Modeling and Perceiving of (Un)Certainty in Articulatory Speech Synthesis

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

[1]Institute of Communication Sciences, University of Bonn, Germany
[2]Institute of Phonetics, Saarland University, Germany
cwo@ifk.uni-bonn.de, evaly@coli.uni-sb.de

Abstract

This paper deals with the role of paralinguistic expression in articulatory speech synthesis. We describe two experiments which investigate the perception of certain vs. uncertain utterances produced by articulatory speech synthesis, using the system developed in [1].

Experiment 1 tests to what extent subjects are able to identify certainty and uncertainty as intended paralinguistic expressions in the acoustical signal by the varying acoustic cues intonation and delay. Further on, we investigate if (un)certainty influences the intelligibility of the synthetic utterances. Results show that the utterances are identified as intended with respect to (un)certainty. Regarding intelligibility, hardly any influence is measurable.

Experiment 2 looks more in detail into the perception of uncertainty by using several levels. Therefore, not only intonation and delay are varied as acoustical cues but also fillers. Results show that our intended different levels of uncertainty indeed evoked different degrees of perceived uncertainty.

1. Introduction

The role of emotion and attitude in human-machine interaction has gained extensive importance in the last few years. One interesting question in this context is to what extent machines are able to recognize emotions in spoken dialogs (e.g. [2], [3], [4]). A typical scenario would be the interaction between a user and a spoken dialog system. Here, emotion detection of the user can be used in order to modify the dialog (cf. [5]). For instance in the case of user frustration or annoyance, the system could react adequately.

On the speech synthesis side, the modeling of emotion and attitude has gained more and more importance as one aims to generate synthetic speech which is as natural and human-like as possible. Emotional TTS systems have been developed by [6] and [7] among others. Most of the emotional TTS systems are based on the prototypical emotions *happiness*, *sadness*, *anger*, *fear*, *surprise* and *disgust* according to [8]. Beyond that, the interface EmoSpeak as part of the TTS System MARY ([7], [9]) uses *evaluation*, *activation* and *power* as basic dimensions for representing emotional states[1]. Thus it is possible to express "... gradual emotional states in a more flexible way than has previously been possible." ([7]: Preface).

[1] The idea that emotional experience can be presented by using gradual dimensions goes back to [10]. An overview of the different models and names for the dimensions is given in [7].

Current emotional TTS systems are based on different techniques. One technique, which has become more and more popular, is *Unit Selection* ([11], [12]). With this technique, a reduced set of units is processed from a large speech corpus for concatenative synthesis (cf. [13]: 279). The advantage of this technique can be described as follows: "The synthesis often perceived as being most natural is unit selection, or large database synthesis, or speech re-sequencing synthesis." ([7]: 91). But the big drawback is that the technique shows a deficit when there are no appropriate units in the synthesizer (cf. [23]: 4).

The approach chosen for this study is articulatory speech synthesis. The 3D-articulatory synthesizer used here [1] allows for a great degree of freedom and precise adjustment of single parameters at the same time. It is not limited by any set of prerecorded utterances. Thus it might be suitable for emotional synthesis, which, by nature, can be very rich in variations. Nevertheless, the modeling of emotion and attitude in articulatory speech synthesis has been barely investigated. There are, however, some recent projects investigating the synthesis of laughter [15] or voice quality variation [16] with articulatory speech synthesis, which let us catch a glimpse of the continuum of possible manipulations with this kind of synthesis.

1.1. Production and perception of (un)certainty in natural speech

In order to simulate emotional ways of speaking or to convey paralinguistic expressions in synthetic speech, it is firstly necessary to know how emotional or paralinguistic cues are produced and perceived in *natural* speech. In the following, we will give an overview of selected studies that deal with the role of (un)certainty in human-human dialog.

