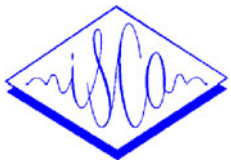
ISCA Archive
http://www.isca-speech.org/archive



7th ISCA Workshop on
Speech Synthesis (SSW-7)
Kyoto, Japan
September 22-24, 2010

Do Prosodic Cues Influence Uncertainty Perception in Articulatory Speech Synthesis?

Eva Lasarcyk[1], *Charlotte Wollermann*[2,3]

[1]Institute of Phonetics, Saarland University, Germany
[2]Institute of Communication Sciences, University of Bonn, Germany
[3]German Linguistics, University of Duisburg-Essen, Germany
evaly@coli.uni-saarland.de, cwo@ifk.uni-bonn.de

Abstract

This study investigates the individual influences of the three prosodic cues *response delay*, the *filler* "hmm", and *rising intonation* on the perception of uncertainty within a fictitious human-computer dialogue. Response delay is the time that the computer waits until it starts to answer a given question. The filler, i.e. the hesitation particle "hmm", can be inserted before the content of the answer starts. The final part of the answer's intonation contour can rise or fall. We hypothesize a hierarchy of influence: Rising intonation has a stronger influence on uncertainty perception than filler; response delay has the weakest effect. In a perception study the uncertainty of utterances generated with articulatory speech synthesis was tested. Results indicate that all cues have an effect on the perception of uncertainty, but the relative impact differs: Delay has a rather weak effect, whereas rising intonation and filler seem to be uncertainty-enhancing acoustic cues, each having strong effects which seem to override the weaker cue of delay. The results can serve as guideline for automatic detection of uncertainty in spoken dialogue systems.

Index Terms: uncertainty, prosody, paralinguistic expression, articulatory speech synthesis

1 Introduction

When people talk to each other, communication is very efficient because we do not only convey a linguistic message, i.e. the words, but also a high amount of paralinguistic information. When asking someone a question, the "factual" answer is e.g. accompanied by information on he speaker's attitude towards the conversational partner and the subject of conversation itself. Does strategies that we use in human-human interaction also work in human-computer dialogues? We aim at exploring this question with regard to the paralinguistic aspect of *uncertainty*: You are expected to answer a question but you are not entirely sure whether your answer is correct – and use paralinguistic means to convey this uncertainty.

The goal of this study is to model different degrees of uncertainty by combining the three acoustic cues *rising intonation*, *delay* and *filler*. We assume a hypothetical scale of uncertainty which is characterized by eight different levels, obtained by the permutations of the three acoustic cues. The scale is evaluated empirically by a perception test. The stimuli are generated with the articulatory speech synthesizer developed by [1].

2 Theoretical background on uncertainty

In this section we first discuss the role of prosodic cues for the production and perception of uncertainty in natural speech (Sec. 2.1), present previous studies on the role of uncertainty in spoken dialogue systems (Sec. 2.2) and refer to the modelling of uncertainty in speech synthesis (Sec. 2.3).

2.1 Production and perception of uncertainty in natural speech

Speakers and listeners use various cues in communication to signal and detect uncertainty. Uncertainty is often characterized as a "non-prototypical" emotive state [2] and/or as a cognitive state [3]. The work of [4] served as source of inspiration for many studies in this field. The authors investigated memory processes in question-answering situations. To test the hypothesis that speakers mark certainty differently from uncertainty, the *Feeling of Knowing* (FOK) paradigm according to [5] was used. With this method, it is possible to elicit metamemory judgements. Their experimental investigation showed that speakers express uncertainty on the lexical level by using phrases like *I guess*, and also by means of prosodic cues like *rising intonation* and *delay*.

In order to test how listeners perceive the FOK of a speaker, [6] defined the *Feeling of Another's Knowing* (FOAK) paradigm. It was shown that the FOAK "[...] was affected by the intonation of answers, the form of answers. […] the latency to response, and the presence of fillers." ([6]: 396). The term *filler* referred to interjections, e.g. *hmm*, *um* and *uh* (cf. [6]: 383). Similar effects were found in [7].

In addition, our empirical study [8] suggested evidence that *rising* intonation as prosodic cue of uncertainty influence the interpretation of pragmatic focus. It was found that the marking of the focus constituent by *falling* intonation combined with a question which is *parallel* to the supposed focus structure favours exhaustive interpretation of answers, whereas *rising* intonation with an *incongruent* question favours non-exhaustivity. In our follow-up study [9] we used *rising* intonation for the focus constituent and sentence-final verb in addition with *pauses* and variation of *macro-context*. The intended certain way of speaking combined with a macro context excluding focus alternatives contributed to exhaustivity; in contrast, the intended uncertain way of speaking in combination with a macro context including focus alternatives reinforced non-exhaustivity. From these empirical data we derived a model of pragmatic focus interpretation [10].

