Disfluencies and uncertainty perception – evidence from a human–machine scenario

Charlotte Wollermann[1], Eva Lasarcyk[2], Ulrich Schade[3] & Bernhard Schröder[1]

[1] German Linguistics, University of Duisburg-Essen, Germany
[2] Institute of Phonetics, Saarland University, Germany
[3] Fraunhofer Institute for Communication, Information Processing and Ergonomics FKIE, Germany

Abstract

This paper deals with the modelling and perception of disfluencies in articulatory speech synthesis. The stimuli are embedded into short dialogues in question-answering situations in a human–machine scenario. The system is supposed to express uncertainty in the answer. We test the influence of *delay*, *intonation*, and *filler* as prosodic indicators of uncertainty on perception in two studies. Study 1 deals with the effect of *delay* and *filler* on uncertainty perception. Results suggest an additive effect of the cues, i.e. the activation of both prosodic cues of uncertainty has a stronger impact on uncertainty perception than the deactivation of a single cue or of both cues. With respect to the effect of single cues, no significant difference can be observed. Study 2 investigates the impact of *delay* and *intonation* on perceived uncertainty. Again, a principle of additivity can be observed. Furthermore as modelled here, *intonation* has a stronger influence than *delay*. In both studies no correlation between the ranking of uncertainty and naturalness of the stimuli is found.

Index Terms: uncertainty, disfluencies, speech synthesis, speech perception.

1. Introduction

Given is a communicative situation with two conversational partners. A is asking a question to B and B is not certain with respect to her answer. This might be due to several reasons, e.g., B only partially knows the answer, B does not know how to formulate the message, B is grasping for information etc. In addition, uncertainty can be regarded as a complex phenomenon. In some works uncertainty is categorized as emotion [1, 2], in other works it is assumed to have a cognitive character [3]. In the context of question-answering situations, the following questions arise: How do speakers signal uncertainty prosodically with respect to answers? Which prosodic cues do hearers use for decoding uncertainty in answers?

2. Communication of uncertainty

In **human–human communication**, it was found that speakers mark uncertainty in question-answering situations by using *rising intonation*, *pauses*, *fillers*, and *lexical hedges* [4]. In [4] the authors used the *Feeling of Knowing* (FOK) paradigm for eliciting metamemory judgments. For taking the hearer's side into account, [5] defined the *Feeling of Another's Knowing* (FOAK). Results from their perception study provided evidence that the FOAK was influenced by the *intonation* and by the *form* of answers as well as by *pauses* and *fillers*. Furthermore, *smiles* and *funny faces* can serve as visual indicators of uncertainty [6].

In the context of **human–machine communication** however, it is less clear if these cues contribute to the perception of uncertainty in a comparable way. In [7] it is argued that the modelling of uncertainty can improve information systems by enriching expressive abilities. With respect to acoustic speech synthesis, *filled pauses* were modelled in [8] on the basis of a 'synthetic disfluent speech model'. For the implementation an unit selection synthesizer was used. The results of the perception study suggest no decrease of the system's naturalness. In the study of [9], utterances were selected from spontaneous conversational speech. Type and placement of *fillers* and *filled pauses* were predicted using a machine learning algorithm. For the synthesis, an unit selection voice was used. The evaluation shows no decrease of naturalness. Moreover, in [10] it was found that *filled pauses* occur more frequently in the human–machine corpora than in the human–human corpus. However, since it is less clear to what extent disfluencies interact with uncertainty perception, further research seems necessary. For investigating this question we use an articulatory speech synthesizer. A motivation is given in the following section.

3. Paralinguistic expression, prosody and articulatory speech synthesis

Natural speech is usually rich in both linguistic and paralinguistic information. Varying the paralinguistic level of a message to match one's needs can entail speech variations that can for instance involve voice quality, segmental duration, or fundamental frequency changes. Thus, modelling expressive speech can be challenging due to these prosodic variations. However, in order to achieve naturalness of the synthetic speech, variability needs to be considered [11].

