



The intonation of echo *wh*-questions

Sophie Repp, Lena Rosin

Humboldt-Universität zu Berlin, Germany

sophie.repp@rz.hu-berlin.de, lena.rosin@cms.hu-berlin.de

Abstract

The acoustic characteristics of German echo questions are explored in a production study. It is shown that there are prosodic differences (F0, duration, intensity) between echo questions signalling a high level of emotional arousal, echo questions signalling that the speaker did not understand the previous utterance, and questions requesting completely new information. The findings are largely compatible with earlier findings on utterances with different levels of emotional arousal, where e.g. a higher F0 signals a higher emotional arousal but do not confirm expectations with respect to phonological differences formulated on the basis of suggestions in the linguistic literature on echo questions.

Index Terms: echo questions, emotional speech, prosody

1. Introduction

Echo questions are questions that "echo", i.e. repeat, part of a previous utterance as in *A: John called Mary. B: John called who?*¹ Structurally, echo questions take the form of the previous utterance – in the example this is the form of a declarative sentence – so that in contrast to normal information-seeking questions, the interrogative *wh*-phrase (*who*) does not normally occur in clause-initial position but *in-situ*, i.e. in the position where the phrase occurs that the *wh*-phrase replaces. Echo questions can be asked for various reasons. The speaker might not have understood the previous utterance due to an auditory failure, the speaker might request clarification about what an expression in the previous utterance referred to (so-called *reference questions* ([1], [2]), e.g. *A: David took him to the vet. B: David took who to the vet?*, [2]: 213), the speaker might not believe what s/he just heard and wants to double-check, or the speaker might be amazed or indignant, i.e. emotionally aroused, about what s/he just heard and wishes to express his/her emotions. Thus, the intent and the emotional underpinnings of an echo question can be rather varied and depend on the situational and linguistic context.

In the linguistics literature it is assumed that echo questions come with special intonational marking. For English as well as for German – the language of investigation in this study – it has been suggested that echo questions obligatorily end with a rise except for reference questions, which end with a fall (e.g. [2], [3], [4], [5]). The *in-situ wh*-phrase is generally thought to be narrowly focussed (e.g. [2], [4], [5]), and to carry the nuclear accent, which can be a (L+)H* or a L* pitch accent. The choice of accent has been suggested to correlate with the intent and the emotional underpinning of the question ([1], [2]: 174, [3]): (L+)H* followed by H-H% typically signals that the speaker did not understand the previous utterance due to an auditory failure. The H% boundary tone might also be replaced by L% in such contexts. Emotional arousal due to disbelief or surprise, in contrast, is typically signalled by a L* accent followed by H-H%. For German, there have been no

suggestions about intonational differences between emotional and auditory failure echo questions.

Controlled empirical investigations of the above proposals for the intonation of echo questions are lacking. From the point of view of findings on the impact of emotional arousal on intonation we expect that there should indeed be differences between auditory failure echo questions and emotionally aroused echo questions. Studies on the production of prosodic cues signalling emotions and the perception of such cues (see e.g. [6] for an overview) investigate, on the one hand, different kinds of emotions, e.g. fear vs. anger vs. joy, and, on the other hand, different levels of emotional arousal, e.g. mild fear vs. panic. For the prosodic realization of echo questions, the kind of emotion can be relevant: the speaker of an echo question expressing disbelief or surprise might be pleasantly or unpleasantly surprised. The level of emotional arousal also is relevant because when uttering an echo question which signals auditory failure the speaker will be aroused relatively little in comparison to when s/he utters an echo question signalling disbelief and indignation. We concentrate on prosodic cues for levels of arousal here because in the experiment reported below we kept the kind of emotion involved constant: we tested only one type of emotional echo question in comparison to plausibly non-emotional echo questions. For emotional arousal, production studies have shown that a speaker's level of emotional arousal correlates with his/her use of vocal cues like F0 level, intensity and voice quality. For instance, Bänziger & Scherer [7] report a production study where speakers uttered non-sense syllable strings for which the level of emotional arousal was a good predictor for the mean F0 of pitch accents: it overall was higher for high than for low levels of emotional arousal. The F0 range also was higher for higher emotional arousal. Furthermore, the slope of rising pitch accents in an utterance was steeper for high levels than for low levels of emotional arousal in some types of emotions. Finally, pitch peak position varied with different levels of the emotion *joy*: the peak occurred later for high emotional levels of joy. Duration and intensity were not investigated in this study. In perception studies, it was shown that listeners use F0 as well as amplitude to identify different emotions ([8] and subsequent literature), and different levels of emotional arousal. For instance, Ladd et al. [9] showed that a greater F0 range correlates with listeners' ratings of higher emotional intensity. Similarly, Scherer et al. [10] found that the F0 level correlates with levels of perceived emotional intensity.

