

The relative lengthening of the syllables directly preceding DSs was measured against the mean duration calculated for all earlier syllables in a given DisIP. This did not applied to situations where a disfluency started directly after the first syllable of DisIP (6 tokens). In six cases, the syllable preceding the DS was shortened. The lengthening of the phrase-final syllables in the utterances under study was measured wherever possible (its duration was compared to the mean duration of four preceding syllables) with the average result of 5.35%. Although the patterns of final lengthening are complex and may influence a number of syllables (e.g., from the nuclus on), this simple measurement was only intended to show that the lengthenings before hesitation pauses are very different from average effects of phrase-final lengthening on the last syllable of the IP [1, 17, 28]. The distribution of the final lengthening values expressed as percentages is presented in Figure 1. The shortenings mentioned above are not included.

It was found that the pitch contour of the syllables directly preceding DSs was, in most cases, flat, independently of the size and type of DS. For the first post-DS syllable, a similar tendency occurred. Falling contours were also frequent in pre-DS and post-DS syllables. As found in earlier studies [13], filled pauses were predominantly realized with a flat melody. In Table 3, the numbers of occurrences of pitch contour categories are presented.

Table 3. *Pitch contour labels on the syllables directly preceding and following DSs*

pitch contour	pre-DS	during DS (if applicable)	post-DS
high	2	0	1
rise	16	1	13
flat	44	31	47
fall	20	2	22
low	1	0	0

Using the measurements of the initial and final f_0 values for the pre- and post-DS syllables, the relative pitch frequency range was calculated for each DS. The distribution of pitch frequency ranges is shown in Fig. 2. The same relative range coefficient was calculated for a sample of fifty UnfIPs using the f_0 value in the endings of their last syllables and the initial f_0 values in the subsequent working units. The comparison between the means for the phrase-internal measurements (DisIP) and between-phrasal measurements (UnfIP) show no significant difference. Accordingly, the pitch frequency change does not seem to help to distinguish between the intra-phrasal disfluencies and the disfluencies that result in an unfinished phrase.

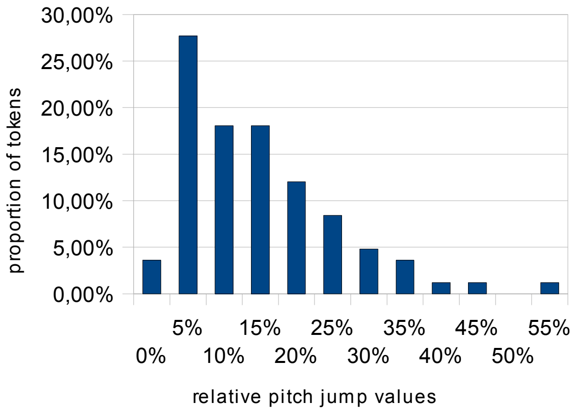

Figure 2. *The distribution relative pitch range values*

3.3. Gestural component

Only 15 DSs are realized without any significant movement or hold executed by hand, head or body posture. The most frequently occurring category accompanying studied DSs is category V (33.6%) which does not belong to the gesture phrase structure and thereby is not recognized as gestural component. Category V stands for the raised hand which takes place between gesture phrases and remains over a span of the whole IP or most of it. This category is often related to some speech production problems (hesitation, difficulty in formulating precise instructions). From the pragmatic point of view this hand position may signal speaker's intent to continue his or her utterance. In Table 4, the proportions of movement units belonging to respective categories are presented.

Most of the DSs are contained in a single movement unit (75.5%), however some of them overlap with multiple movement units (24.5%) consisting of two, three or even four subsequent units (e.g. Category IV, I, and II). Therefore, for the total number of 85 DSs, there are 98 overlapping movement units taken into consideration in this study.

The recorded material was also examined from a perspective of the temporal alignment between DSs and movement units belonging to the gesture phrase (Categories I – IV).

Table 4. *The proportions of particular categories of movement units in the analyzed material*

Category of Movement Unit	Proportion
Category I	13.3%
Category II	18.4%
Category III	4.1%
Category IV	15.3%
Category V	33.6%
Category VI	15.3%
Sum	100% (98 movement units)

The data show no strong tendency concerning time relations between DSs and gesture phase. However it is possible to point out that DSs more often co-occur with these gesture phases that contain movement. The DSs overlap most often with the movement units belonging to Category II (18.4%). It was noticed that 83% of pauses correlated with movement units belonging to Category II lasted less than 300ms. Since the durations of the DSs were relatively short, it may be hypothesized that co-occurring movements were not intentionally used to compensate for the speech disfluence. The results of the examinations conducted by other researches indicate the strong synchronization between speech pauses and gesture holds [8]. No evident co-occurrence of speech disfluencies and gesture holds was confirmed in presented material. 13.3% of DSs overlapped with the pre-stroke hold or post-stroke hold. In 85% of analyzed cases of DSs and movement units of Category I co-occurences, the beginning of hold precedes the beginning of DS or both of them begin exactly at the same time.