To generate such highly variable speech, we use the articulatory synthesis system VocalTractLab [12]. The system produces utterances of high acoustic quality. It processes a timeline of articulatory gestures which are translated into trajectories of the articulators in the virtual three-dimensional vocal tract [13]. In an aerodynamic-acoustic simulation step, the speech signals are generated. Since the utterance is created 'from scratch', the system is very versatile and offers large degrees of freedom. The abovementioned paralinguistic demands on the manner of speaking can be integrated at the foundation of the utterance planning, and no post-hoc signal processing needs to be applied.

4. Previous work

In an intial investigation of modelling and perceiving of uncertainty [14], also the articulatory speech synthesizer of [12] was used. Four different degrees of intended uncertainty were generated by varying the cues *intonation* (rising vs.

falling), *delay* (present vs. absent), and the *filler* 'hmm' (present vs. absent). The scenario was a fictitious telephone dialogue between a weather expert system and a user. The answer of the system was marked by different degrees of uncertainty. Results show that the activation of all uncertainty cues has a stronger impact on the perceived uncertainty than *rising intonation* alone and *delay* combined with *rising intonation*.

In a follow-up study [15], all eight possible combinations of the three cues were used for conveying different degrees of uncertainty, and the stimuli were presented in a modified scenario, an interaction between a robot for image recognition and a user. The user 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. Depending on a fictitious recognition confidence score, the system conveyed (un)certainty in its answer by using the cues mentioned above. Results provide evidence for additivity of all three uncertainty cues with respect to uncertainty perception. Compared to the effect of *rising intonation* and *filler*, the influence of *delay* was relatively weak. The following questions remain open: Does a much longer duration of the *delay* contribute more strongly to the perception of uncertainty? To what extent does the filler 'uh' effect the perception of uncertainty? We address these questions in the current paper. Therefore, we modify the material used in [15].

5. Material

Our stimuli consist of four different one-word phrases in German ('Melonen'/*melons*, 'Bananen'/*bananas*, 'Tomaten'/*tomatoes*, 'Kartoffeln'/*potatoes*) , each one is generated in eight different levels of uncertainty by varying *intonation* (rising vs. falling), *delay* (absent vs. present) and the *filler* 'uh' (absent vs. present). The variation of **intonation** takes place at the last syllable of each word: for *rising intonation* fundamental frequency increases to around 200 Hz, for *falling intonation* it decreases to around 70 Hz. The **delay** refers to the time between the user's question ('Was siehst Du?'/*What do you see?*) and the system's response ('Bananen', 'Tomaten', ...). In each case there is a default *delay* of 1000 ms. In the case of a long *delay* there are two subcases: i) when *filler* is absent the additional *delay* is 4000 ms, ii) when *filler* is present we have the default *delay* (1000 ms) + *filler* 'uh' (duration of 370 ms) + *delay* (3630 ms). For the **filler** we choose the particle 'uh' this time, since 'uh' is the *filler* which occurs most often in the Verbmobil corpus for German [16]. For distracting the subject from our interest we use four further items ('Bohnen'/*beans*, 'Paprika'/*sweet pepper*, 'Gurken'/*cucumber*, 'Knoblauch'/*garlic*).

6. Perception study I

The goal of this study is to test the impact of *filler* and/or *delay* on the perception of uncertainty and naturalness.

Table 1: *Cues of uncertainty. Left: Study I. Right: Study II.*

level	filler	delay	level	intonation	delay
U0	−	−	U0	−	−
U3	−	+	U3	−	+
U4	+	−	U8	+	−
U7	+	+	U11	+	+

6.1. Material and hypothesis

We use four different levels of intended uncertainty (cf. Table 1, left side). To explain the structure of the stimuli we use the example *bananas*:

U0: 'Was siehst Du?' [delay 1000 ms] 'Bananen.'

U3: 'Was siehst Du?' [delay 1000 ms] [delay 4000 ms] 'Bananen.'

U4: 'Was siehst Du?' [delay 1000 ms] ['uh' 370 ms] 'Bananen.'