2.2 Automatic detection of uncertainty

With respect to spoken dialogue systems the automatic detection of uncertainty from natural speech plays an important role since uncertainty is a phenomenon which often occurs when users interact with dialogue systems functioning as tutors: Students' production of uncertainty is often observable during computer tutoring (cf. [11]). In the study of [12] uncertainty was detected by using the corpus ITSPOKE (Intelligent Tutoring Spoken Dialogue System) [13]. For theses purposes a combination of acoustic-prosodic features

extracted at two levels of intonational analysis, i.e. breath groups and turn, were used. Results show a classification accuracy of 76%, i.e. a 16% relative improvement over baseline performance. Similar results were reported by [14] who showed that using prosodic features of the word or phrase responsible for the level of certainty and of its surrounding context improves the prediction accuracy when compared to using only features taken from the utterance as a whole. Furthermore, it was suggested that the adaptation of a dialogue system to the students uncertainty increases the tutors effectiveness [15]. While the scope of our study does not cover automatic detection itself, the detailed analysis of our set of acoustic cues can be used to improve such systems.

2.3 Modelling uncertainty

In [16] an approach is presented for synthesizing filled pauses. Based on their "synthetic disfluent speech model", the authors analysed the features that describe filled pauses and propose a model to predict them. After implementing the model in a unit selection synthesis system [17], a perception test was carried out. Results show no decreasing of the naturalness of the system, but also no significant increase. Furthermore, [18] included selected utterances from spontaneous conversational speech in a unit selection voice [19]. Utterances were synthesized by using this voice and by automatic prediction of type and placement of fillers and filled pauses. Results of a perception study suggest that synthetic speech sounds more conversational without degrading naturalness.

To our knowledge there has been barely any research on the role of uncertainty for *articulatory* speech synthesis. Our study investigates this question by using the system developed by [1]. We aim at combining different acoustic cues for modelling uncertainty and carry out a perception test.

3 Articulatory speech synthesis for modelling uncertainty

To generate our perception test material, we used the articulatory speech synthesis system VocalTractLab [1]. In this section we discuss the characteristics and potential advantages of articulatory speech synthesis (Sec. 3.1) and describe previous studies on paralinguistic aspects that were carried out with this synthesizer (Sec. 3.2).

3.1 Articulatory speech synthesis

Emotional speech is often very rich, and sometimes extreme, in variations. This can pose a problem to synthesis methods such as unit selection because they are based on a database of speech utterances which mostly need to undergo signal manipulation to meet the acoustic demands of the target utterance. This signal manipulation can introduce synthesis artefacts. The method of synthesis used here simulates the whole mechanical process of speech production exactly as needed for a target utterance. Therefore, extreme paralinguistic variation does not pose a particular problem for articulatory synthesis.

The synthesis process is controlled by a set of temporally aligned articulatory gestures for segmental articulation. They define the shape of a virtual three-dimensional vocal tract (see Fig. 1) by controlling the movements of virtual articulators. An aerodynamic-acoustic simulation of airflow produces the speech output [1]. While the control of supraglottal articulators such as the jaw, tongue, lips and velum primarily define the segmental, phonemic, or *linguistic* content of the speech output, *paralinguistic* features of the target speech can also be controlled. In our study, we primarily make use of the intonation contour control. The speech output sounds potentially very natural since the acoustic interaction of, e.g. high fundamental frequency and vowel formants, is calculated at the same time. In this case, even with intonation rising very high, the impact of acoustic artefacts is very low.

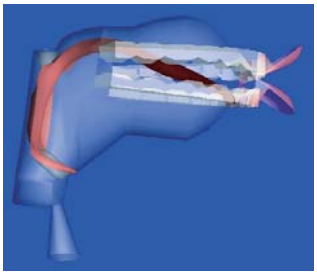


Figure 1: 3D model of the vocal tract of the articulatory speech synthesizer [1].

Apart from modelling paralinguistics with articulatory synthesis systems, they have for instance been used as virtual language trainers. The system ARTUR [20] e.g. is a virtual speech tutor who can use three-dimensional animations of the face and internal parts of the mouth. i.e. tongue, palate, jaw etc. to give feedback on the deviation between the user's pronunciation and the correct manner of articulation (cf. [21]).

3.2 Perception studies with articulatory synthesis

Articulatory synthesis in high audio quality is a relatively new field of research and many questions of gesture control are still open. Especially the modelling of emotion and attitude in articulatory speech synthesis has been barely investigated.

Recent studies with this synthesizer regarding paralinguistic aspects have shown that it is possible to synthesize conversational laughter which is rated as "natural" as human laughter in conversational context [22]. Additional studies investigated which features of articulation contribute to an increased perception of amusement when presenting speech that is overlayed with laughter (speech laughs), or speech that shows a paralinguistic quality of smile by manipulating different articulatory factors [23].

