PITCH CHARACTERISTICS OF FILLED PAUSES

Malte Belz[1], Uwe D. Reichel[2]

[1] Humboldt-Universität zu Berlin, [2] Ludwig-Maximilians-Universität München
malte.belz@hu-berlin.de, reichelu@phonetik.uni-muenchen.de

ABSTRACT

We investigate the pitch characteristics of filled pauses in order to distinguish between hesitational and floor-holding functions of filled pauses. A corpus of spontaneous dialogues is explored using a parametric bottom-up approach to extract intonation contours. We find that subjects tend to utter filled pauses more prominently when they cannot see each other, which indicates an increased floor-holding usage of filled pauses in this condition.

Keywords: Disfluencies, Filled Pauses, Intonation, Floor-holding

1. TURN-HOLDING AND INTONATION

Filled pauses such as *uh* or *um* have been assigned different functionalities. In this study, we explore the pitch characteristics of filled pauses in spontaneous dialogues and whether different f_0 contours distinguish between different functions.

Different hypotheses try to account for hesitations. The floor-holding hypothesis [7, 8] claims that filled pauses are used to hold the turn during speaking. This seems to be more challenging in heated debates. Lallgee and Cook [6] confront interlocutors with a high pressure situation by means of a quickly interrupting confederate. However, this setting does not effect the frequency of filled pauses as opposed to low pressure situations. The authors suggest that speakers use other means for signalling turn-holding, for example by raising their voice.

A factor that might have an additional effect on the frequency of filled pauses is visual contact. Kasl and Mahl [5] note that filled pauses occur more frequently in situations where interlocutors cannot see each other. This indicates that some compensation processes take place in remote communication sitations, as speakers in natural face to face dialogues mark turn-holding by non-verbal cues as well [4]. When bereft of this modality, they might compensate not only by increasing the filled pause frequency, but also by using other verbal cues more often, e.g., a high f_0 level [19] (for Swedish and English).

Nevertheless, the mere frequency information of filled pauses might distort the picture, as the distribution of filled pauses is widespread and other explanations such as planning problems are possible. Therefore, a closer look at the intonation of filled pauses might help discriminate the filled pauses that might enable, and account for, turn-holding. We will explore the link between a rising fundamental frequency (f_0) and filled pauses, two features of spontaneous dialogues that are used for turn management. Thus, intonation patterns of filled pauses may help with identifying the functional implementations of filled pauses.

Previous research on the intonation of filled pauses implies that they show gradual downsteps of f_0 [10], at least within clauses [17]. This is in accordance with Tseng [18] , who states that filled pauses usually show flat intonational contours. This notion may also be described as showing only little or no prominence. We explore whether there are prominent contours for filled pauses and whether speakers make use of them to compensate in remote dialogue situtations. We will use a bottom-up approach for extracting the relevant intonation contours.

2. METHOD AND DATA

2.1. f_0 Preprocessing

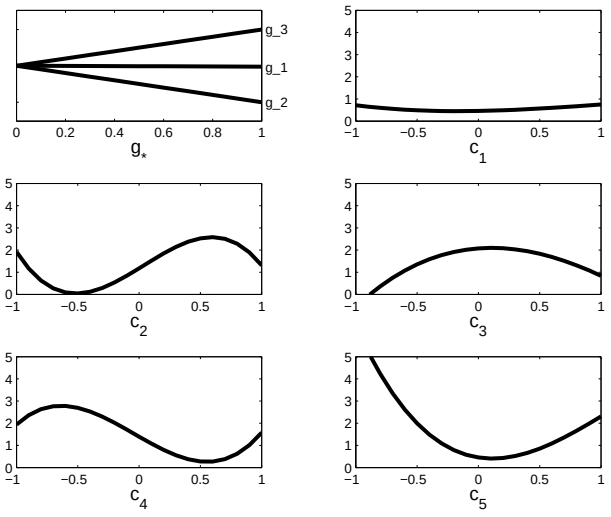
f_0 was extracted by autocorrelation (PRAAT 5.3.16 [3], sample rate 100 Hz). Voiceless utterance parts and f_0 outliers were bridged by linear interpolation. The contour was then smoothed by Savitzky-Golay filtering using third order polynomials in 5 sample windows and transformed to semitones relative to a base value [15]. This base value was set to the f_0 median below the 5th percentile of an utterance and serves to normalize f_0 with respect to its overall level.