4. Discussion

The lengthening of the pre-DS syllable turned out to be one of the most prominent among the prosodic features of the surroundings of DSs. However, this phenomenon is not unique to the analyzed category of disfluent phrases and, as mentioned, lengthenings are themselves parts of disfluencies. Pre-pausal lengthening is considered to be a common phenomenon [23] and it is mentioned also for Polish [17]. Syllable lengthening is often considered as typical of phrase-final syllables and described as one of phrase boundary indicators [26, 28]. On the other hand, pre-DS syllables in the material under study typically have a flat intonational contour as opposite to the variety of contours normally met in phrase-final syllables in Polish (e.g., [13]). Accordingly, these two features (lengthening and flat pitch contour) considered together may be used to predict following pause sequences (DS). Nevertheless, they still do not seem to provide any cues to distinguish between the attempts of phrase recovery (DisIP) and the case of phrase “abandonment” (UnfIP). The flat pitch contour of the post-DS syllable may be considered as a cue for the identification of a post-pausal continuation of a disrupted phrase.

No significant difference between the values of pitch change for (a) pre- to post-DSs and (a) final UnfIP to next IP token. This can be partially explained by the fact that unfinished units were frequently closed with intonationally flat, lengthened syllables (similar to the pre-DS ones in DisIPs) at an average pitch level for a given speaker. As new phrases tended to start with non-extreme pitch values, the pitch change was relatively small.

No correlation was found among the amount of syllable lengthening, the durations of DSs and the value of the pitch frequency jump. While it does not prove that there are no interdependencies among these phenomena, there may be some other strong factors that overdrive potential correlations.

Some researchers stress the temporal alignment between the disfluency in speech and certain phases of the gestural phrase. McNeil [22] points to the stutterers who tend to hold gesturing during the phase of stuttering in speech. In artificially evoked stuttering, disfluencies occurring in speech also tend to bring about holds in gesturing [*ibidem*]. Esposito and Marinaro [8] strongly underlines the co-occurrence of overlaps of gestural holds and pauses in speech, both in children and adults. The present study does not confirm this claim. It was found that strokes accompany disfluencies even more frequently than by intra-phrasal holds. However, we are not formulating any strong hypothesis at this point because of the limited set of analyzed data as well as because of the fact that only a specific category of disfluencies was taken into consideration in the present work.

5. Conclusions

This pilot study is focused on a narrow category of disfluent but successfully closed intonational phrases. The characteristics of the analyzed disfluencies is certainly under the influence of the specific dialogue task (Shriberg's [2001] "Ecological Claim"). Therefore, our findings may not apply to the properties of speech disfluencies and accompanying gestures in general. Nevertheless, even with the small number of analyzed tokens, some indications have been found for directions of further research.

A substantial lengthening and a predominantly flat melody turned out to be the most noticeable features of pre-pausal syllables. The relative lengthening exceeded even four hundred percent. However, these phenomena are not specific only for the studied category of disfluencies and their surroundings. A flat pitch contour of the post-pausal syllable may be hypothesized as indicating the continuation of a disrupted phrase. The pitch frequency change between the final pre-DS and initial post-DS value extremely varied but it was not signitifantly different from the values measured for the final parts of unfinished phrases and the phrases that followed them.

Strict co-occurrence of gestural holds and disfluencies in speech has been not confirmed. In most cases, it was noted that when a speech disfluency is associated with a hold, it is the hold that starts earlier.

As the present study indicates, further research should be based on a significantly larger corpus that would offer a larger number of instances of disfluencies. If the intention of closing an IP is somehow signalized by the speaker, more sources of such cues should be taken into consideration. They may include more subtle body movements, facial expressions and gaze direction. All these factors may be analyzed in the new high-quality recordings from DiaGest2 project.

6. Acknowledgements

This research is a part of DiaGest2 project (N N104 010337) supported by the Polish Ministry of Research and Higher Education.

7. References

[1] Beckman, M. E., Edwards, J. 1990. Lengthening and shortening and the nature of prosodic constituency [In:] J. Kingston and M. E. Beckman (Eds.) *Laboratory Phonology I*, Cambridge: CUP, pp.152-178.

[2] Boersma, Paul 2001. Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341-345.

[3] Bortfeld, H., Silva, D., Bloom, J. E., Schober, M. F., Brennan, S. E. 2000. Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech* 2000, 2001, 434 (23), pp. 12239 –124579.

[4] Brennan, S. 2001. How Listeners Compensate for Disfluencies in Spontaneous Speech. *Journal of Memory and Language*, 44, pp. 274–296.