The design of these studies take advantage of the fact that, with articulatory synthesis, virtually every aspect of speech production is controllable independently from one another and the articulatory manipulations are reflected in detail in the acoustic output (see also e.g. voice quality variation and formant structure [24]). This makes it a tool suitable for testing the impact of different features of speech production on the perception in human listeners.

4 Previous study

In a previous study [25] we presented an initial investigation of the modelling of uncertainty by using the articulatory synthesizer in [1]. Furthermore, a perception experiment was carried out to test how the intended uncertainty was perceived. We generated several degrees of uncertainty by varying *intonation* (high vs. low), *delay* (presence vs. absence) and the *filler* "hmmm" (presence vs. absence).

Delay values were 1000 ms in an unmarked question-response case and 2200 ms for a "delayed" answer in uncertain stimuli without fillers. The *uncertain* stimuli that contained fillers had a delay structure of 1500 ms before the filler and another 1000 ms after the filler, before the two-word answer started. Variation of the F0 contour took place at the end of a stimulus, basically on the last word. It was characterized either by a rising or a falling F0 contour.

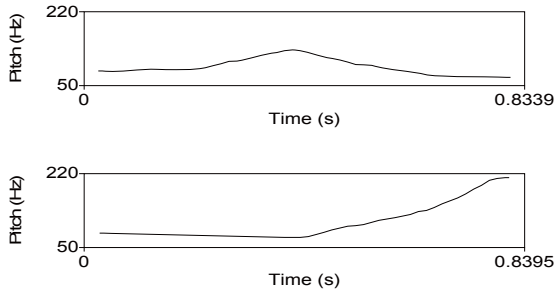


Figure 2: Illustrations of falling (top) and rising intonation (bottom) in the word "Melonen" (*melons*).

Our stimuli consisted of the two phrases "ziemlich kühl" (*pretty chilly)* and "eher kalt" (*rather cold*). For each phrase we generated four versions: The *certain* level was characterized by falling intonation and absence of delay and filler. *Uncertain 1* was marked by rising intonation exclusively, *uncertain 2* by rising intonation and delay and *uncertain 3* by rising intonation, delay, and filler. Altogether, there were eight relevant dialogues and two filler dialogues.

For embedding the stimuli in a context, we chose the interaction between a caller and a telephone weather expert system. The caller asked the question: "Wie wird das Wetter nächste Woche in X?" (*How is the weather going to be next week in X?*) and the program gave the answer. 34 students rated the dialogues between the caller and the weather expert system, scoring the answer of the system regarding its certainty and also its intelligibility on a 5-point Likert-Scale with 1 meaning *uncertain/unintelligible* and 5 meaning *certain/intelligible*, respectively. The results suggested significant effects for *uncertain 1*, but not for *uncertain 2*. Furthermore, we observed for *uncertain 3* the strongest effect on the perception of uncertainty. However, from the data of this first experiment, it was not clear to what extent delay alone and filler alone influence the perception of uncertainty. In addition, it could be argued that the choice of the phrases ("ziemlich" and "eher" as adverbs denoting vagueness) could also have conveyed different levels of certainty in themselves.

5 Perception study

5.1 Goal

While the previous study did not allow for the interpretation of all three factors independently, the goal of the current study is to model the different levels of uncertainty by using *all* possible combinations of the three cues *rising intonation*, *delay,* and *filler,* and to test if these intended levels of uncertainty have an impact on perception. For this, we assume a hierarchy of impact factors and evaluate it by a perception test.

5.2 Material

Our target stimuli contain four different one-word phrases, each one is generated in eight different levels of uncertainty (see Tab. 1) by permutation of the factors *intonation, delay*, and *filler*. They are appended to a question asked by a human and serve as answer in a fictitious dialogue (see Sec. 5.2.2).

The one-word phrases are from the semantic field of fruits and vegetables. The words each are three syllables long and carry the main stress on the second syllable to ensure rhythmic comparability. All of them have a voiced last syllable ending on <-en> or <-eln> to be able to carry intonational information equally well: "Melonen" (word 1, *melons*), "Bananen" (word 2, *bananas*), "Tomaten" (word 3, *tomatoes*), and "Kartoffeln" (word 4, *potatoes*). In a pretest,

Table 1: Dialogue stimuli features; levels of certainty: C: certain, U1..7: uncertain 1..7

ID	Question (human)	Answer (synthetic)	Level of certainty	Rising intonation	Filler	Delay
1	"Was siehst Du?" (*What do you see?*)	Wording 1/2/3/4	C	-	-	-
2			U1	-	-	+
3			U2	-	+	-
4			U3	-	+	+
5			U4	+	-	-
6			U5	+	-	+
7			U6	+	+	-
8			U7	+	+	+

we evaluated a set of words for their intelligibility. The stimuli used here received the top scores (based on 22 participants). As opposed to our previous study, we chose one-word sentences and excluded adverbs conveying vagueness (such as "ziemlich" and "eher") for better comparability and validity of the data.