The work of [17] serves as source of inspiration for many studies on this field (e.g. [18], [19]). The authors investigated memory processes in question answering situations. Question-answering in their framework is regarded as a social process, which is characterized by information exchange and also self-presentation (cf. [19]). For testing the hypothesis that uncertainty of a speaker is marked differently than certainty, they use Hart's [20] so-called *Feeling of Knowing (FOK)* paradigm. With this method, it is possible to elicit meta-memory judgments. Their experimental investigation brought to light that uncertainty is not only signaled by using linguistic hedges like "I guess", but also by prosodic features like rising intonation and delay (cf. [19]).

In order to investigate how people perceive the *FOK* of another speaker, [18] defined the *Feeling of Another's Knowing (FOAK)*. Their study showed that the *FOAK* "... was affected by the intonation of answers, the form of answers ...

the latency to response, and the presence of fillers." ([18]: 396). The term *filler* is defined as "interjections such as 'um', 'uh', 'hmm'" ([18]: 383).

As the studies mentioned focused on the role of (un)certainty in the *acoustical* signal, the question remained open about which role (un)certainty plays in the *audiovisual* modality. With respect to production, Swerts, Krahmer and colleagues ([21], [19]) found that several characteristics are used for the production of uncertainty: *delay*, *pause* and *fillers* for the audio modality; *smiles*, *"funny faces"* etc. for the visual one. Tests on the perception side showed that subjects were able to distinguish certain from uncertain utterances for all three conditions (audio-only, visual-only, and audiovisual); the identification was even easier in the bimodal condition than in the unimodal conditions (cf. [19]).

1.2. Characteristics and goal of the current study

The current study deals with the modeling and perception of different degrees of certainty in articulatory speech synthesis. The stimuli are characterized by a high/low degree of certainty (experiment 1) as well as by several more fine-grained degrees of uncertainty (experiment 2).

For generating our stimuli, we use the articulatory speech synthesis system developed in [1]. It offers a high degree of speech quality combined with a very high degree of high level control over all articulatory parameters (cf. Sec. 2.2).

The goal of experiment 1 is to investigate if subjects are able to distinguish intended *certain* utterances from intended *uncertain* ones under the audio condition. Another purpose is to survey if certainty influences the intelligibility of the synthetic utterances.

The purpose of experiment 2 is to look into the perception of (un)certainty in articulatory synthesis more in detail by using several degrees of uncertainty. Thus, it should be determined which acoustical cues exactly are relevant for perceiving an utterance as *uncertain*.

2. Modeling (un)certainty in articulatory speech synthesis

2.1. Acoustical criteria for modeling (un)certainty

According to [21], uncertainty will be distinguished from certainty acoustically along the dimensions *delay* (presence or absence), *intonation* (high or low) and, later on (in experiment 2), also *fillers* (presence or absence), as shown in Table 1.

Delay values are 1000 ms in an unmarked question-response case (for the *certain* stimuli) (cf. [17], who report an average silence of 0.97 s), and 2200 ms for a "delayed" answer in uncertain stimuli without fillers. The *uncertain* stimuli that do contain fillers (we chose the sound "hmmm") have a delay structure of 1500 ms before the filler and another 1000 ms after the filler, before the actual two-word answer starts.

Variation of the F0 contour takes place at the end of a stimulus, basically on the last word. It is characterized either by a rising F0 contour (*high intonation*, according to [21]) or a falling F0 contour (*low intonation*).

Table 1: Acoustical criteria for (un)certainty according to [21].

Acoustic cue	Certain	Uncertain
Delay	-	+
High intonation	-	+
Low intonation	+	-
Filler	-	+

2.2. The articulatory speech synthesis system

As described above, the test stimuli are generated with the speech synthesizer in [1]. This synthesizer uses a three-dimensional model of the vocal tract (see Fig. 1). Based on its geometry, an aerodynamic-acoustic simulation generates the speech output. The shape of this geometrical model is controlled by a *gestural score* on which the pertinent parameters are varied according to the intended articulation.

