In: Proceedings 10th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2013), Marseille, France, October 2013.

On the Modelling of Prosodic Cues in Synthetic Speech – What are the Effects on Perceived Uncertainty and Naturalness?

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

[1]Institute of Phonetics, Saarland University, Germany
evaly@coli.uni-saarland.de
[2]German Linguistics, University of Duisburg-Essen, Germany
{charlotte.wollermann, bernhard.schroeder}@uni-due.de
[3]Fraunhofer Institute for Communication, Information Processing and Ergonomics FKIE, Germany
ulrich.schade@fkie.fraunhofer.de

Abstract

In this paper we present work on the modelling of uncertainty by means of prosodic cues in an articulatory speech synthesizer. Our stimuli are embedded into short dialogues in question-answering situations in a human-machine scenario. The answers of the robot vary with respect to the intended level of (un)certainty, the independent variables are *intonation* (rising vs. falling) and *filler* (absent vs. present). We perform a perception study in order to test the relative impact of the prosodic cues of uncertainty on the perception of uncertainty and also of naturalness. Our data indicate that the cues of uncertainty are additive. If both prosodic cues of uncertainty are present, the perceived level of uncertainty is higher as opposed to the deactivation of a single cue. Regarding the relative contribution of *intonation* vs. *filler* our results do not show a significant difference between judgments. Moreover, the correlation between the judgment of uncertainty and of naturalness is not significant.

1 Introduction

The general topic of this paper is the role of uncertainty in question-answering situations. Suppose a communicative situation with two conversational partners. A asks B a question, and B is not certain with respect to her answer. Why is B uncertain in this situation? There might be several reasons: i) B only partially knows the answer,

ii) B cannot judge what the listener already knows, iii) B does not know how to formulate the message etc. For a detailed presentation of the process of language production in general and its possible troubles cf. Levelt (1989).

In addition, uncertainty can be regarded as a complex phenomenon. In some works uncertainty is categorized as emotion (Rozin, Cohen, 2003; Keltner, Shiota, 2003), in other works it is assumed to have a cognitive character (Kuhltau, 1993). In the context of question-answering situations, the following questions arise: Which prosodic cues do speakers use for encoding uncertainty in answers? Which prosodic cues contribute to the perception of uncertainty?

2 Communication of uncertainty

In this section we firstly discuss the role of uncertainty in human-human communication (Section 2.1). Afterwards we refer to previous studies on the role of uncertainty in human-machine communication (Section 2.2). In Section 2.3, we give a general motivation for investigating uncertainty as an expressive ability of machines.

2.1 Uncertainty in human-human communication

In human-human communication, conversation partners use several prosodic cues in order to signal and also to perceive uncertainty in answers. With respect to speech production in the study of Smith and Clark (1993), metamemory judgments in question-answering situation were elicited by using the *Feeling of Knowing* (FOK) paradigm. Results suggest that speakers mark uncertainty by using *rising intonation*, *pauses*, *fillers* and *lexical hedges*. For investigating the hearer's side as well, in Brennan and Williams (1995) the *Feeling of Another's Knowing* (FOAK) was defined. Results from their perception study show for the acoustic channel that the *intonation*, the *form* of answers, *pauses* and also *fillers* effect the FOAK.

Furthermore, *fillers* and *pauses* have been found as relevant cues with respect to self-repair in speech, especially to those self-repairs that do not contain lexical material (coined *c-repairs*) (Goldman-Eisler, 1967; Levelt, 1983). These repairs occur if the speaker recognises and corrects the slip of the tongue even before a speech signal is produced. A connectionist model of such a kind of repairs can be found in Schade and Eikmeyer (1991).

Swerts and Krahmer (2005) replicated the study of Smith and Clark (1993) and extended the design to the visual aspect. For the audio channel, *delay*, *pauses* and *fillers* were found as being relevant for marking uncertainty; for the visual modality, *smiles* and *funny faces*. In order to test the relevance of these cues for speech perception, audio-only, visual-only, and audiovisual stimuli were presented to subjects and had to be judged with respect to uncertainty. Results suggest that subjects were able to distinguish certain from uncertain utterances in all three conditions, but identification was easier in the bimodal condition than in the unimodal conditions.