The synthetic answers are based on copy-synthesis of recordings of a male speaker and show the following acoustic characteristics. **Falling vs. rising intonation:** The intonation contours of our human speaker were copied to generate one pattern for falling intonation and one for rising intonation. As in the previous study, the main difference in intonation is situated on the last part of the utterance, in this case the final syllable and parts of the pre-final syllable. In the case of rising intonation, the fundamental frequency increases to around 200 Hz, for falling intonation it decreases to around 70 Hz. The other parts of the utterance are in an unmarked pitch range of around 80 Hz. An illustration is shown in Fig. 2. **Delay:** Delay is defined as the time the system waits until it starts to respond after the question has finished. We defined a *short* delay of 1000 ms and a *long* delay of 2200 ms, similar to our previous study. **Filler:** The filler we use in the present study is worded "hmm". It can be present or absent. If present, it is placed before the stimulus followed by a pause of 1000 ms.

Thus delay and/or filler occur before the system says the actual content word; intonation (rising "?" or falling ".") occurs at the end of the system's response. E.g. (cf. Tab. 1):
C: "What do you see?" [silence 1000 ms] "Bananen."
U1: "What do you see?" [2200 ms] "Bananen."
U2: "What do you see?" [1000 ms] "Hmm [1000 ms] Bananen."
U7: "What do you see?" [2200 ms] "Hmm [1000 ms] Bananen?"

Rising intonation is assumed to support perception of uncertainty, whereas falling intonation is not. Similarly, long delay and filler should support uncertainty, whereas short delay and no filler are assumed to represent certainty.

The experiment also contains distractor stimuli. As for the target stimuli, they belong to the semantic field of produce. Their structure always represents the unmarked level of certainty with falling intonation, short delay, and no filler. The distractors are "Bohnen", "Gurken", "Paprika", "Knoblauch" (*beans, cucumber, sweet pepper, garlic*).

5.2.1 Hypothesis

Since in our previous study rising intonation had a significant influence on the uncertainty ratings, we assume the following scale of uncertainty for our target stimuli feature combination: Intonation > Filler > Delay. Based on this hierarchy, we designed the feature combination matrix shown in Tab. 1 which gives an overview of the activation or de-activation of each feature for each level of uncertainty. Since the previous study did not distinguish between the different combinations

Table 2: Distribution of stimuli dialogues and distractor dialogues for each of the four sets. The sequence within the sets was randomized for each group of participants.

ID	Set 1	Set 2	Set 3	Set 4
1	Word1-C	Word1-U1	Word1-U2	Word1-U3
2	Word1-U4	Word1-U5	Word1-U6	Word1-U7
3	Word2-U1	Word2-U2	Word2-U3	Word2-C
4	Word2-U5	Word2-U6	Word2-U7	Word2-U4
5	Word3-U2	Word3-U3	Word3-C	Word3-U1
6	Word3-U6	Word3-U7	Word3-U4	Word3-U5
7	Word4-U3	Word4-C	Word4-U1	Word4-U2
8	Word4-U7	Word4-U4	Word4-U5	Word4-U6
9	Dist1	Dist1	Dist1	Dist1
10	Dist2	Dist2	Dist2	Dist2
11	Dist3	Dist3	Dist3	Dist3
12	Dist4	Dist4	Dist4	Dist4

of filler and delay, the present study is an enhancement of the previous set-up as it will allow for analysis across all three factors individually. Additionally, the stimuli have the same syllable structure, and the segmental duration and intonation were obtained from a real speaker.

5.2.2 Scenario

For embedding the stimuli in a context, we chose an interaction between a research assistant and a robot that does image recognition. In our scenario, we explain to our participants that the image recognition depends on the quality of the images, such that high image quality results in a high recognition confidence on the part of the robot. The assistant shows images of fruits and vegetables to the robot, asking it each time "Was siehst Du?" (*What do you see?*). The robot gives the respective answer and is supposed to be conveying its (un)certainty in the paralinguistics of the answer.

5.2.3 Sets of stimuli

Since the four different one word phrases of the answers are generated at eight different levels of certainty, we obtain 32 target stimuli (4 x 8). To split up the load on the participants of the perception test, we generated four sub-sets of these stimuli. Thus, since we have eight levels of certainty, each set contains each level exactly once. Since we have four different one word phrases, each set contains each one word phrase twice. Tab. 2 shows the distribution of the stimuli for each set. Four additional dialogues are used as distractor stimuli.