The movements of the supraglottal articulators (such as lips, tongue, jaw, velum) can be subsumed under vocalic, consonantal, and velar gestures. Their basic interaction is held constant to convey the phonemic (*linguistic*) content of the utterance, i.e. the words. On top of that, *paralinguistic* features are changed according to the degree of certainty aimed at. In this way, the F0 movements are specified on an F0 tier, matching the desired intonational patterns.

Fillers can be inserted when needed by simply adding the corresponding gestures in the score.

The variations in response delay are accounted for during preparation of the *complete* stimulus as a question-answer pair (see next section below).

In addition to the audio part, the system is characterized by a three-dimensional visualization of the articulatory gestures (see Fig. 1). The lips can be seen together with the front teeth and parts of the tongue. Other facial parts such as eyes or eyebrows are not displayed.

As our goal for the current study is to have a *first* look into the perception of (un)certainty in articulatory speech synthesis, we will focus in our experiments on the pure audio condition. Future work is also going to consider the visual part.

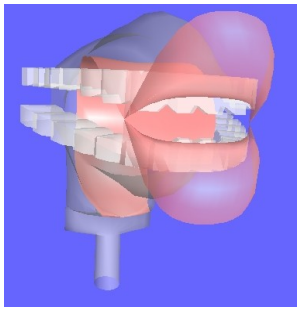

Figure 1: 3D model of the vocal tract of the articulatory speech synthesizer [1].

Table 2: Features of the dialogs presented in experiment 1 (with the starred IDs, 6 in total), and experiment 2 (ID 1 - 4; 7 - 10). Levels of certainty: C: *certain*, U1: *uncertain 1*, U2: *uncertain 2*, U3: *uncertain 3*. For further explanations cf. Sec. 2.

ID	Caller's question	System's answer	Level of certainty	Intonation high	Delay	Filler
1*	"Wie wird das Wetter nächste Woche in X?"	"Ziemlich kühl"	C	-	-	-
2			U1	+	-	-
3*			U2	+	+	-
4			U3	+	+	+
5*		"Relativ heiss"	C	-	-	-
6*			U(2)	+	+	-
7*		"Eher kalt"	C	-	-	-
8			U1	+	-	-
9*			U2	+	+	-
10			U3	+	+	+

2.3. Scenario

For embedding the stimuli in a context, we chose the interaction between a caller and a telephone weather expert system. The caller asks the question: "Wie wird das Wetter nächste Woche in X?" (*How is the weather going to be next week in X?*) and the program gives an answer. Since this study is meant as an initial investigation, the answers are very short (two-word sentences) and there are only three different wordings: "ziemlich kühl" (*pretty chilly)*, "relativ heiss" (*relatively hot*) and "eher kalt" (*rather cold*).

For experiment 1, each wording is generated in two versions: in a *certain* and an *uncertain* way of speaking. All in all there are six dialogs. They are shown with a starred ID in Tab. 2.

For experiment 2, we leave out the wording "relativ heiss" due to the low intelligibility[2] measured in experiment 1 (cf. Sec. 3.1.3). The other two wordings ("ziemlich kühl", "eher kalt") are generated in four versions: One *certain* way of speaking, and three *uncertain* ones. They are intended to capture different degrees or different acoustic aspects of uncertainty. The final intonation is always high but there are differences regarding delay times and fillers. The presumably weakest version concerning the level of uncertainty (*uncertain 1*) has no more marked features (only high intonation), a middle version (*uncertain 2*) possesses the delay structure mentioned in Sec. 2.1 and high intonation, and the strongest version (*uncertain 3*) incorporates all acoustic signals of uncertainty concentrated on in this paper (i.e. intonation, delay, and filler). Altogether, there are eight relevant dialogs (see Tab. 2, IDs 1 - 4 and 7 - 10).