U7: 'Was siehst Du?' [delay 1000 ms] ['uh' 370 ms] [delay 3630 ms] 'Bananen.'

The stimuli were divided into four sets. In each group we presented eight stimuli: The four items, and the four distractor items. Each stimulus occurred exactly once with respect to the overall data.

We assume that U0 yields to the lowest level of perceived uncertainty, whereas U7 leads to the highest level of perceived uncertainty. Based on our previous studies [14, 15], the following hierarchy is hypothesized: U0<U3<U4<U7.

6.2. Procedure

Subjects were 74 undergraduate students (62 f, 12 m) from the University of Duisburg-Essen, all of them native speakers of German. They were tested in four groups (g1: $N = 25$, g2: $N = 15$, g3: $N = 19$, g4: $N = 16$). For each group one set was presented. The dialogues were played back over loudspeakers. The procedure started with an example stimulus. For each dialogue, subjects had to judge the answer of the system on a questionnaire. There were two 5-point-Likert scales, and subjects were asked to judge how (un)certain the answer sounded and also how natural it sounded.

For statistical analysis, we use different tests. With the Kruskal-Wallis Rank Sum Test we test the overall difference between judgments with respect to uncertainty and naturalness, respectively. In a next step the Wilcoxon Signed Rank with Bonferroni correction is performed for single comparisons between the different levels. Finally, we calculate with the Spearman's Rho Test whether there is a correlation between the uncertainty ratings and the naturalness ratings.[1]

6.3. Results

Firstly, we present the results for the recipients judgments with respect to the perception of **uncertainty**. According to the Kruskal-Wallis Rank Sum Test the overall difference between judgments is $p < 0.0001$ (level of significance: 5%). Figure 1 shows the results in more detail. The Wilcoxon Signed Rank Test with Bonferroni correction (level of significance: $1/6 \times 5\%$) results in $p < 0.0001$ for all comparisons, but there is one exception. The judgments between U3 and U4 do not differ significantly from each other ($p > 0.008$).

In a next step we focus on the **naturalness** ratings. The Kruskal-Wallis Rank Sum Test does not show a difference between judgments when we look at the data overall ($p > 0.05$). For each of the four different levels of uncertainty the median is 3 (cf. Fig. 2). There is no significant difference between the judgments of the single levels as the Wilcoxon Signed Rank Test with Bonferroni correction shows (each time $p > 0.008$).

Furthermore, the Spearman's Rho Test computes a coefficient of 0.05, i.e. our data do not provide evidence for a correlation between the ratings of uncertainty and the ratings of naturalness.

[1] Results of the judgments for the perceived uncertainty exclusively were presented at the Workshop *Fluent Speech* in November 2012 in Utrecht (without publication).

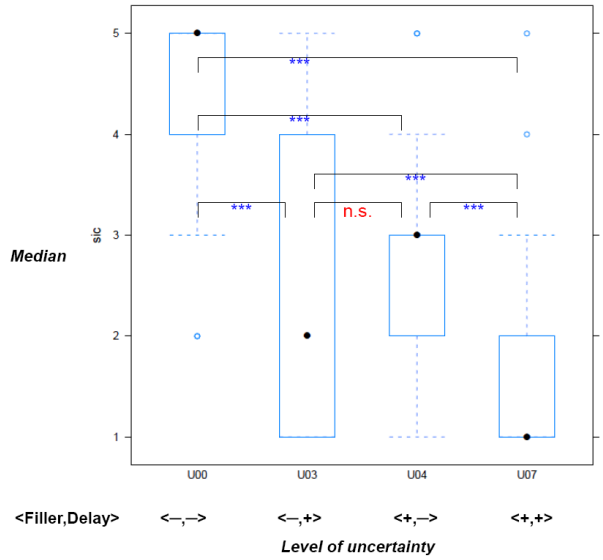

Figure 1: Study I: *Uncertainty judgments;* p < *0.008:*,* p < *0.001:**,* p < *0.0001:****