5.3 Procedure

Subjects were 95 students (73 female, 22 male) of the University of Bonn with an average age of about 23 years (SD = 4.7 yrs; range: 18 to 59 yrs). All of them were native speakers of German. They were tested in four group experiments. Since we conducted the experiment at the beginning of different university courses, the number of participants per session varies: 19 participants in group 1, 29 in group 2, 29 in group 3, and 18 in group 4. For each group one of the four sets of stimuli was presented in a differently randomized order. The audio stimuli were played back over loudspeakers. An example stimulus was presented to the participants to make them familiar with the task and give them the chance to ask questions. During the main test, the participants could not ask any questions nor was any feedback given. For each dialogue between the research assistant and the robot, the subjects were asked to score the answer of the system regarding its certainty. We used a 7-point Likert-Scale with 1 meaning *uncertain* and 7 meaning *certain* respectively.

The results were statistically analysed using the Wilcoxon Signed Rank Test. This test was chosen since our dependent data were measured on an ordinal scale and the number of participants varies. The ratings of the stimuli were compared in pairs to test if there were significant differences in rating the intended uncertain and certain utterances. The null hypothesis (H0) was as follows: There is no dependency between the rating of the utterances as certain and their intended certainty and uncertainty respectively. The alternative hypothesis (H1) was: The rating of the utterances as certain/intelligible depends on their intended certainty and uncertainty respectively. The level of significance was 0.05.

5.4 Results

5.4.1 Target word ratings

Fig. 3 a) to d) show the median values of the participants' certainty ratings of the four target words. Overall, each word's ratings of certainty follow in most cases the main tendency as hypothesized in our hierarchy (see Tab. 3). Only in two out of 140 cases of pairwise comparisons, the medians show a reversed tendency in perceived uncertainty in a significant way (see below). I.e. the median of an intendedly more uncertain stimulus (according to the proposed hierarchy) is higher than a stimulus of a more certain level.

When we cluster the data by pooling the judgements for all four stimuli per level of uncertainty, we find nevertheless strong evidence for our hierarchy (cf. Fig. 3e, Tab. 3): Highly significant differences between rankings occur for most comparisons ($p<0.01$). For illustrating the relative contribution of each prosodic cue on the perception of uncertainty, the most important findings are shown Tab. 4.

i) **Delay:** Our hierarchy assumes *Intonation* > *Filler* > *Delay*. Tab. 4 shows the effects of delay by presenting pairwise comparisons that each differ only in the feature of *delay*. The results for the individual stimuli show for C vs. U1 only in the case of "Tomaten" a significant effect of delay as exclusive indicator of uncertainty ($p<0.05$). When delay is combined with filler (U2 vs. U3), our data reveal a marginally significant effect as opposed to filler alone. This can be observed for "Bananen" and "Kartoffeln" ($p<0.1$). When we pool the data, the impact is significant ($p<0.05$). In a similar way, delay combined with rising intonation contributes to a stronger perception of uncertainty than rising intonation alone (U4 vs. U5). This effect is highly significant for "Bananen", significant for "Tomaten" ($p<0.05$) and marginally significant for the pooled data ($p<0.1$). Furthermore, filler and rising intonation alone do in most cases not contribute to a stronger level of perceived uncertainty than rising intonation, filler and delay (U6 vs. U7). Only for "Melonen" the effect is marginally significant ($p<0.1$).

ii) **Intonation and filler**: The hierarchy proposes that *Intonation* > *Filler*. Tab. 4 shows the effects on the ratings in stimuli pairs where intonation *or* filler are the only features that change. For U2 vs. U4, the ratings of "Kartoffeln" show a significant difference ($p<0.05$) in agreement with our hierarchy, whereas "Tomaten" shows a significant difference with reversed tendency (testing with reversed hypothesis yields $p<0.01$). All other comparisons in this group (U2 vs. U4, U3 vs. U5) show no significant differences between judgements.

iii) **Intonation (+ delay) and filler (+ delay):** Tab. 4 shows the effects of filler combined with delay vs. rising intonation alone (U3 vs. U4) and of filler alone vs. rising intonation combined with delay (U2 vs. U5). U3 vs. U4 comparisons show no significant differences except for