The current study is a production study investigating the acoustic characteristics of echo questions in German. It addresses the issue of whether or not speakers distinguish by prosodic means between (1) questions that contextually function like ordinary information-seeking questions but come in the form of echo questions, (2) echo questions signalling auditory failure, and (3) echo questions signalling indignation (i.e. a high level of emotional arousal), and if so, how they do it. In contrast to many earlier production studies on emotional utter-

ances the speakers in the present study were naive speakers (i.e. not actors), and they were not instructed to act out a particular (level of) emotion but they took part in short dialogues where they were required to react in a natural, situationally appropriate way to an utterance of another speaker. Thus, speakers did not have to choose consciously an intonation that expressed e.g. indignation, but they chose an intonation that fitted the context, see below for details. In accordance with the literature on prosodic reflexes of emotions, we predicted that echo questions signalling indignation would come with higher F0 levels and a greater F0 range than the information-seeking question and than the echo question signalling auditory failure. In accordance with the literature on the prosody of echo questions, we predicted that all echo questions would end in a rise. Furthermore, since English and German are comparable in their intonational characteristics, we predicted from the claims about English echo questions that the *wh*-word and the subsequent clause-final region in German should also be realized with a different contour in echo questions signalling auditory failure in comparison to echo questions signalling indignation ((L+)H* H-% vs. L*H-H% / L*H-L%). With respect to the information-seeking question, and potential differences in comparison to the auditory failure question, the study was of an exploratory nature. The non-*wh*-part in the information-seeking question could have a higher F0 than in the auditory failure echo question because the former is not a repetition of the immediately preceding utterance (see Table 1 for illustration), which might have an influence on its givenness status and thus on the flatness / deaccentuation of the contour (see e.g. [11] on the prosody of givenness). Still, the information-seeking question like the other questions did have an antecedent in the left linguistic context – it just appeared earlier –, so the non-*wh*-part can still be considered given and therefore need not be different from the auditory failure question. With respect to emotional arousal, it is plausible to assume that an auditory failure question does not necessarily have a higher level of emotional arousal than an information-seeking questions. However, speakers might want to convey that an auditory failure question has a specific intent, namely to make clear that the previous utterance was not heard properly. Naively, one might assume that increasing the intensity and the duration in the question could be used to cue the addressee to be particularly clear in his/her repetition.

2. Method

2.1. Design and Materials

The experiment had a one-factorial design with three conditions that corresponded to three types of questions which all had the structure of echo questions. They were declarative clauses with an in-situ *wh*-word (*wen* 'whom'), marked with a question mark at the end, e.g.:²

Target question: Und Anja will wen ermahnen?
(example) and Anja wants who.ACC reprimand
'And Anja wants to reprimand whom?'

The target question was part of a short (fake telephone) dialogue between two speakers, where the first speaker described a state-of-affairs, and the second speaker reacted by uttering the target question and by giving some additional information which allowed conclusions about the intent of the question and the emotional arousal of the speaker (if any). The three con-

ditions are illustrated in Table 1. In the first condition, NEW.INFO, the second speaker requests information that was not given before. In the second condition, REPEAT.INFO, the second speaker requests information that was just given but that was difficult to understand. In the third condition, INDIGNANT, the second speaker expresses her indignation about the state-of-affairs that was just described by the first speaker.

The test materials consisted of eight different dialogue sets (lexicalizations) where the target sentence had the same metrical structure as in the example. There were $8 \times 3 = 24$ experimental items as well as 32 fillers (discourses with *wh*-exclamatives and with ordinary *wh*-questions). The first speaker's turn was presented both auditorily from a pre-recorded audio file, as well as in written form. In condition 2, REPEAT.INFO, the object in the antecedent clause, which was the requested information in the echo question, was replaced by a masking noise in the audio file, and by the information [*Rauschen*] ('noise') in the written text. The second speaker's turn was presented in written form only.