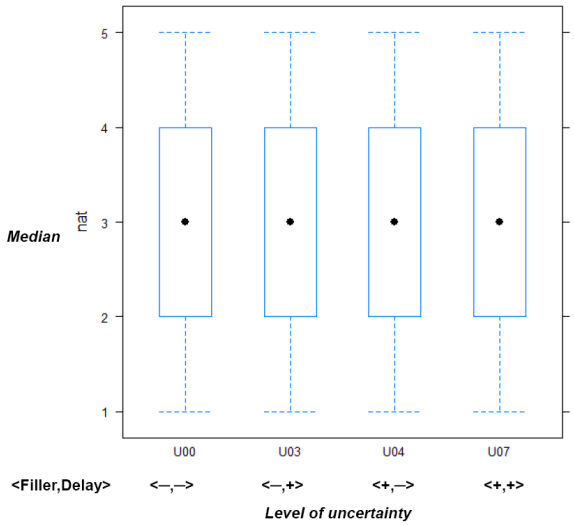

Figure 2: *Study I: Naturalness judgments*

6.4. Discussion

Our results suggest an additive principle of the uncertainty cues. If both cues, i.e. *delay* and *filler*, are activated the perceived uncertainty is significantly higher than for the deactivation of one cue or of both cues. With respect to the relative contribution of the cues, our data show no significant difference between the effect of *delay* vs. *filler*. Moreover, no significancies occur regarding differences in naturalness ratings. Also, no correlation between uncertainty and natural jugments is found.

7. Perception study II

The goal of the second experiment is to investigate the effect of *intonation* and/or *delay* on the perception of uncertainty and naturalness.

7.1. Material and hypothesis

Like in the first study, we use four different levels of intended uncertainty (cf. Table 1, right side). For explaining the structure of the stimuli we use the example *bananas* again:

U0: 'Was siehst Du?'/ [delay 1000 ms] 'Bananen.'

U3: 'Was siehst Du?' [delay 1000 ms] [delay 4000 ms] 'Bananen.'

U8: 'Was siehst Du?' [delay 1000 ms] 'Bananen?'

U11: 'Was siehst Du?' [delay 1000 ms] [delay 4000 ms] 'Bananen?'

As in the previous experiment the stimuli were divided into four sets, each containing eight stimuli (4 items, 4 distractors). Overall, each item stimulus occurred exactly once.

We assume that U0 yields to the lowest level of perceived uncertainty, whereas U11 leads to the highest level of perceived uncertainty. Based on our previous studies [14, 15] the following hierarchy is hypothesized: U0<U3<U8<U11.

7.2. Procedure

Seventy-nine undergraduate students (62 f, 12 m) from the University of Duisburg-Essen took part in the experiment, all of them native speakers of German. They were tested in four groups (g1: $N = 21$, g2: $N = 27$, g3: $N = 15$, g4: $N = 16$). The procedure was the same as in the previous study. For each dialogue, subjects had to judge the answer of the system on a questionnaire with respect to uncertainty and naturalness.

For the statistical analysis we again used the Kruskal-Wallis Rank Sum Test, the Wilcoxon Signed Rank Test with Bonferroni correction, and the Spearman's Rho Test.

7.3. Results

Regarding the perception of **uncertainty** our data reveal (cf. Figure 3) an overall difference between judgments ($p < 0.0001$, Kruskal-Wallis Rank Sum Test). Pairwise comparisons (Wilcoxon Signed Rank Test with Bonferroni correction) show significant differences between judgments for all comparisons ($p < 0.0001$ each time). There is only one comparison with $p < 0.008$ (U3 vs. U8).

Results for the perveiced **naturalness** are illustrated in Figure 4. There are no significant differences between judgments (Kruskal-Wallis Rank Sum Test: $p > 0.05$, Wilcoxon Signed Rank Tests with Bonferroni correction: each time $p > 0.008$). Spearman's Rho Test yields a coefficient of -0.04, indicating no correlation between the uncertainty and naturalness judgments.