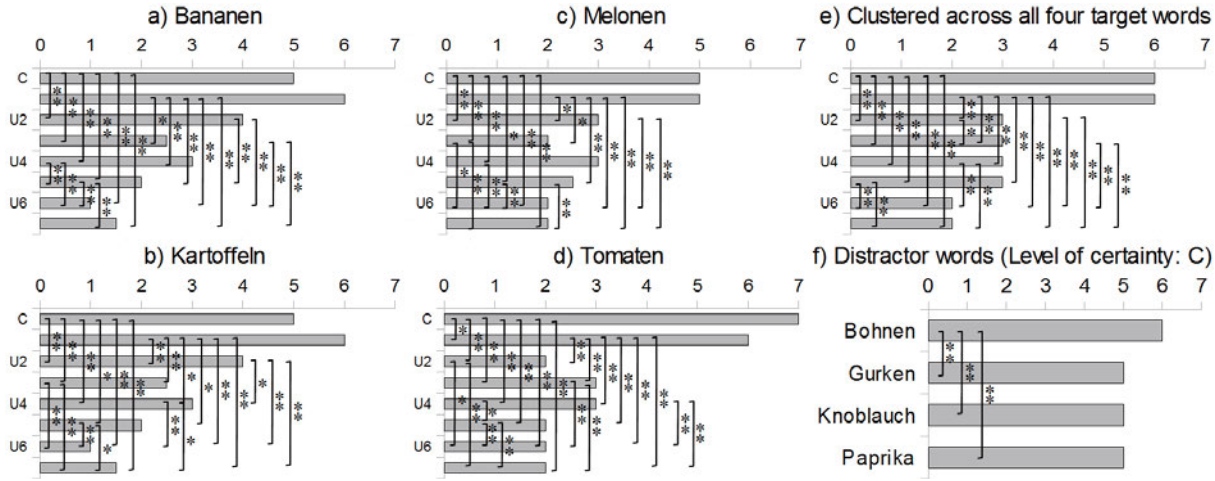


Figure 3 Medians of certainty perception. 3 a) to d) Individual results of the four target words "Bananen", "Kartoffeln", "Melonen", "Tomaten". 3 e) Results over all target words. 3 f) Results of distractor words. 1 = uncertain, 7 = certain. *: significant differences between judgements, **: highly significant differences between judgements.

"Bananen" where U4 is marginally significantly more certain than U3 (reversed tendency test yields $p<0.1$). For U2 vs. U5, our data show a highly significant effect for "Bananen" ($p<0.01$) and a marginally significant effect for the pooled data ($p<0.1$).

To summarize, our results firstly suggest that delay in responses increases the perception of uncertainty in general if either filler or rising intonation or both features are deactivated. Secondly, our data show that in most cases filler alone is not ranked significantly different from rising intonation alone. Finally, our data suggest some evidence that rising intonation combined with delay contributes to a stronger level of perceived uncertainty than filler alone, but the effect seems too weak for a generalization.

5.4.2 Distractors

Fig. 3 f) shows the ratings of the four distractor words. Each distractor represents the intended certainty of level C, being characterized by absence of all three prosodic cues. "Bohnen" is ranked with a median of 6; "Gurken", "Knoblauch" and "Paprika" achieve a median of 5 each. The comparison between "Bohnen" and the three other distractors shows a highly significant difference (each time $p<0.01$). Our data show that "Paprika" is judged as more certain than "Knoblauch" in a marginally significant way ($p<0.1$). The remaining two comparisons do not show significant differences ($p>0.05$). To conclude, it can be said that the distractors used in this experiment achieve rankings of 5 or 6, similar to the rankings for the certain level of our stimuli, where judgements range from 5 to 7.

6 Discussion

We presented a study on the perception of uncertainty expressed by a combination of the three different prosodic cues response delay, rising intonation, and filler. In our scenario we used a fictitious human-robot interaction on the topic of picture recognition. We assumed a hierarchy of uncertainty: *Intonation > Filler > Delay*. Our data show evidence that the intended level of certainty, characterized by absence of all three cues, is always judged as more certain than the strongest level of intended uncertainty, marked by presence of all three cues. This occurrence is highly significant. As assumed, effects of delay on the perception of uncertainty are relatively weak and mostly occur when filler or rising intonation or both cues are absent. If both cues are present, the effect of delay decreases. Against our expectation our data suggest evidence that filler and rising intonation have a similarly strong effect on the perception of uncertainty. The proposed hierarchy then needs to be reformulated to: *Intonation = Filler > Delay*.

This study as well as our previous study [25] suggest empirical evidence that articulatory synthesis can be used for analysis-by-synthesis perception test paradigms. This synthesis technique would even enable testing of e.g. speech rate variations without introducing audio signal artefacts. Additional extensions could be to investigate visual effects on the perception of uncertainty by using the visualization possibilities of the vocal tract and articulatory gestures. Furthermore, the results of our study can serve as guideline for automatic detection of uncertainty in spoken dialogue systems or tutoring systems. A higher detection accuracy of uncertain answers of a student by using the revised hierarchy could help to improve dialogue handling in these systems.

Acknowledgements: Many thanks to Berthold Crysmann, Bernd Möbius, Ulrich Schade and Bernhard Schröder for helpful comments.

References

[1] Birkholz, P. (2005). *3-D Artikulatorische Sprachsynthese*. Berlin: Logos Verlag.