Table 1. *Experimental Materials. Example*

<p>Condition 1: NEW.INFO Speaker A: Tina will Sven und Mark wegen der häufigen Prügeleien ermahnen. Aber Anja will jemand anderen ermahnen, der viel mehr Unruhe stiftet, und darüber diskutieren sie und Tina nun schon seit Tagen! (<i>Tina wants to reprimand Sven and Mark because of their frequent fights. But Anja wants to reprimand a different person, who causes much more trouble, and Tina and she have been arguing about this issue for days!</i>) Speaker B: Und Anja will wen ermahnen? Wer ist denn ihrer Meinung nach der schlimmere Unruhestifter? (<i>And Anja wants to reprimand whom? Who is a worse trouble-maker, according to her?</i>)</p>
<p>Condition 2: REPEAT.INFO Speaker A: Tina will Mark und Sven wegen der häufigen Prügeleien ermahnen. Und Anja will [Rauschen] ermahnen. (<i>Tina wants to reprimand Sven and Mark because of their frequent fights. And Anja wants to reprimand [NOISE].</i>) Speaker B: Und Anja will wen ermahnen? Tut mir leid, ich bin gerade in die U-Bahn gestiegen. Da ist die Verbindung manchmal schlecht. (<i>And Anja wants to reprimand whom? I'm sorry. I just got on the underground. The connection is sometimes bad there.</i>)</p>
<p>Condition 3: INDIGNANT Speaker A: Tina will Mark und Sven wegen der häufigen Prügeleien ermahnen. Und Anja will Lotta ermahnen. (<i>Tina wants to reprimand Sven and Mark because of their frequent fights. And Anja wants to reprimand Lotta.</i>) Speaker B: Und Anja will wen ermahnen? Ich fass es nicht, warum will sie das denn machen? Lotta kann doch keiner Fliege was zuleide tun! (<i>And Anja wants to reprimand whom? I don't believe this! Why does she want to do that? Lotta wouldn't hurt a fly!</i>)</p>

2.2. Participants and procedure

9 female speakers (mean age 25.1) of Standard German from the Berlin-Brandenburg area took part in the experiment. The experiment was run using the software *Presentation* (Neurobehavioral systems). Participants took the role of the second speaker in the dialogues. First they heard and read the pre-recorded text of the first speaker. Then they quietly read the reply of the second speaker. When they felt they had understood

the reply they recorded it. They were asked to speak in the way that they found most natural in the given context. It was pointed out to them that some utterances might be passionate (*emotional* in German), e.g. surprised, indignant and the like. Items were recorded in a pseudo-randomized order.

2.3. Analysis

10 of 216 utterances (4.6 % data points) were discarded due to hesitations or speech errors. This left 206 utterances for analysis. They were annotated by hand for syllable and word boundaries in PRAAT [12]. Voice pulses were corrected manually. Acoustic measures were drawn from the data using the PRAAT script ProsodyPro [13], which was also used to create the time-normalized F0 data displayed in Figure 1.

For the acoustic investigation, maximum, minimum and mean F0 (henceforth $F0_{max}$, $F0_{min}$, $F0_{mean}$), duration and intensity were taken for the stressed first syllable and the unstressed second syllable of the subject, for the auxiliary, the stressed *wh*-word, for the unstressed first syllable of the main verb, and for the conjoined stressed second and the unstressed third syllable of the main verb (since speakers realized the third syllable of the main verb in only 60.7% of the utterances (uncorrelated with conditions), analysis was run for both syllables together). We only report the results for the most relevant syllables here (subject, *wh*-word, main verb). F0 excursion ($F0_{exc}$) was calculated for the two rises in the utterance (see Figure 1): the rise on the subject, and the rise on the *wh*-word up to the verb's first syllable. For these rises, peak position also was determined. The statistical analysis was carried out on raw data. We applied general mixed effects models [14] with question type as fixed factor. The simplest best-fitting models (identified by likelihood ratio tests) included random intercepts for participants, and usually random slopes for the INDIGNANT condition per participant and lexicalization as well as random intercepts for lexicalization. In the INDIGNANT condition, there was greater speaker variation in comparison to the other conditions. Single comparisons were run for the best-fitting models with the generalized linear hypothesis test `glht` with Tukey correction [15].