7.4. Discussion

In line with the findings of study I, we observe an additivity of the uncertainty cues, but this time for *intonation* and *delay*. Our data also suggest that *rising intonation* alone contributes more strongly to the perception of uncertainty than *delay* alone. With respect to the perception of naturalness our data do not provide evidence for a difference in ratings. In a similar way, no correlation between uncertainty perception and naturalness perception can be observed.

8. Conclusion

We presented two perception studies on the influence of disfluencies in uncertainty perception. The utterances were characterized by different combinations of uncertainty cues and were generated by an articulatory synthesizer. Results show in general significant differences of perceived uncertainty. Our data provide evidence for an additivity of the cues with respect to uncertainty perception. However, we cannot observe an effect of prosodic cues of uncertainty on the perception of naturalness of the synthetic utterances. In addition, no significant correlation could be observed between the judgments of perceived uncertainty and perceived naturalness. In previous studies [8, 9], it was also found that *fillers* and *filled pauses* do not significantly decrease the naturalness of synthetic speech.

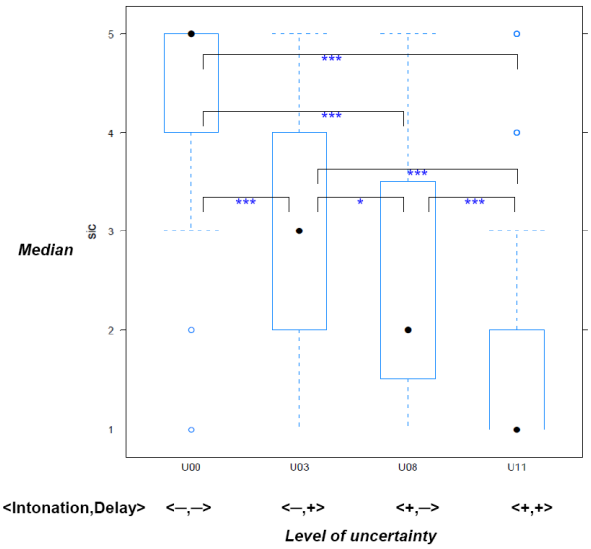

Figure 3: *Study II: Uncertainty judgments;* p < *0.008:*,* p < *0.001:**,* p < *0.0001:****

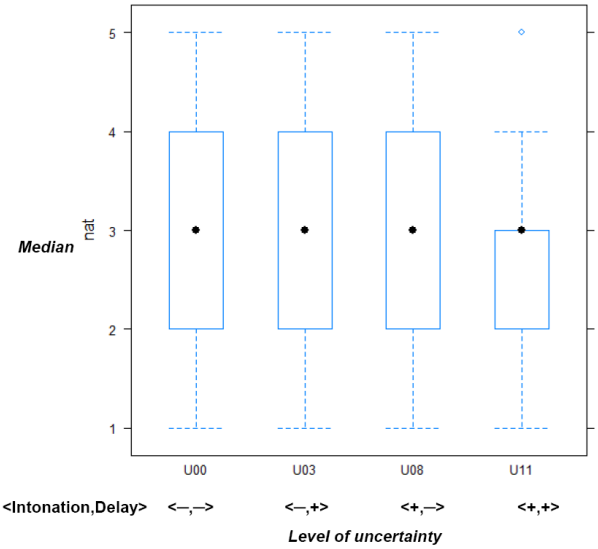

Figure 4: *Study II: Naturalness judgments*

For future work, we deem it necessary to further investigate the interplay between uncertainty perception and naturalness perception. Also, different scenarios need to be considered in order to test the role of disfluencies for the expression of uncertainty and its benefit for human–machine communication.

Many studies have shown that prosody is not only conveyed but also perceived in the visual channel (for synthetic speech e.g. [17, 18]) and the role of visual prosody and uncertainty has been studied for instance in [6, 7, 19]. In our future work we would also like to further investigate *audiovisual* prosody of uncertainty and its interplay with naturalness perception.

9. Acknowledgements

We thank Bernhard Fisseni and Denis Arnold for helpful comments on this paper.