[2] Rozin, P. and Cohen, A.B. (2003). High frequency of facial expressions corresponding to confusion, concentration, and worry, in an analysis of naturally occurring facial expressions in Americans. In *Emotion*, 3, 68-75.

[3] Kuhlthau, C. C. (1993). A principle of uncertainty for information seeking. In *Journal of Documentation*, 49(4), 339-355.

[4] Smith, V. and Clark, H. (1993). On the course of answering questions. In *Journal of Memory and Language*, 32, 25-38.

[5] Hart, J. T. (1965). Memory and the feeling-of-knowing experience. In *Journal of Educational Psychology*, 56, 208–216.

[6] Brennan, S. E. and Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. In *Journal of Memory and Language*, 34, 383-398.

[7] Swerts, M. and Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. In *Journal of Memory and Language 53 1*, 81-94.

[8] Wollermann, C. and Schröder, B. (2008). Does Uncertainty Effect the Case of Exhaustive Interpretation? In *Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics*. Athens, Greece, 233-236.

[9] Wollermann, C. and Schröder, B., (2008). Certainty, Context and Exhaustivity of Answers. *Paper presented at Speech and Face to Face communication*, Grenoble, France.

Table 3: Significance values of pairwise comparisons between different levels of certainty using Wilcoxon Signed Rank Test. Significant results are marked in **bold**.

Levels of comparison	Bananen		Kartoffeln		Melonen		Tomaten		Clustered	
	W	p	W	p	W	p	W	p	W	p
C vs. U1	273.5	0.52	201	0.17	301.5	0.18	533	**0.03**	5083	0.06
C vs. U2	439.5	**<0.01**	480	**<0.01**	544	**<0.01**	512	**<0.01**	8266.5	**<0.01**
C vs. U3	291	**<0.01**	477	**<0.01**	772	**<0.01**	528.5	**<0.01**	8302	**<0.01**
C vs. U4	272.5	**<0.01**	298.5	**<0.01**	772	**<0.01**	776	**<0.01**	7990	**<0.01**
C vs. U5	536.5	**<0.01**	306	**0.01**	443	**0.03**	817	**<0.01**	8366.5	**<0.01**
C vs. U6	551	**<0.01**	522	**<0.01**	551	**<0.01**	514	**<0.01**	8968	**<0.01**
C vs. U7	318.5	**<0.01**	517.5	**<0.01**	837.5	**<0.01**	551	**<0.01**	8898.5	**<0.01**
U1 vs. U2	664.5	0.05	495	**<0.01**	300	**0.03**	497.5	**<0.01**	7884	**<0.01**
U1 vs. U3	445.5	**0.02**	491	**<0.01**	439	**0.04**	511.5	**<0.01**	7998	**<0.01**
U1 vs. U4	416	**<0.01**	308	**0.01**	429	0.09	723.5	**<0.01**	7649.5	**<0.01**
U1 vs. U5	801	**<0.01**	315.5	**0.03**	250.5	**<0.01**	777.5	**<0.01**	8034	**<0.01**
U1 vs. U6	841	**<0.01**	551	**<0.01**	328.5	**<0.01**	505	**<0.01**	8909	**<0.01**
U1 vs. U7	485.5	**<0.01**	545	**<0.01**	498	**<0.01**	551	**<0.01**	8799.5	**<0.01**
U2 vs. U3	326	0.07	509	0.08	285	0.42	136	0.87	5145	**0.04**
U2 vs. U4	267.5	0.57	356.5	**0.02**	244	0.76	147	0.99	4527.5	0.48
U2 vs. U5	593	**<0.01**	266.5	0.58	152.5	0.73	238	0.71	5025	0.08
U2 vs. U6	777	**<0.01**	762	**<0.01**	259	**<0.01**	220	**0.03**	7539	**<0.01**
U2 vs. U7	432	0.07	708.5	**<0.01**	411	**<0.01**	243.5	**<0.01**	7240.5	**<0.01**
U3 vs. U4	123	0.94	294.5	0.23	362.5	0.83	204.5	0.94	3897	0.95
U3 vs. U5	296	0.22	215.5	0.9	224	0.81	301.5	0.29	4396	0.62
U3 vs. U6	462.5	**<0.01**	641	**<0.01**	351.5	**0.04**	257.5	**<0.01**	6946	**<0.01**
U3 vs. U7	252.5	**<0.01**	608.5	**<0.01**	591.5	**<0.01**	293	**<0.01**	6667.5	**<0.01**
U4 vs. U5	392.5	**<0.01**	111	0.97	257.5	0.54	558.5	**0.01**	5019.5	0.08
U4 vs. U6	526	**<0.01**	364.5	**<0.01**	410	**<0.01**	436	**<0.01**	7519	**<0.01**
U4 vs. U7	293.5	0.07	350	**0.02**	653.5	0.08	499.5	**<0.01**	7225.5	**<0.01**
U5 vs. U6	693.5	**<0.01**	486	**<0.01**	245	**<0.01**	374	**<0.01**	7051.5	**<0.01**
U5 vs. U7	372.5	**<0.01**	456	**0.04**	400.5	**<0.01**	410.5	**<0.01**	6774.5	**<0.01**
U6 vs. U7	210.5	0.9	410	0.57	334	0.09	161	0.64	4312	0.72