2.2. Stylization

For intonation stylization we adopt the parametric CoPaSul approach of Reichel [13]. Within this framework, intonation is stylized as a superposition of linear global contours representing the f_0 baseline and third order polynomial local contours rep-

resenting the local pitch movements related to accent groups. The domain of global contours approximately related to intonation phrases is determined automatically by placing prosodic boundaries at speech pauses and punctuation in the aligned transcript. Within this domain, the baseline is fitted in a robust way through a sequence of f_0 medians of values below the 10^{th} percentile [14]. The domain of local contours is determined by placing boundaries behind each content word determined by POS tagging [12]. Thus, these local contour domains roughly correspond to syntactic chunks [1] and generally contain at most one pitch accent. Additionally to this parametric analysis we carried out a kmeans clustering of the stylization coefficient vectors as in [9, 13] to describe the f_0 shapes also in categorical terms. The resulting contour classes are displayed in Fig. 1.

Figure 1: Global (g) and local (c_n) stylized intonation contour classes. The vertical axis represents semitones, the horizontal axis represents normalized time.



2.3. Data

Our database consists of GECO [16], a corpus of 48 German spontaneous mono- and multimodal dialogues (25 min each) of 13 women aged from 20 to 25 years who are strangers to each other. Only the multimodal condition allows subjects to see each other. Seven subjects participate in both conditions.

3. RESULTS

3.1. Frequencies

First, we test the hypothesis whether subjects produce more filled pauses in the monomodal condition (no visual cues) as predicted by the compensation hypothesis. The seven subjects that participate in both modality conditions, however, produce significantly less filled pauses in the monomodal condition, as computed by a one-sample t-test on the mean of the filled pauses frequencies ($\mu_0 = 0.5$, $t = -3.7$, $df = 6$, $p < 0.01$). Thus, frequencies do not confirm the compensation hypothesis. Nevertheless, the intonational features may still show such an effect.

3.2. Pitch characteristics

Table 1 shows that all subjects utter filled pauses with a higher proportion of steady intonation (global class g_1) in comparison to words with two to three characters ($t_{paired} = -16.1, df = 12, p < 0.001$). At first glance, it seems that the findings of [10] and [17] are corroborated, as 88% of filled pauses are uttered with a steady f_0 contour.

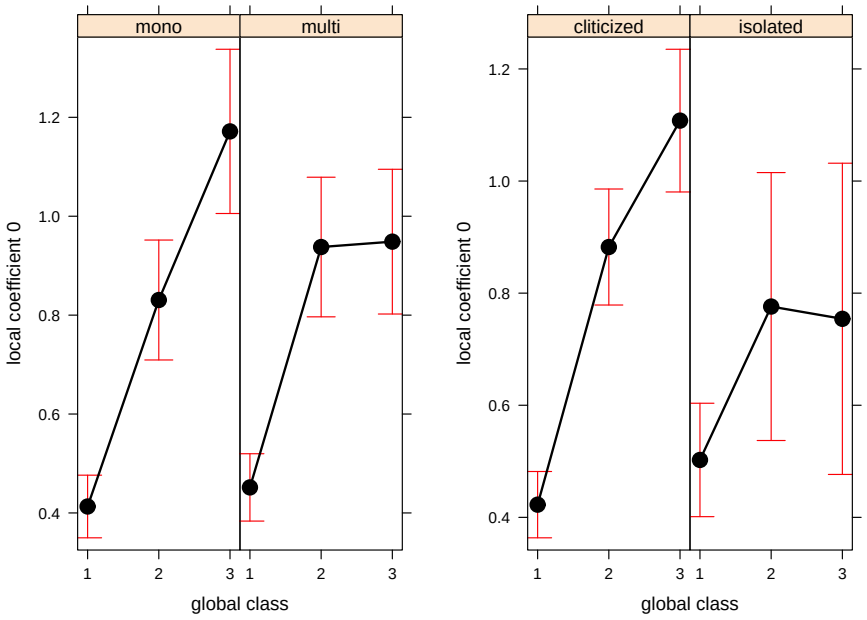
Table 1: Frequencies of global and local intonation classes, filled pauses and 2–3 character words.

Global class	1	2	3
Words	0.52	0.25	0.23
Filled pauses	0.88	0.07	0.05

Rising and falling f_0 baselines are reflected by global classes g_3 and g_2, respectively (cf. Fig 1). As every global class is described in more detail by overlaying local classes, we may now investigate prominence. Prominence is mainly modelled by the zeroth order polynomial coefficient of the local contour (coefficient 0). High values indicate a prominence-lending f_0 deviation from the baseline. To explore our data, we start with a full linear regression model of prominence, using the predictors modality (mono vs. multi), isolation (clitic vs. non-clitic), global class (steady vs. falling vs. rising) and their interactions as fixed effects and subjects as random effects in R [11] using the *lme4*-package (v.1.1-7). The isolation factor allows us to account for filled pauses in the context of a left and right silent pause, as opposed to cliticized filled pauses. We insert this factor as a fixed effect because its isolated position might lend the filled pauses more prominence *qua* isolation. We use the same cut-off for the left and right tail of the distribution, which removes 2.2% of the data. The model is reduced by a stepwise backward selection based on AIC [2].

