The following paper was presented at The 9th Workshop on Disfluency in Spontaneous Speech (DiSS 2019) held at ELTE Eötvös Loránd University in Budapest, Hungary on 12–13 September, 2019.

Title: On the role of disfluent speech for uncertainty in articulatory speech synthesis

Author(s): Charlotte Bellinghausen, Thomas Fangmeier, Bernhard Schröder, Johanna Keller, Susanne Drechsel, Peter Birkholz, Ludger Tebartz van Elst and Andreas Riedel

Abstract: In this paper we present a perception study on the role of disfluent speech in forms of prosodic cues of uncertainty in question-answering situations. In our scenario the answer to each question was modeled by varying three prosodic cues: pause, intonation, and hesitation. The utterances were generated by means of an articulatory speech synthesizer. Subjects were asked to rate each answer on a Likert scale with respect to uncertainty, naturalness and understandability. Results showed evidence for an additive principle of the prosodic cues, i.e. the more cues were activated the higher the perceived level of uncertainty. Overall, the effect of intonation and hesitation was more evident than the effect of pause.

DOI: https://doi.org/10.21862/diss-09-011-bell-etal

Citation (JIPA): Bellinghausen, Charlotte, Thomas Fangmeier, Bernhard Schröder, Johanna Keller, Susanne Drechsel, Peter Birkholz, Ludger Tebartz van Elst & Andreas Riedel. 2019. On the role of disfluent speech for uncertainty in articulatory speech synthesis. In: R. L. Rose & R. Eklund (eds.), *Proceedings of DiSS 2019, The 9th Workshop on Disfluency in Spontaneous Speech*, 12–13 September, 2019, Budapest, Hungary, 39–42.

The complete proceedings for DiSS 2019 are available as follows.

ISBN: 978-963-489-063-8

DOI: https://doi.org/10.21862/diss-09

DiSS 2019 was sponsored by The Faculty of Humanities, ELTE Eötvös Loránd University and the International Speech Communication Association (ISCA).

On the role of disfluent speech for uncertainty in articulatory speech synthesis

Charlotte Bellinghausen [1], Thomas Fangmeier [2], Bernhard Schröder [1], Johanna Keller [2],
Susanne Drechsel [3], Peter Birkholz [4], Ludger Tebartz van Elst [2] and Andreas Riedel [2]
[1]Institute of German Studies, University of Duisburg-Essen, Essen, Germany
[2]Institute of Psychiatry and Psychotherapy, Medical Center, University of Freiburg, Freiburg, Germany
[3]Department of Speech Science and Phonetics, Martin Luther University Halle-Wittenberg, Halle, Germany
[4]Institute of Acoustics and Speech Communication, TU Dresden, Dresden, Germany

Abstract

In this paper we present a perception study on the role of disfluent speech in forms of prosodic cues of uncertainty in question-answering situations. In our scenario the answer to each question was modeled by varying three prosodic cues: pause, intonation, and hesitation. The utterances were generated by means of an articulatory speech synthesizer. Subjects were asked to rate each answer on a Likert scale with respect to uncertainty, naturalness and understandability. Results showed evidence for an additive principle of the prosodic cues, i.e. the more cues were activated the higher the perceived level of uncertainty. Overall, the effect of intonation and hesitation was more evident than the effect of pause.

Background of the study

The communication of uncertainty

The expression and perception of uncertainty is an essential part in communication (cf. Oh, 2006: 8). In general uncertainty can be regarded as a non-prototypical emotion (Rozin & Cohen, 2003) or as a cognitive state (Kulthau, 1993). We focus on the role of uncertainty in answers following questions. For the acoustic channel several studies suggested evidence that uncertainty is not only expressed, but also perceived by prosodic means like rising intonation, pauses, and hesitations (Smith & Clark, 1993; Brennan & Williams, 1995; Swerts & Krahmer, 2005).

With respect to disfluent speech in acoustic speech synthesis, the synthesis of filled pauses (Adell, Bonafonte & Escudero-Mancebo, 2010) and also of filled pauses and hesitations (Andersson et al., 2010) in Unit Selection Synthesis does not show an increase of naturalness. In Hönemann and Wagner (2016) uncertainty is modelled as one of four emotional states by means of prosodic and voice quality parameters. Furthermore, decreased vocal effort, filled pauses and prolongation of function words contribute to uncertainty perception in synthetic speech using a corpus based-method (Śzekely, Mendelson & Gustafson, 2017).