3. Results

Figure 1 shows the time-normalized F0 contour for the three question types across all speakers, which is based on ten equally distributed F0 measurements per syllable for each utterance [13]. Figure 1 shows that all three question types were realized in a similar way. They all ended with a final rise. In the prenuclear region there was a rising pitch accent (L*+H) on the subject. The nuclear contour started with a L* accent on the *wh*-word, followed by a steep rise to a H-^H% phrasal accent and boundary tone.

Statistical analysis revealed the following significant differences between the three question types (see Figures 2 and 3 for a visualization of the duration and intensity results). On the **stressed first syllable of the subject** INDIGNANT questions had higher means than NEW.INFO questions for $F0_{max}$ ($z = -2.4$, $p < .05$) and for $F0_{mean}$ ($z = -3.1$, $p < .01$). INDIGNANT questions had higher means than REPEAT.INFO questions for duration ($z = -2.6$, $p < .05$). REPEAT.INFO questions had marginally higher means than NEW.INFO questions for $F0_{max}$ ($z = -2.2$, $p = .07$), and marginally lower means for duration ($z = 2.2$, $p = .07$).

On the **unstressed second syllable of the subject**, INDIGNANT questions had higher means than NEW.INFO ques-

tions for $F0_{min}$ ($z = -2.5$, $p < .05$) and for intensity ($z = -3.0$, $p < .01$). REPEAT.INFO questions had higher means than NEW.INFO questions for $F0_{max}$ ($z = -2.3$, $p < .05$), $F0_{min}$ ($z = -3.1$, $p < .01$), $F0_{mean}$ ($z = -2.4$, $p < .05$) and duration ($z = 3.1$, $p < .01$). The **pitch excursion of the rise on the subject**, i.e. the difference between the minimum F0 in the first syllable and the maximum F0 in the second syllable did not differ between conditions.

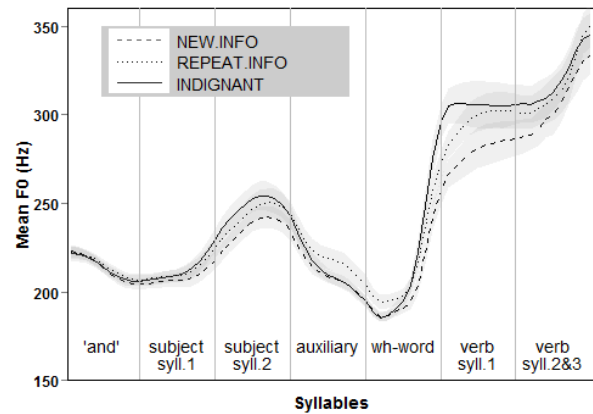


Figure 1: Time-normalized F0 contour averaged across all speakers with 95% confidence interval

On the ***wh*-word**, INDIGNANT questions had higher means than NEW.INFO questions for $F0_{max}$ ($z = -3.6$, $p < .001$), $F0_{mean}$ ($z = -3.3$, $p < .01$), duration ($z = -4.3$, $p < .001$), and intensity ($z = -2.4$, $p < .05$). INDIGNANT questions had higher means than REPEAT.INFO questions for $F0_{max}$ ($z = 2.3$, $p < .05$) and duration ($z = -5.0$, $p < .001$), and lower means for $F0_{min}$ ($z = 2.4$, $p < .05$). REPEAT.INFO questions had higher means than NEW.INFO questions for $F0_{min}$ ($z = -2.4$, $p < .05$), $F0_{mean}$ ($z = -4.7$, $p < .001$) and for intensity ($z = -3.5$, $p < .01$).