Table 4: Effects between different levels of certainty; pairwise comparisons using Wilcoxon Signed Rank Test. Significant results are marked in **bold.**

Levels	Influences of delay				Direct comparison: filler vs. intonation		Conjugated comparison: filler vs. intonation (with delay)	
Wordings	**C vs. U1**	**U2 vs. U3**	**U4 vs. U5**	**U6 vs. U7**	**U2 vs. U4**	**U3 vs. U5**	**U3 vs. U4**	**U2 vs. U5**
Bananen	>0.05	<0.1	**<0.01**	>0.05	>0.05	>0.05	>0.05	**<0.01**
Kartoffeln	>0.05	<0.1	>0.05	>0.05	**<0.05**	>0.05	>0.05	>0.05
Melonen	>0.05	>0.05	>0.05	<0.1	>0.05	>0.05	>0.05	>0.05
Tomaten	**<0.05**	>0.05	**<0.05**	>0.05	>0.05	>0.05	>0.05	>0.05
All	<0.1	**<0.05**	<0.1	>0.05	>0.05	>0.05	>0.05	<0.1

[10] Wollermann, C., Schade, U., Fisseni, B. and Schröder, B. (2010). Accentuation, Uncertainty and Exhaustivity - Towards a Model of Pragmatic Focus Interpretation. In *Proceedings of Speech Prosody 2010*, Chicago, IL.

[11] Forbes-Riley, K. and Litman, D. (2008). Adapting to Student Uncertainty Improves Tutoring Dialogues. *In Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED)*, Brighton, UK, 60-69.

[12] Liscombe, J., Hirschberg, J. and Venditti, J. J. (2005). Detecting certainness in spoken tutorial dialogues. In *Proceedings of Interspeech,* Lisbon, Portugal, 1837-1840.

[13] Litman, D. and Silliman, S. (2004). ITSPOKE: An intelligent tutoring spoken dialogue system. In *Companion Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (HLT/NAACL), Boston, MA.

[14] Pon-Barry, H. and Shieber, S. (2009). The importance of sub-utterance prosody in predicting level of certainty. *Proceedings of NAACL-HLT Short papers,* Boulder, Colorado, 105-108.

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

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

[17] Bonafonte, A., Agüero, P. D., Adell, J., Pérez, J. and A. Moreno (2006). Ogmios: The upc text-to-speech synthesis system for spoken translation. In *Proceedings of the TC-STAR Workshop on Speech-to-Speech Translation*.

[18] Andersson, S., Georgila, K., Traum, D., Aylett, M. and Clark, R. A. J. (2010). Prediction and Realisation of Conversational Characteristics by Utilising Spontaneous Speech for Unit Selection. In *Proceedings of Speech Prosody 2010*, Chicago.

[19] Andersson, J., Badino, L., Watts, O. and Aylett, M. (2008), The CSTR/CereProc Blizzard entry 2008: The inconvenient data. In *The Blizzard Challenge*, Brisbane, Australia.

[20] Engwall, O. (2008). Can audio-visual instructions help learners improve their articulation? - An ultrasound study of short term changes. *Proceedings of Interspeech,* Brisbane, 2631-2634.

[21] ARTUR - the ARticulation TutoR. http://www.speech.kth.se/multimodal/ARTUR, last retrieved: 13th of april 2010.

[22] Lasarcyk, E. and Trouvain, J. (2007). Imitating conversational laughter with an articulatory speech synthesizer. In *Proceedings of the Interdisciplinary Workshop on the Phonetics of Laughter*, Saarbrücken, Germany, 43-48.

[23] Lasarcyk, E. (2007). Investigating Larynx Height With An Articulatory Speech Synthesizer. In *Proceedings of the 16th ICPhS*, Saarbrücken, Germany,

[24] Lasarcyk, E. and Trouvain, J. (2008). Spread lips + raised larynx + higher F0 = smiled speech? - An articulatory synthesis approach. In *Proceedings of the 8th International Speech Production Seminar* (ISSP '08), Strasbourg, France, 345-348.

[25] Wollermann, C. and Lasarcyk, E. (2007). Modeling and perceiving of different degrees of certainty in articulatory speech synthesis. In *Proceedings 6th ISCA Workshop on Speech Synthesis*, Bonn, 40-45.