Articulatory speech synthesis

In our approach we used the articulatory synthesizer VocalTractLab (Birkholz, 2017), which allows to generate high quality speech sounds while manipulating parameters of the time varying laryngeal and supra-laryngeal actions. The synthesizer consists of a geometric 3D model of a male vocal tract (Birkholz, 2013) controlled by 23 parameters to simulate the articulation. The voice source is generated by a self-oscillating model of the vocal folds (Birkholz, Kröger & Neuschaefer-Rube, 2011) which is controlled by six parameters to specify the subglottal pressure, fundamental frequency, and the rest shape of the glottis. The movements of the 3D model and the fundamental frequency are controlled by a gestural score. For each synthetic word the articulatory movements are adjusted manually and generated with different prosodic features.

Previous work

In our previous work (Lasarcyk et al., 2013; Wollermann et al., 2013) we investigated perceived uncertainty by using prosodic cues. The stimuli were question-answer pairs in a human-machine scenario. The answer varied with respect to the combination of the cues pause (absent vs. present), intonation (falling vs. rising) and hesitation (absent vs. present). The experiment design was characterized by three blocks, each time with a 2 × 2 design with pause vs. intonation, intonation vs. hesitation and hesitation vs. pause as independent variables.

261 students of University of Duisburg-Essen (all native speakers of German) listened to the question-answer pairs. They had to rate each time on a 5-point Likert scale how uncertain the answer sounds, how natural, and how understandable it sounded. Results showed in general that the cues of uncertainty were additive with respect to perceived uncertainty.

Perception of uncertainty in ASD

In our interdisciplinary project we investigate the perception of prosodic indicators of uncertainty in Autism Spectrum Disorder (ASD). The aim of our

https://doi.org/10.21862/diss-09-011-bell-etal

current perception study (see below) is to validate the material by presenting it to neurotypically developed subjects. According to diagnostic criteria of DSM-5 (Falkai & Wittchen, 2015) ASD is a neurodevelopmental disease characterized by difficulties in social communication, unusually restricted, repetitive behavior and interests, and specific differences in language and perception. It is mainly accompanied by qualitative deviations in mutual interactions and patterns of communication.

There is an increasing number of studies investigating the role of prosody in ASD. Diehl and Paul (2011) found differences between the perception and imitation of prosodic patterns in children with ASD compared to the control group. Furthermore, in the context of perceiving information status adult listeners with Asperger Syndrome made less use of prosody than the control group, they, however, rely more on lexical information like word frequency and semantic information (Grice, Krüger & Vogeley, 2016).

With respect to emotion perception in articulatory speech synthesis, Hsu and Xsu (2014) showed that high-functioning autistic listeners were less sensitive with respect to emotional prosody by manipulating voice quality as compared to the control group.

Perception study

Goal

The following questions were addressed: Are subjects able to discriminate different intended levels of uncertainty expressed by the three prosodic cues pause, intonation, and hesitation? Is there a correlation between the judgments of uncertainty and the judgments of naturalness as well as of understandability?

Material

Our stimuli were question-answer pairs between a research assistant and a robot for image recognition which were part of a human-machine scenario. The assistant showed pictures of fruits and vegetables to the robot and asked the robot "Was siehst Du?" / *What do you see?* The robot recognized the objects with a certain confidence score, depending on the quality of the picture. Thus, the system was able to express uncertainty about recognition in its answer which was a one-word sentence. As our critical stimuli we chose four one-word trisyllabic phrases in German: "Bananen", "Limetten", "Melonen", "Tomaten" / *bananas*, *limes*, *melons*, and *tomatoes*. For each critical stimulus nine different intended levels of uncertainty were generated (see Table 1).