Also with respect to audiovisual cues of uncertainty, Borràs-Comes et al. (2011) tested the relative contribution of *facial gestures*, *intonation* and *lexical choice* on uncertainty perception. Results suggest that all three cues have a significant effect on

perceived uncertainty. Furthermore, in the case of a mismatch between *gesture* and *intonation*, *gesture* has a stronger impact.

2.2 Uncertainty in human-machine communication

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. Marsi and van Rooden (2007) argue that the modelling of uncertainty can improve information systems by enriching expressive abilities. With respect to acoustic speech synthesis, Adell et al. (2010) modelled *filled pauses* on the basis of a 'synthetic disfluent speech model'. For these purposes an unit-selection synthesizer was used. In a next step a perception study was performed in order to test whether *filled pauses* can be generated without decreasing the system's quality. The results show no significant decrease of the system's naturalness. In the study of Andersson et al. (2010) utterances were selected from spontaneous conversational speech. The goal was to generate *fillers* without affecting the system's naturalness in a negative way. By using a machine-learning algorithm, type and placement of *fillers* and of *filled pauses* were predicted. Again, a unit-selection voice was used. Similar to the findings of Adell et al. (2010), no significant decrease of naturalness was observed during the evaluation.

In addition, the role of uncertainty in human-machine communication has also been investigated with respect to visual speech synthesis. The results of Oh (2006) suggest that the variation of *facial expressions* and *head movements* affects the recognition of uncertainty. According to Marsi and van Rooden (2007) *head movement* alone, and also combined with *eyebrow movement*, affects the perception of uncertainty as well.

The automatic detection of uncertainty in utterances by dialogue systems is for instance useful for systems that function as tutors. The study of Pon-Barry et al. (2006) suggests that the learning process of the student can be affected positively if the system adapts to the student's uncertainty. For training these systems, corpora consisting of natural conversations between tutors and students are often used. Uncertain utterances have been detected with an accuracy of ca. 75% by the usage of prosodic cues covering *fundamental frequency*, *intensity*, *tempo* and *duration* (Liscombe et al., 2005; Pon-Barry, Shieber, 2009).

2.3 Motivation

As already mentioned in the previous section, the modelling of uncertainty can be useful to create systems with expressive abilities (Marsi, van Rooden, 2007). Why is it useful to have systems equipped with those abilities? Natural language is characterized by a high degree of variability (Murray, Arnott, 1996). Speech does not only differ from speaker to speaker, but also within an individual speaker. This variability is caused by different factors, e.g. by speaking style and by emotion and mood (cf. Murray, Arnott, 1996). If one aims to develop speech synthesis systems with an as natural as possible speech output, this variability needs to be taken into account. We regard the expression of uncertainty as one factor which can contribute to the variability of synthetic speech.

Moreover, we are interested in simulating uncertainty as a human meta-cognitive state by an artificial system which is able to express this uncertainty in the synthetic

signal. Also, we would like to investigate whether human listeners ascribe this meta-cognitive state to the machine.

In our work we model different degrees of uncertainty by means of prosodic cues, using an articulatory speech synthesizer to generate the utterances. A motivation is given in the following section. We perform a perception study to test to what extent the intended uncertainty indeed affects speech perception.

3 Articulatory speech synthesis

To generate the highly variable speech, we use the articulatory synthesis system VocalTractLab (Birkholz, 2006). The system produces utterances of high acoustic quality. It processes a timeline of articulatory gestures which are translated into trajectories of speech articulators in a virtual three-dimensional vocal tract (Birkholz et al., 2011). In an aerodynamic-acoustic simulation step, the speech signals are generated. Since each utterance is created 'from scratch', the system is very versatile and offers large degrees of freedom for variation. The prosodic 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 Related work

An initial investigation on the modelling and perception of uncertainty using the articulatory speech synthesizer by Birkholz (2006) was presented in Wollermann and Lasarcyk (2007). 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 (Lasarcyk, Wollermann, 2010), all eight possible combinations of the three cues were used for conveying different degrees of uncertainty. Moreover, 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 effects of *rising intonation* and *filler*, the influence of *delay* was relatively weak.