10. References

[1] P. Rozin and A.B. Cohen, "High Frequency of Facial Expressions Corresponding to Confusion, Concentration, and Worry in an Analysis of Naturally Occurring Facial Expressions of Americans". In: *Emotion* 3(1), pp. 68–75, 2003.

[2] D. Keltner and M.N. Shiota, "New Displays and New Emotions: A Commentary on Rozin and Cohen." In: *Emotion* 3(1), pp. 86–91, 2003.

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

[4] V.L. Smith and HH. Clark, "On the Course of Answering Questions". In: *Journal of Memory and Language* 32, pp. 25–38, 1993.

[5] S.E. Brennan and M. Williams, "The feeling of anothers knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers". In: *Journal of Memory and Language* 34, pp. 383–398, 1995.

[6] M. Swerts and E. Krahmer, "Audiovisual prosody and feeling of knowing". In: *Journal of Memory and Language* 53, pp. 81–94, 2005.

[7] E. Marsi and F. van Rooden, "Expressing Uncertainty with a Talking Head in a Multimodal Question-Answering System". In: *Proceedings of the Workshop on Multimodal Output Generation*. Enschede, Netherlands, 105–116, 2007.

[8] J. Adell, A. Bonafonte and D. Escudero-Mancebo, "Modelling Filled Pauses Prosody to Synthesise Disfluent Speech". In: *Proceedings of Speech Prosody 2010*, Chicago, IL, 2010.

[9] S. Andersson, L. Georgila, D. Traum, M. Aylett, R.A.J. Clark, "Prediction and Realisation of Conversational Characteristics by Utilising Spontaneous Speech". In: *Proceedings of Speech Prosody 2010*, Chicago, IL, 2010.

[10] R. Eklund, *Disfluency in Swedish human–human and human–machine travel booking dialogues*. PhD thesis, Linköping Studies in Science and Technology, Dissertation No. 882, Linköping University, Sweden, 2004.

[11] I.R. Murray and J.L. Arnott, "Synthesizing Emotions in Speech: Is it Time to Get Excited?" In: *Proceedings of the International Conference on Spoken Language Processing* 1996, Philadelphia, PA, pp. 1816–1819, 1996.

[12] P. Birkholz, *3D-Artikulatorische Sprachsynthese*, Berlin: Logos, 2006.

[13] P. Birkholz, B.J. Kröger and C.J. Neuschaefer-Rube, "Model-Based Reproduction of Articulatory Trajectories for Consonant-Vowel-Sequence". In: *EEE Transactions on Audio, Speech, and Language Processing*, 19(5), pp. 1422–1433, 2011.

[14] C. Wollermann and E. Lasarcyk, "Modeling and Perceiving of (Un)Certainty in Articulatory Speech Synthesis". In: *Proceedings of the 6th ISCA Workshop on Speech Synthesis*. Bonn, Germany, pp. 40–45, 2007.

[15] E. Lasarcyk and C. Wollermann, "Do prosodic cues influence uncertainty perception in articulatory speech synthesis?" In: *Proceedings of the 7th ISCA Workshop on Speech Synthesis*. Kyoto, Japan, pp. 230–235, 2010.

[16] A. Batliner, A. Kießling, S. Burger and E. Nöth, "Filled Pauses In Spontaneous Speech". In: *Proceedings of 13th Intl. Congress of Phonetic Sciences* 3, pp. 472–475, 1995.

[17] E. Krahmer, Z. Ruttkay, M. Swerts and W. Wesselink, "Pitch, Eyebrows and the Perception of Focus". In: *Proceedings of Speech Prosody* 2002. Aix-en-Provence, France, pp. 443–446, 2002.

[18] B. Granström and D. House, "Inside out – acoustic and visual aspects of verbal and non-verbal communication". In: *Proceedings of the 16th International Congress of Phonetic Sciences* 2007, Saarbrücken, Germany, pp. 11–18, 2007.

[19] I. Oh, *Modeling Believable Human–Computer Interaction with an Embodied Conversational Agent: Face-to-Face Communication of Uncertainty*. PhD thesis, Rutgers The State University of New Jersey, NJ, 2006.