a) **Pause** refers to the time between the question and the answer. For every level of intended uncertainty we used a default silence pause of 1 s. In the case of pause[+] we used either a silent pause of 4 s or a filled pause, i.e. the hesitation "äh" / uh which took 0.37 s followed by a silent pause of 3.632 s, such that the total duration of the whole pause was 4 s. Since it was not clear from the literature which length of pause is exactly adequate for our research question, we chose an obviously marked pause of 4 s to see whether there is any effect at all on uncertainty perception. b) As **hesitation** particle we chose the particle "äh" / uh since this particle occurred most often for the Verbmobil corpus in German (Batliner et al., 1995). It was either activated (hes[+]) or deactivated (hes[−]). c) The **intonation** of the intended level of certainty showed a peak (measured in semitones) on the stressed syllable of the word with 37 ST. To express uncertainty the last syllable was either characterized by 38 ST for slight uncertainty (level of uncertainty 1) and by 44 ST for strong uncertainty (level of uncertainty 2). Figure 1 shows the different intonation contours for the critical stimulus "Bananen" (each time the question is preceding).

In addition to the critical stimuli, we used nine further one-word phrases as distractors. The utterances were "Birnen", "Blaubeeren", "Bohnen", "Erdbeeren", "Gurken", "Knoblauch", "Mandarinen", "Orangen", and "Paprika" / *pears*, *blueberries*, *beans*, *strawberries*, *cucumbers*, *garlic*, *mandarins*, *oranges*, and *paprika*. The distractors were all characterized by the absence of all three uncertainty cues. We used them in order to distract subjects from the critical stimuli.

Hypothesis

Based on our previous findings (see "Previous work" above) we expected that the prosodic cues of uncertainty have an additive effect, i.e. the activation of all three cues yields a higher degree of perceived uncertainty.

Design

In total we used 36 critical stimuli (4 critical stimuli × 9 different levels of intended uncertainty), 9 stimuli as distractors and one example stimulus. In order to minimize learning effects of the subjects we divided the stimuli into four subsets, each with 19 question-answer pairs in a different random order.

Table 1. Nine different levels of intended uncertainty

pause	hesitation	intonation	level
-	-	-	certainty (c)
-	+	-	hesitation (hes)
+	-	-	pause (pau)
-	-	+	intonation 1 (into 1)
-	-	+	intonation 2 (into 2)
+	+	-	hes & pau
-	+	+	hes & into 2
+	-	+	pau & into
+	+	+	pau & into & hes

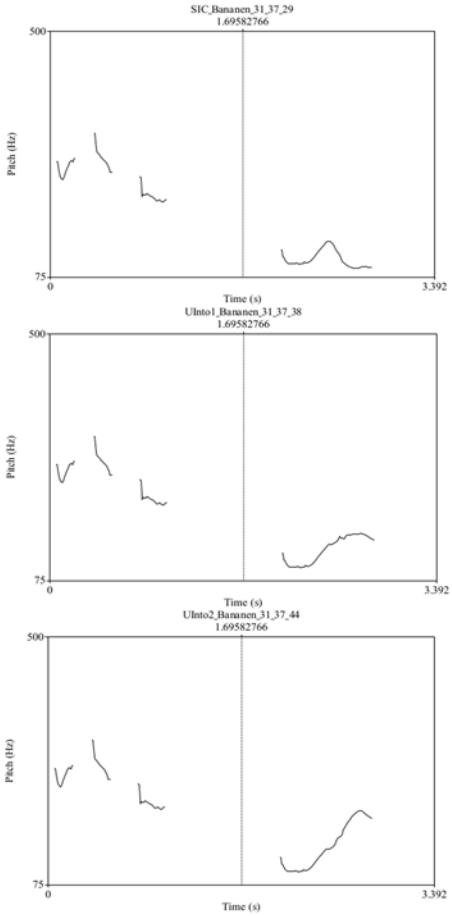

Figure 1. Intonation contour for intended level of a) pattern 31-37-29 for certainty (top), b) pattern 31-37-38 for intonation 1 (middle), c) pattern 31-37-44 for intonation 2 (bottom).

Procedure

Subjects were 36 undergraduate students (23f, 13m; average age: 25 years) of Duisburg-Essen University. All of them were native speakers of German. The number of students per group was as follows: G1: $N = 10$; G2: $N = 7$; G3: $N = 9$; G4: $N = 10$. The procedure started with the presentation of the example stimuli in a seminar room. After each of the 19 question-answer pairs was played subjects had to rate on three 5-point Likert scales a) how uncertain the answer of the robot sounded, b) how natural it sounded, and c) how understandable it was. In addition, subjects had to list the word which they perceived in the answer.