From our findings we infer the following questions which need to be further investigated: i) Does a much longer duration of the cue *delay* contribute more strongly to the perception of uncertainty? ii) To what extent does the filler 'uh' affect the perception of uncertainty? iii) Does the expression of uncertainty influence the naturalness of the synthetic utterances? We address these questions in the current paper. To do this, we modify the speech material used in Lasarcyk and Wollermann (2010).

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 in 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 apply 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 (Batliner et al., 1995).

To distract the subjects from our interest we use four distractor items ('Bohnen'/*beans*, 'Paprika'/*sweet pepper*, 'Gurken'/*cucumber*, 'Knoblauch'/*garlic*). To generate the distractor items, we use *falling intonation*, default *delay*, and no *filler*. By using the distractor items it should be precluded that the subjects' linguistic awareness is focused on the tested question.

6 Experimental design

Our overall experimental design consists of three experimental blocks. In each experimental block we vary two of our three prosodic factors. In the first block we test the relative contribution of *filler* vs. *delay* on the perception of uncertainty (cf. Table 1, left side). In the second block we investigate the influence of *intonation* vs. *delay* (cf. Table 1, middle). In the last block, the relative impact of *intonation* vs. *filler* is tested (cf. Table 1, right side). Furthermore, in all three cases we calculate whether there is a correlation between the perception of uncertainty and the perception of naturalness.

The results of block I and II are described in detail in Wollermann et al. (2013) and will only briefly be summarized here. In block I, the stimuli were presented to 74 subjects. They rated the degree of uncertainty and naturalness of each stimulus on 5-point Likert scales. Results suggest an effect of additivity of the uncertainty cues. If both *filler* and *delay* are present, the level of perceived uncertainty is higher as opposed to when one of the cues is deactivated. Furthermore, there was no significant difference between the effects of *filler* and *delay* – as single cues – on the perceived level of uncertainty. Moreover, our data do not suggest evidence for a correlation of uncertainty ratings and naturalness ratings in a significant way.

In block II, the stimuli were evaluated by 79 participants. Similar to block I, a principle of additivity can be observed since *rising intonation* combined with *delay* has a stronger impact on the perceived uncertainty than *rising intonation* alone or *delay* alone. When comparing the effects of the single cues against each other, our data indicate that *rising intonation* yields a stronger level of perceived uncertainty than *delay*. Again, no significant correlation between the perception of uncertainty and naturalness is found.

Table 1: Cues of uncertainty. Left: Block I. Middle: Block II. Right: Block III.

Level	Filler	Delay	Level	Intonation	Delay	Level	Intonation	Filler
C	–	–	C	–	–	C	–	–
U3	–	+	U3	–	+	U4	–	+
U4	+	–	U8	+	–	U8	+	–
U7	+	+	U11	+	+	U12	+	+

Table 2: Ordering of the stimuli (highlighted in yellow), as presented in the perception test groups 1 to 4. Positions 2, 3, 6, and 7 are filled with distractor items.

Position	Group 1	Group 2	Group 3	Group 4
1	C-Kartoffeln	U4-Bananen	U8-Tomaten	U12-Melonen
2	C-Bohnen	C-Knoblauch	C-Paprika	C-Gurken
3	C-Gurken	C-Paprika	C-Bohnen	C-Knoblauch
4	U8-Bananen	U12-Kartoffeln	C-Melonen	U4-Tomaten
5	U12-Tomaten	U8-Melonen	U4-Kartoffeln	C-Bananen
6	C-Knoblauch	C-Bohnen	C-Gurken	C-Paprika
7	C-Paprika	C-Gurken	C-Knoblauch	C-Bohnen
8	U4-Melonen	C-Tomaten	U12-Bananen	U8-Kartoffeln

7 Perception study

In the following section we present the experimental design, the procedure and the results of block III. The goal of this study is to test the impact of *intonation* and/or *filler* on the perception of uncertainty and naturalness.