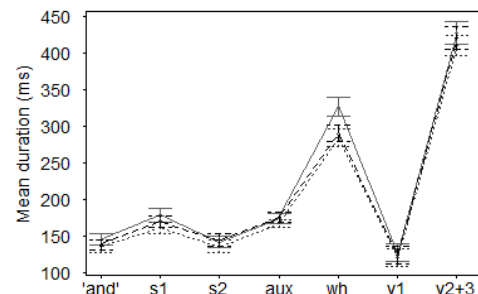


Figure 2: Mean duration with 95% confidence interval. See Figure 1 for legend

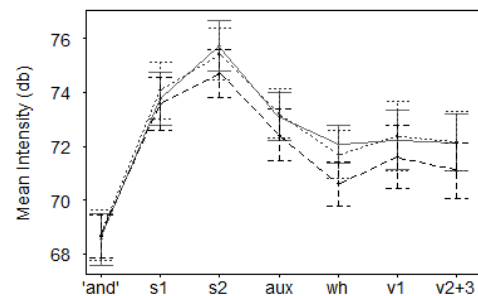


Figure 3: Mean intensity with 95% confidence interval. See Figure 1 for legend

On the **main verb's unstressed first syllable**, INDIGNANT questions had higher means than NEW.INFO questions for $F0_{\max}$ ($z = -2.8, p < .05$), $F0_{\min}$ ($z = -3.3, p < .01$) and $F0_{\text{mean}}$ ($z = -2.9, p < .01$). INDIGNANT questions had higher means than REPEAT.INFO questions for duration ($z = -2.4, p < .05$). REPEAT.INFO questions had significantly higher means than NEW.INFO questions for $F0_{\max}$ ($z = -3.2, p < .01$), $F0_{\min}$ ($z = -3.2, p < .01$), and $F0_{\text{mean}}$ ($z = -2.4, p < .05$).

The **pitch excursion of the rise on the *wh*-word** up to the first syllable of the main verb was larger in INDIGNANT than in REPEAT.INFO questions ($z = -3.4, p < .01$) and than in NEW.INFO questions ($z = -3.8, p < .001$). The peak was reached marginally earlier in INDIGNANT than in NEW.INFO questions ($z = 2.2, p = .08$).

On the **last syllable(s) of the main verb**, INDIGNANT questions had higher means than NEW.INFO questions for $F0_{\min}$ ($z = -2.5, p < .05$) and for $F0_{\text{mean}}$ ($z = -2.4, p < .05$). INDIGNANT questions had higher means than REPEAT.INFO questions for duration ($z = -2.6, p < .05$). REPEAT.INFO questions had higher means than NEW.INFO questions for $F0_{\max}$ ($z = -3.2, p < .01$), $F0_{\text{mean}}$ ($z = -3.0, p < .01$) and for intensity ($z = -3.0, p < .01$).

4. Discussion

The results indicate that speakers distinguish reliably between the three question types under investigation. Differences can be found for several of the F0 measures, for intensity and for duration. The greatest differences can be found on the *wh*-word and subsequent clause-final region but other regions in the utterances differ as well. The most important results are the following. All three types of questions end in a rise. On the subject, speakers realize a L*+H accent which has a higher F0 (max/min) in indignant and auditory failure questions than in information-seeking questions. The former two questions only differ reliably in the duration, which is longer in indignant questions. On the focussed *wh*-word, speakers realize a L* accent in all questions, i.e. the predictions from the theoretical linguistic literature with respect to accentual differences between auditory failure (H*) and indignant questions (L*) in English ([1], [2], [3]) could not be confirmed for German. Phonetically, the L* accent shows differences between question types. It has a lower minimum F0 in indignant and in information-seeking questions than in auditory failure questions. The rise following the L* accent reaches a plateau on the next syllable and then continues towards the end of the question ending in an upstepped boundary tone (nuclear contour: L* H-^H%). Again, the suggestions in the linguistic literature on different boundary tones for different types of echo questions in English ([1], [2], [3]) could not be confirmed for German. Turning to the acoustics of the phrase and boundary tones, results show that the plateau is highest in indignant questions followed by auditory failure questions and then by information-seeking questions. It is reached earlier in indignant than in new information questions. The pitch excursion from the L* up to the plateau is highest for indignant questions, which may be considered as being a result both of the low L* accent and the high H- tone in this condition. Apart from the pitch measures, duration was different for the question types in this region of the utterance: indignant questions had longer syllables than auditory failure questions. Finally, intensity was higher in auditory failure questions than in information-seeking questions.

Overall, the results are compatible with previous findings on prosodic reflexes of emotional arousal. There are clear pro-

sodic differences – both in F0 measurements and in duration – between the two non-emotional question types and the emotionally aroused indignant questions. F0 maxima are higher, duration is longer and intensity is greater for the latter – to different extents in different regions of the utterance, with greatest differences on the focussed *wh*-word and subsequent syllable. We also observed, however, that F0 minima are lower for emotionally aroused questions, at least in comparison to auditory failure questions (but, surprisingly, not in comparison to information-seeking questions). The lower minimum contributes to a greater pitch excursion, which is compatible with previous findings on a greater pitch range for high levels of emotional arousal [7].