Statistical analysis

For making comparisons between the rankings of the different levels of uncertainty we performed Wilcoxon Matched Pairs Tests with Bonferroni correction. Since we had 30 comparisons our alpha was 0.05/30=0.00167. Furthermore, we tested by means of Spearman Rho Test whether there was a correlation between uncertainty perception and perception of a) naturalness and b) understandability.

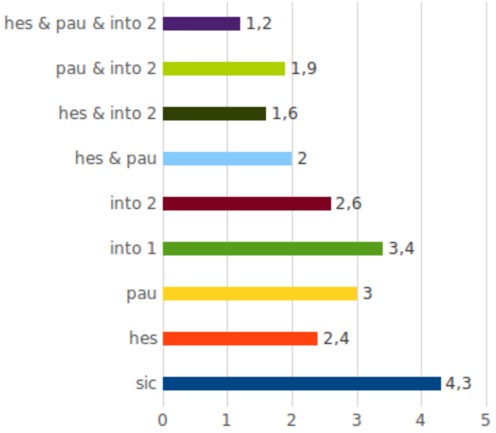

Figure 2. Judgments for perceived uncertainty; x-axis: mean; y-axis: intended level of uncertainty, pau: pause, hes: hesitation, into 2: intonation 2 (see also Figure 1)

Table 2. Significance values of pairwise comparisons using Wilcoxon Matched Paired Test. Significant results with $p < 0.00167$ are marked by x.

	sic	hes	pau	into 1	into 2
hes	x				
pau	x	n.s.			
into 1	x	x	n.s.		
into 2	x	n.s.	n.s.	n.s.	
hes & pau	x	n.s.	x		n.s.
hes & into 2	x	x	x		x
pau & into 2	x	n.s.	x		x
hes & pau & into 2	x	x	x	x	x

	hes & pau & into 2
hes & pau	n.s.
pau & into 2	x
hes & into 2	x

Results

The results for the perceived uncertainty are shown in Figure 2. In Table 2 results for the pairwise comparisons are shown.

The level of intended certainty was ranked significantly different from all other levels of intended uncertainty. When only one prosodic cue was activated the perceived level of uncertainty was always lower in a significant way as opposed to the activation of all three cues. Comparing the activation of a single cue with the activation of two cues, the additional pause combined with hesitation did not contribute significantly to perceived uncertainty. For the other cases, the additional effect was significant. When two activated cues of intended uncertainty are compared to three activated cues the results are as follows: pause and hesitation have an additive effect on the perceived uncertainty in a significant way, but intonation 2 does not. With respect to the comparisons between single cues, our data showed in general no significant differences between judgments except for hesitation vs. intonation 1. For the correlation between the judgment of uncertainty and a) naturalnesss and also of b) understandability we had 10 calculations such that our a was .01/10 = 0.001. The Spearman's Rho Test computed for case a) $p = 0.692$ and for case b) $p = 0.003$. Thus, no significant correlation was found.

Discussion

We presented a study investigating the role of prosodic indicators for uncertainty perception. Results in general suggest an additive effect of prosodic cues and are in line with our previous findings (Lasarcyk et al., 2013; Wollermann et al., 2013). The relative contribution of the pause to uncertainty perception is not clear from our data. For future work we would like to test in a more fine-grained way the effect of pause length. The current study shows that the tested prosodic features of intended uncertainty provide us with a sufficient number of degrees of uncertainty as perceived by neurotypical subjects, so that the stimulus material can be considered suitable to test whether there are differences between neurotypical and autistic hearers.

Acknowledgments

The project is funded by the "Programm zur Förderung des exzellenten wissenschaftlichen Nachwuses" of University of Duisburg-Essen.

References

Adell, J., A. Bonafonte & D. Escudero-Mancebo. 2010. Modelling Filled Pauses Prosody to Synthesize Disfluent Speech. In: *Proceedings of Speech Prosody 2010*, 10–14 May 2010, Chicago, IL, 100624, 1–4.

Andersson, S., K. Georgila, D. Traum, D., M. Aylett & R. A. J. Clark. 2019. Prediction and Realisation of Conversational Characteristics by Utilising Spontaneous Speech. In: *Proceedings of Speech Prosody 2010*, 10–14 May 2010, Chicago, IL, 100116, 1–4.