[5] Christenfeld, N., Schachter, S., Bilous, F. 1991. Filled Pauses and Gestures: It's Not Coincidence. *Journal of Psycholinguistic Research*, Vol. 20, No. L.

[6] Clark, H. H., Fox Tree, J. E. 2002. Using uh and um in spontaneous speaking. Cognition 84, pp. 73–111.

[7] Demenko, G., Jassem, W. 1997. Phonetic and syntactic coherence of the phrase. [In:] W. Jassem, Cz. Basztura (Eds.) *Spech and Language Technology*, vol. I, Wrocław: Format.

[8] Esposito, A., Marinaro, M. 2007. What Pauses Tell Us About Speech and Gesture Partnership. [In:] A. Esposito et al. (Eds.) F*undamentals of Verbal and Nonverbal Communication and the Biometric Issue*. Amsterdam - Berlin - Oxford - Tokyo - Washington: IOS Press, pp.45-58.

[9] Fox, A. 1999. *Prosodic Features and Prosodic Structures: The Phonology of Supras egmentals*. Oxford: OUP.

[10] Fromkin, V.A. (Ed.) 1973. *Speech errors as linguistic evidence*. The Hague, Paris: Mouton.

[11] Goldman-Eisler, F. 1968. *Psycholinguistics: Experiments in Spontaneous Speech*. New York: Academic Press.

[12] Jarmołowicz-Nowikow, E., Karpiński, M., Malisz, Z., Szczyszek, M. 2007. Gesture, Prosody and Lexicon in Task-oriented Dialogues: Multimedia Corpus Recording and Labelling. [In:] A. Esposito, M. Faundez-Zanuy, E. Keller, M. Marinaro (Eds.) *Verbal and Nonverbal Communication Behaviours*. LNAI 4775. Berlin-Heidelberg: Springer-Verlag, pp. 99-110.

[13] Karpiński, M. 2007. Selected quasi-lexical and non-lexical units in Polish map task dialogues, *Archives of Acoustics*, 32, 1, pp. 51-65.

[14] Karpiński, M., Jarmołowicz-Nowikow, E., Malisz, Z. 2008. Aspects of gestural and prosodic structure of multimodal utterances in Polish task-oriented dialogues [In:] G. Demenko, K. Jassem, S. Szpakowicz (Eds.) *Speech and Language Technology*, vol. XI, Polish Phonetic Association: Poznań, pp. 113-122.

[15] Kendon, A. *Gesture. Visible Action as Utterance.* Cambridge: CUP.

[16] Kita, S. 1990. The Temporal Relationship between Gesture and Speech: A Study of Japanese-English Bilinguals. MA thesis, University of Chicago.

[17] Klessa, K., Szymański, M., Breuer, S., Demenko, G. 2007. Optimization of Polish Segmental Duration Prediction with CART, *SSW6,* Bonn.

[18] Korolko, M. 1998. Sztuka retoryki. Przewodnik encyklopedyczny. (The Art of Rhetorics. Encyclopedic Guide.) Warszawa: Wiedza Powszechna.

[19] Lickley, R. J. 1995. Missing disfluencies. In *Proceedings of the International Congress of Phonetic Sciences* (Vol. 4) pp. 192–195.

[20] Lickley, R. J., Bard, E. G. 1996. On not recognizing disfluencies in dialogue. Proc. ICSLP 96, pp. 1876-1879.

[21] Martin, J. G., Strange, W. 1968. The perception of hesitation in spontaneous speech. *Perception and Psychophysiology* 3(6), pp. 427-438.

[22] McNeill, D. 2007. Gesture & Thought. University of Chicago Press.

[23] Shriberg, E. E. 1994. Preliminaries to a Theory of Speech Disfluencies. PhD thesis, University of California at Berkeley.

[24] Shriberg, E. E. 2001. To "Errr" is Human: Ecology and Acoustics of Speech Disfluencies.*Journal of the International Phonetic Association* 31(1), pp. 153-169.

[25] Swerts, M. 1998. Filled Pauses as Markers of Discourse Structure. *Journal of Pragmatics*, Volume 30, Number 4, October 1998 , pp. 485-496(12)Pon-Barry, H., Shieber, S. 2009. The Importance of Sub-Utterance Prosody in Predicting Level of Certainty.Proceedings of NAACL HLT 2009: Short Papers, pp. 105–108.

[26] Turk, A. E., Shattuck-Hufnagel, S. 2007. Multiple targets of phrase-final lengthening in American English words, *Journal of Phonetics*, Volume 35, Issue 4, pp. 445-472.

[27] Wagner, A. 2008. A comprehensive model of Polish intonation. PhD thesis presented at Adam Mickiewicz University.

[28] Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., Price, P. J. 1992. Segmental Durations in the Vicinity of Prosodic Phrase Boundaries", *JASA 91(3),* pp. 1707-1717,