7.1 Material and hypothesis

We used the four different levels of intended (un)certainty shown in Table 1, right side.[1] To illustrate the structure of the stimuli, the simulated interactions between the human and the machine concerning *bananas* are listed below. A question mark at the end of a phrase indicates rising intonation.

C: Human: *'Was siehst Du?'* (What do you see?) – Machine: [delay 1 s] *'Bananen.'* (Bananas.)
U4: Human: *'Was siehst Du?'* – Machine: [delay 1 s] ['uh' 370 ms] *'Bananen.'*
U8: Human: *'Was siehst Du?'* – Machine: [delay 1 s] *'Bananen?'*
U12: Human: *'Was siehst Du?'* – Machine: [delay 1 s] ['uh' 370 ms] *'Bananen?'*

The stimuli were divided into four sets, as shown in Table 2. In each group we presented eight stimuli: four items and the four distractor items. Each stimulus occurred exactly once with respect to the overall data.

We assume that prosodic indicators of uncertainty have an additive effect with respect to uncertainty perception, i.e. the more uncertainty cues are activated, the higher the level of perceived uncertainty. Our detailed assumption is as follows: C will receive,

[1] We plan to test more than these four levels of uncertainty. To make the current stimuli comparable to future experiments, the coding of the levels is not done using a straight count.

relatively to the other levels, the highest rating of perceived certainty. U4, U8, and U12 are intended levels of uncertainty. We expect that U12 will lead to the highest rating of perceived *un*certainty. We further assume that U4 and U8 will be rated between C and U12.

Our goal is to model different levels of intended uncertainty which are closely connected to a relatively high level of naturalness. We refer to naturalness as *relatively* high because we assume that the human listeners will identify the artificial nature of the synthesized speech (as opposed to the human speech) and thus will not ascribe absolute naturalness to the system's utterances. Our expectation is as follows: If we are able to model uncertainty by prosodic cues without decreasing the naturalness of the system, the prosodic cues are adequate to trigger different degrees of intended uncertainty. Therefore we expect no significant correlation between perceived naturalness and perceived uncertainty.

7.2 Procedure

108 undergraduate students (82 f, 26 m) from the University of Duisburg-Essen took part in the perception study. All of them were native speakers of German. The subjects were tested in four groups (g1: N=25, g2: N=17, g3: N=31, g4: N=35). In each group a subset of the stimuli was presented and also a different order of the items was used to neutralise the impact of learning effects.

The dialogues consisted of the question-answer pairs described in the previous section and were played back over loudspeakers. The procedure started with an example stimulus. For each dialogue, subjects were instructed to judge the answer of the system on a questionnaire, using two 5-point Likert scales to indicate how (un)certain the answer sounded and also how natural it sounded (5=certain, 1=uncertain; 5=natural, 1=unnatural).

For statistical analysis, we firstly test the overall difference between judgments with respect to uncertainty and naturalness, respectively, using the Kruskal-Wallis Rank Sum Test. Secondly, we perform the Wilcoxon Signed Rank test with Bonferroni correction to calculate single comparisons between the different levels. Finally, we use Spearman's Rho Test to test if there is a correlation between the uncertainty ratings and the naturalness ratings.[2]

8 Results

In the following we present the results of the perception of uncertainty (Section 8.1) and of the perception of naturalness (Section 8.2).

8.1 Uncertainty

The Kruskal-Wallis Rank Sum Test indicates that the overall difference between uncertainty judgments is highly significant ($p<0.0001$, level of significance: 5%). Figure 1

[2]Results of perceived uncertainty alone were presented at the Workshop of the *Scandinavian Association for Language and Cognition* in June 2013 in Joensuu, Finland (without publication).