Batliner, A., A. Kieling, S. Burger, & E. Nöth. 1995. Filled Pauses In Spontaneous Speech. In: *Proceedings of 13th International Congress of Phonetic Sciences (ICPhS)*, 14–19 August 1995, Stockholm, Sweden, vol. 3, 472–475.

Birkholz, P. 2013. Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLoS ONE* 8:e60603. https://doi.org/10.1371/journal.pone.0060603

Birkholz, P. 2017. Vocal Tract Lab (version 2.2). http://www.vocaltractlab.de/ (accessed 03.09.2019).

Birkholz, P., B. J. Kröger, C. Neuschaefer-Rube. 2011. Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis. In: *Proceedings of Interspeech*, 27–31 August 2011, Florence, Italy, 2681–2684.

Brennan, S. E. & M. Williams, M. 1995. The feeling of anothers knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language* 34, 383–398. https://doi.org/10.1006/jmla.1995.1017

Diehl, J. & R. Paul. 2011. Acoustic differences in the imitation of prosodic patterns in children with autism spectrum disorder. *Research in Autism Spectrum Disorders* 6(1): 123–134. https://doi.org/10.1016/j.rasd.2011.03.012

Falkai, P. & H.-U. Wittchen (eds.). 2015. *Diagnostisches und statistisches Manual psychischer Störungen: DSM-5* [Diagnostic and statistical manual of mental disorders: DSM-5]. Göttingen: Hogrefe, 64ff.

Grice, M., M. Krüger & K. Vogeley. 2016. Adults with Asperger syndrome are less sensitive to intonation than control persons when listening to speech, *Culture and Brain* 4(1): 38–50. https://doi.org/10.1007/s40167-016-0035-6

Hönemann, A. & P. Wagner. 2016. *Synthesizing Attitudes in German".* In: *Proceedings of the 16th Speech Science and Technology Conference*, 6–9 December 2016, Sydney, Australia, 209–213.

Hsu, C. & Y. Xu. 2014. Can adolescents with autism perceive emotional prosody? In: *Proceedings of Interspeech 2014*, 14–18 September, Singapore, 1924–1928.

Kuhlthau, C. C. 1993. *Seeking Meaning: A Process Approach to Library and Information Services*, Norwood, NJ: Ablex.

Lasarcyk, E., C. Wollermann, B. Schröder & U. Schade. 2013. On the Modelling of Prosodic Cues in Synthetic Speech – What are the Effects on Perceived Uncertainty and Naturalness? In: *Proceedings of the 10th International Workshop on Natural Language Processing and Cognitive Science*, 15–16 October 2013, Marseille, France, 117–128.

Oh, I. 2006. *Modeling Believable Human-Computer Interaction with an Embodied Conversational Agent: Face-to-Face Communication of Uncertainty*, Rutgers The State University, Dissertation.

Rozin, P. & A. B. Cohen. 2003. High Frequency of Facial Expressions Corresponding to Confusion, Concentration, and Worry in an Analysis of Naturally Occurring Facial Expressions of Americans. *Emotion* 3(1): 68–75. https://doi.org/10.1037/1528-3542.3.1.68

Smith, V. L. & H. H. Clark. 1993. On the Course of Answering Questions. *Journal of Memory and Language* 32(1), 25–38. https://doi.org/10.1006/jmla.1993.1002

Swerts, M. & E. Krahmer. 2005. Audiovisual prosody and feeling of knowing. *Journal of Memory and Language* 53(1), 81–94. https://doi.org/10.1016/j.jml.2005.02.003

Śzekely, E., J. Mendelson & J. Gustafson. 2017. Synthesising uncertainty: The interplay of vocal effort and hesitation disfluencies. In: P*roceedings of Interspeech*, 20–24 August 2017, Stockholm, Sweden, 804–808. https://doi.org/10.21437/Interspeech.2017-1507

Wollermann, C., E. Lasarcyk, U. Schade & B. Schröder. 2013. Disfluencies and Uncertainty Perception – Evidence from a Human-Machine Scenario. In: R. Eklund (ed.): *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech*, 21–23 August 2013, Stockholm, Sweden, 73–76.