Two two-way interactions remain in the model: global class interacts with modality ($\chi^2 = 6.8$, $df = 2$, $p = 0.03$) as well as with isolation ($\chi^2 = 9.0$, $df = 2$, $p = 0.01$), as confirmed by a log-likelihood

Figure 2: Effects of the multi- vs. monomodal and the cliticized vs. isolated predictor. The local coefficient 0 is used as a measure of prominence over global classes 1 (steady), 2 (falling) and 3 (rising).



test (cf. Fig. 2). The effect in the modality condition suggests that filled pauses are marked more prominently when interlocutors cannot see each other. The effect in the isolation condition does not hold for modality, but suggests that filled pauses in cliticized positions are marked more prominently than in isolated positions.

4. CONCLUSION

From these first exploratory findings we can conclude that filled pauses might in fact differ in their intonational appearances, and that these vary systematically. Instances in which filled pauses are uttered prominently—either with a rising or falling f_0 contour—indicate that these specific filled pauses can be used for holding the turn, especially so if interlocutors cannot see each other. Thus, our analysis implies that speakers are compensating the lack of visual contact in the monomodal condition with the help of intonation.

For further research, interesting insights to the functions and distributions of disfluencies may come from comparing the context of filled pauses showing a 'turn holding contour' to those of other instances. Additionally, annotating the data for turn-holding signals may reveal whether the detected instances are in fact perceived as holding the turn.

5. REFERENCES

[1] Abney, S. 1991. Parsing by chunks. In: Berwick, R., Abney, S., Tenny, C., (eds), *Principle-Based Parsing*. 257–278.

[2] Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* AC-19, 716–723.

[3] Boersma, P., Weenink, D. 1999. PRAAT, a system for doing phonetics by computer. Technical report Institute of Phonetic Sciences of the University of Amsterdam. 132–182.

[4] Goodwin, C. 1981. *Conversational Organization*. Academic Press.

[5] Kasl, S. V., Mahl, G. F. 1965. Relationship of disturbances and hesitations in spontaneous speech to anxiety. *J Pers Soc Psychol* 1, 425–433.

[6] Lallgee, M. G., Cook, M. 1969. An experimental investigation of the function of filled pauses in speech. *Language and Speech* 12, 24–28.

[7] Maclay, H., Osgood, C. E. 1959. Hesitation phenomena in spontaneous English speech. *Word* 5, 19–44.

[8] Malandro, L. A., Barker, L., Barker, D. A. 1989. *Nonverbal Communication*. Random House.

[9] Möhler, G., Conkie, A. 1998. Parametric modeling of intonation using vector quantization. *Proc. 3rd ESCA Workshop on Speech Synthesis* 311–316.

[10] O'Shaughnessy, D. 1992. Recognition of hesitations in spontaneous speech. *Acoustics, Speech, and Signal Processing, ICASSP-92*. volume 1.

IEEE 521–524.
[11] R Core Team, 2012. R: A language and environment for statistical computing.
[12] Reichel, U. D. 2007. Improving data driven part-of-speech tagging by morphologic knowledge induction. *Proc. AST Workshop* 65–73.
[13] Reichel, U. D. 2014. Linking bottom-up intonation stylization to discourse structure. *Computer, Speech, and Language* 28, 1340–1365.
[14] Reichel, U. D., Mády, K. 2014. Comparing parameterizations of pitch register and its discontinuities at prosodic boundaries for Hungarian. *Proc. Interspeech* 111–115.
[15] Savitzky, A., Golay, M. 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry* 36(8), 1627–1639.
[16] Schweitzer, A., Lewandowski, N. 2013. Convergence of articulation rate in spontaneous speech. *Proc. Interspeech* 525–529.
[17] Shriberg, E. E., Lickley, R. J. 1993. Intonation of clause-internal filled pauses. *Phonetica* 50, 172–179.
[18] Tseng, S.-C. 1999. *Grammar, Prosody and Speech Disfluencies in Spoken Dialogues*. University of Bielefeld, Germany.
[19] Zellers, M. 2014. Duration and pitch in perception of turn transition by Swedish and English listeners. *Proc. from FONETIK* 41–46.