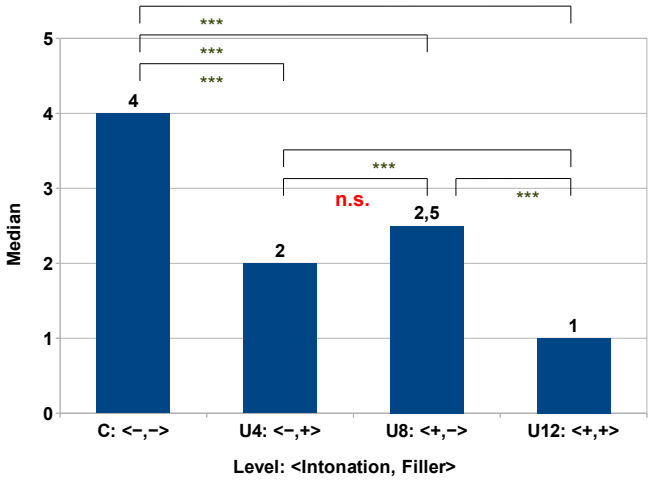

Figure 1: Clustered data – uncertainty judgments; $p<0.008$:*, $p<0.001$:**, $p<0.0001$:***

shows the results for the clustered data, i.e. aggregated for all four stimuli of each level of uncertainty. The Wilcoxon Signed Rank Test with Bonferroni correction (level of significance: 1/6 x 5%) results in $p<0.0001$ for all comparisons, except for the comparison U4 vs. U8. In the latter case there is no significant difference between judgments ($p>0.008$).

In a next step, we analyse the judgments for the individual stimuli. The results are illustrated in Figure 2. For all four wordings, the following comparisons show a significant difference between judgments: C vs. U4, C vs. U8, C vs. U12, and U8 vs. U12. The levels U4 (*filler* activated individually) vs. U8 (*intonation* activated individually) are never rated significantly differently. U4 vs. U12 only shows a significant difference for *Bananen* und *Tomaten*, but not for *Kartoffeln* and *Melonen*.

8.2 Naturalness

For naturalness, the Kruskal-Wallis Rank Sum Test does not show a significant difference between judgments when we look at the data overall ($p>0.05$). It can be observed that each of the four different levels of (un)certainty is judged with a median of 4 (cf. Figure 3). The Wilcoxon Signed Rank Test with Bonferroni correction indicates for each of the six inter-level comparisons that judgments do not differ significantly from each other ($p>0.008$ in all cases). Regarding a possible correlation of the ratings of uncertainty and naturalness, the Spearman's Rho Test results in a correlation coefficient of -0.11 ($p>0.05$). Thus, as expected, our data do not suggest evidence for a correlation in a significant way.