Turning to the comparison between auditory failure question and information-seeking questions, we found that the former had higher duration and intensity values than the latter in several syllables across the utterance but maximum F0 also often was higher. These findings are not compatible with a difference in givenness status but they show that speakers signal that the auditory failure question is not an ‘ordinary’ information-seeking question, as was hypothesized in the introduction. We cannot say at present whether this is to signal to the addressee that s/he should speak in a particularly clear way, i.e. is due to the intent of the question, or whether some level of emotional arousal is involved. Overall, the results suggest a graded continuum of the strength of phonetic cues applied in the three question types, especially in the nuclear region, with indignant questions having the strongest cues (i.e. higher values), followed by auditory failure questions, followed by information-seeking questions. However, it is an issue for future research to determine whether this continuum reflects a continuum of pragmatic meaning or emotional arousal, or whether it does not.

5. Conclusions

The present study has shown that different types of echo questions show different prosodic characteristics depending on the emotional arousal of the speaker and/or the intent of the question, as triggered by the linguistic context for naive speakers. The prosodic differences are of a gradient phonetic rather than a phonological nature, contrary to expectations on the basis of claims in earlier literature on echo questions. The findings overall are compatible with earlier findings on prosodic reflexes of different levels of emotional arousal. The study has shown that F0 and intensity as well as duration are manipulated by speakers to mark the respective questions.

6. Acknowledgements

This work was supported by the German Research Foundation DFG as part of the Collaborative Research Centre (SFB) 632 ‘Information Structure’ at the Humboldt-Universität zu Berlin. We thank Andreas Haida for speaking the text of the first speaker, and Felix Golcher for advice on the statistics.

Notes

¹ This is a simplification. Below, we show that *wh*-in-situ questions can also be used to request information that was not given in the immediate context (also cf. [5]).

² *Wh*-words also can have an indefinite reading in German (*wen* = ‘someone’), if they are unstressed. The experimental contexts did not support such a reading.

7. References

- [1] E. Rando, "Intonation in discourse", in *The Melody of Language*, L.R. Waugh and C.H. van Schooneveled, Eds. Baltimore: University Park Press, 1980, pp. 243-278.
- [2] C. Bartels, *The Intonation of English Statements and Questions. A Compositional Interpretation*. Garland Publishing, 1999.
- [3] D. Bolinger, *Interrogative Structures of American English: The Direct Question. Amer. Dialect Soc. 28*. Birmingham: University of Alabama Press, 1957.
- [4] R. Artstein, "Parts of words: compositional semantics for prosodic constituents," Ph.D. dissertation, New Brunswick Rutgers, The State University of New Jersey, NJ, 2002.
- [5] M. Reis, "On the analysis of echo questions," *Tampa Papers in Linguistics*, vol. 3, pp. 1-24, 2012.
- [6] P. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?," *Psychol. Bull.*, vol. 129, no. 5, pp. 770-814, 2003.
- [7] T. Bänziger and K. Scherer, "The role of intonation in emotional expressions," *Speech Commun.*, vol. 46, no. 3-4, pp. 252-267, 2005.
- [8] P. Lieberman and S. Michaels, "Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech," *J. Acoust. Soc. Amer.*, vol. 34, no. 7, pp. 922-927, 1962.
- [9] D. Ladd, K. Silverman, F. Tolkmitt, G. Bergmann, and K. Scherer, "Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect," *J. Acoust. Soc. Amer.*, vol. 78, no. 2, pp. 435-444, 1985.
- [10] K. Scherer, D. Ladd, and K. Silverman, "Vocal cues to speaker affect: Testing two models," *J. Acoust. Soc. Amer.*, vol. 76, no. 5, pp. 1346-1356, 1984.
- [11] S. Baumann, *The Intonation of Givenness – Evidence from German*. Tübingen: Niemeyer, 2006.
- [12] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," Available: <http://www.fon.hum.uva.nl/praat/>, 2014.
- [13] Y. Xu, "ProsodyPro — A tool for large-scale systematic prosody analysis," *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP)*, Aix-en-Provence, 2013.
- [14] D. Bates, B. Bolker, M. Maechler and S. Walker, "Linear mixed-effects models using Eigen and S4," R Package lme4, Version: 1.0-4, 2013.
- [15] T. Hothorn, F. Bretz and P. Westfall, "Simultaneous inference in general parametric models," R Package multcomp, Version: 1.3-2, 2014.