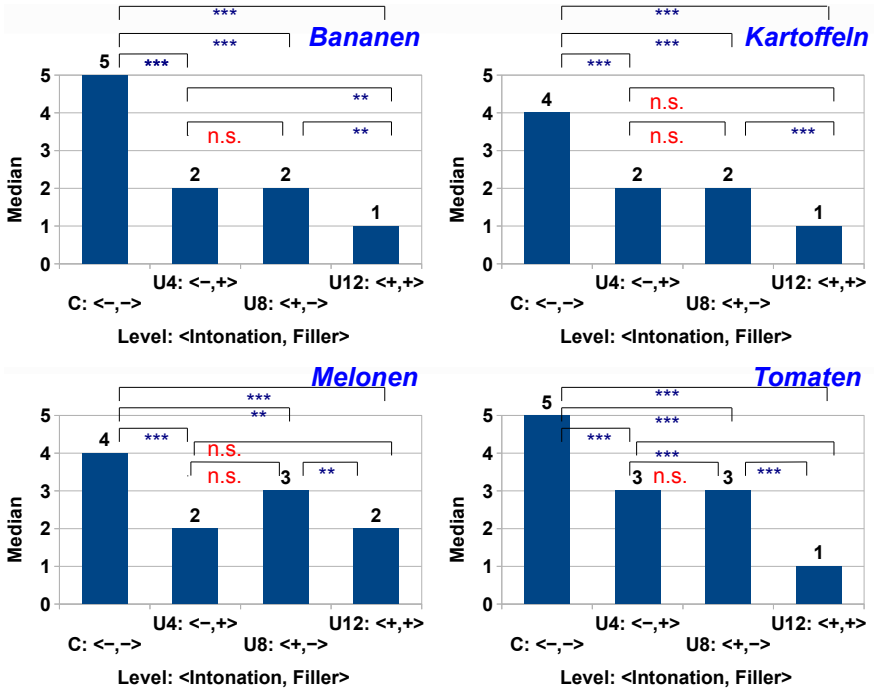

Figure 2: Individual stimuli – uncertainty judgments; $p<0.008$:*, $p<0.001$:**, $p<0.0001$:***

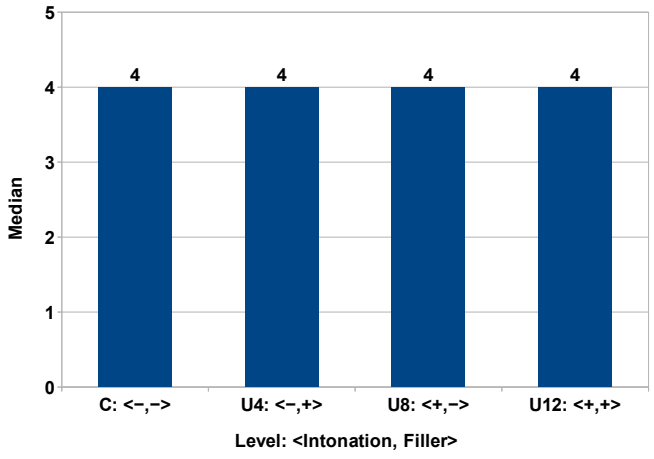

Figure 3: Clustered data – naturalness judgments

9 Discussion

In this paper we presented a study on the modelling of uncertainty by prosodic cues in articulatory speech synthesis. We varied *intonation* and *filler* and tested the relative impact of these cues on perceived uncertainty and perceived naturalness. Regarding uncertainty, the results of the experiment suggest that the cues are additive, i.e. the more uncertainty cues are activated the higher the perceived level of uncertainty. However, our data do not suggest evidence for a stronger effect of *filler* or *intonation* since the subjects' judgments do not differ significantly when these cues are activated individually (U4 vs. U8).

With respect to the perception of naturalness, we do not observe a significant effect of the prosodic cues. In a similar way, the correlation between perceived uncertainty and perceived naturalness is not significant. This result is in line with our assumptions because it indicates that *filler* and *delay* increase the perceived level of uncertainty but do not reduce naturalness. If that were the case, it would be problematic since it could indicate that listeners perceived high levels of uncertainty due to low naturalness, and not due to prosodic variation.

We conclude that different degrees of uncertainty can be expressed by the variation of prosodic cues. As modelled here, varying prosody neither increases nor decreases the naturalness of the utterances. Thus, we assume that – for our scenario – listeners decode uncertainty in the answers of the system and ascribe a meta-cognitive state to the machine.

For future work, we regard it as important to evaluate for different scenarios whether the modelling of uncertainty is a benefit for human-machine communication. Also, we would like to take into account the visual aspect of speech. In several studies, visual prosodic cues have been synthesized (e.g. Krahmer et al., 2002; Granström, House, 2007), and uncertainty in particular has been modelled by means of audiovisual prosody (Oh 2006; Marsi, van Rooden, 2007). We would like to further investigate the interplay between audio and visual prosody and its relevance for perceived uncertainty.

10 Acknowledgments

Many thanks to Denis Arnold and Bernhard Fisseni for helpful comments.

References

[1] Adell, J., Bonafonte, A., Escudero-Mancebo, D. (2010) Modelling Filled Pauses Prosody to Synthesise Disfluent Speech. In: *Proceedings of Speech Prosody 2010*, Chicago, IL, pp. 100624:1-4.

[2] Andersson, S., Georgila, K., Traum, D., Aylett, M., Clark, R. A. J. (2010) Prediction and Realisation of Conversational Characteristics by Utilising Spontaneous Speech. In: *Proceedings of Speech Prosody 2010*, Chicago, IL, pp. 100116:1-4.

[3] Batliner, A., Kieling, A., Burger, S., Nöth, E. (1995) Filled Pauses in Spontaneous Speech. In: *Proceedings of 13th International Congress of Phonetic Sciences*, 3, Stockholm, Sweden, pp. 472-475.

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

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

[6] Borràs-Comes, J., C., Roseano, P., del Mar Vanrell, M., Chen, A., Pietro, P. (2011) Perceiving uncertainty: facial gestures, intonation, and lexical choice. In: *Proceedings of the Workshop on Gesture and Speech in Interaction*. Bielefeld, Germany.

[7] Brennan, S.E., Williams, M. (1995) 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.

[8] Goldman-Eisler, F. (1967) Sequential temporal patterns and cognitive processes in speech. In: *Language and speech* 10(2), pp. 122-132.

[9] Granström, B., House, D. (2007) 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.

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

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

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

[13] Lasarcyk, E., Wollermann, C. (2010) 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.

[14] Levelt, W.J.M. (1983) Monitoring and self-repair in speech. In: *Cognition* 14, pp. 41-104.

[15] Levelt, W.J.M. (1989) *Speaking: From Intention to Articulation*. Cambridge: MIT Press.

[16] Liscombe, J., Hirschberg, J., Venditti, J.J. (2005) Detecting certainness in spoken tutorial dialogues. In: *Proceedings of Interspeech 2005*. Lisboa, Portugal, pp. 1837-1840.

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

[18] Murray, I.R., Arnott, J.L. (1996) 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.

[19] Oh, I. (2006) Modeling Believable Human-Computer Interaction with an Embodied Conversational Agent: Face-to-Face Communication of Uncertainty. PhD thesis, Rutgers University, New Brunswick, NJ, USA.

[20] Pon-Barry, H., Schultz, K., Bratt, E.O., Clark, B., Peters, S. (2006) Responding to Student Uncertainty in Spoken Tutorial Dialogue Systems. In: *International Journal of Artificial Intelligence in Education* 16(2), pp. 171-194.

[21] Pon-Barry, H., Shieber, S. (2009). The Importance of Subutterance Prosody in Predicting Level of Certainty. In: *Proceedings of the Companion Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL) 2009*. Boulder, CO, pp. 105-108.

[22] Rozin, P., Cohen, A.B. (2003) 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.

[23] Schade, U., Eikmeyer, H.-J. (1991) "wahrscheinlich sind meine Beispiele soo sprunghaft und und und eh ehm zu zu telegraph" – Konnektionistische Modellierung von "covert repairs". In: Th. Christaller (ed.): *GWAI-91 1. Fachtagung für Künstliche Intelligenz*. Berlin: Springer Verlag, pp. 264-273.

[24] Smith, V.L., Clark, H.H. (1993) On the Course of Answering Questions. In: *Journal of Memory and Language*, 32, pp. 25-38.

[25] Swerts, M., Krahmer, E. (2005) Audiovisual prosody and feeling of knowing. In: *Journal of Memory and Language*, 53, pp. 81-94.

[26] Wollermann, C., Lasarcyk, E. (2007) 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.

[27] Wollermann, C., Lasarcyk, E., Schade, U., Schröder (2013) Disfluencies and Uncertainty Perception Evidence from a Human-Machine Scenario. In: *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech*. Stockholm, Sweden, pp. 73-76.