Two additional wordings are generated: "Trocken" (*dry*) and "heiss" (*hot*). These wordings are embedded in different contexts and the resulting dialogs serve as filler items[3]. In addition, "trocken" embedded in another dialog is used as an example for making subjects familiar with the stimuli. In these cases, the level of certainty of the wordings is intended to be neutral. We define *neutrality* as an unmarked version between *certain* and *uncertain*: The acoustic features show regular delay, regularly low intonation (not as deep as in the certain version), and contain no fillers.

[2] Technical problems with the initial consonant in "relativ" seemed to be the reason for this.

[3] *Filler items* in this case are items which are not intended to be relevant to this experiment. Thus, the ratings for the example dialog and also for the filler items will not be considered in the analyses.

3. Perception studies

3.1. Experiment 1

3.1.1. Goal

The goal of experiment 1 is to determine if subjects are able to recognize certainty and uncertainty as intended paralinguistic expressions in articulatory speech synthesis under audio condition. Another purpose is to investigate if certainty affects the intelligibility of articulatory synthetic utterances.

3.1.2. Method

Subjects were 38 students of the Universities of Bonn and Saarbrücken with an average age of about 25.5 years. 18 participants were female, 20 male, all of them German native speakers. They were tested in group experiments or individually: The audio stimuli were presented to them in two different random orders[4] over a loudspeaker. When the example stimulus was presented to the subjects, they had the chance to ask questions. After the procedure started, subjects were neither supposed to ask any questions nor was any feedback given. For each dialog between the caller and the weather expert system, the subjects were asked to score 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 were statistically analyzed using the Wilcoxon Signed Rank Test. This test was chosen since our dependent data were measured on an ordinal scale. 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 (H_0) was as follows: There is no dependency between the rating of the utterances as *certain*/*intelligible* and their intended certainty and uncertainty respectively. The alternative hypothesis (H_1) was: The rating of the utterances as *certain*/*intelligible* depends on their intended certainty and uncertainty respectively. The level of significance was 5 %.

3.1.3. Results

Results for the perception of *certain* and *uncertain* utterances regarding their certainty are visualized in Fig. 2 and Tab. 3. The intended *certain* versions of "ziemlich kühl", "relativ heiss", and "eher kalt" were all rated with a median of 4, whereas the median for the *uncertain* versions received a lower median of 3. The comparison of both data series showed a highly significant difference for each wording

[4] In order to minimize the influence of the sequence of the stimuli when calculating the overall results.

("ziemlich kühl": V = 342.5; p < 0.001, "relativ heiss": V = 308.5; p < 0.001, "eher kalt": V = 351; p < 0.001).

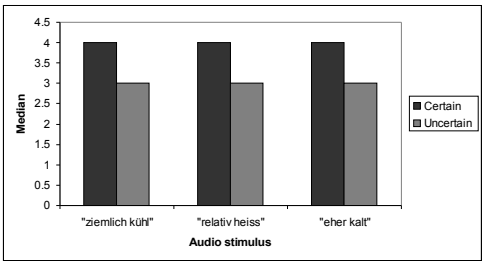

Figure 2: Medians for perceiving certainty of the three wordings in the *certain* and the *uncertain* way of speaking (experiment 1).

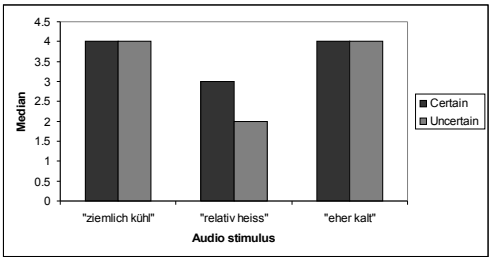

Figure 3: Medians for perceiving intelligibility of the three wordings in the *certain* and the *uncertain* way of speaking (experiment 1).

Table 3: V and *p* values of the pairwise comparison of *certain* vs. uncertain utterances; significant differences are marked in bold face.

Wording	Certainty	Intelligibility
"Ziemlich kühl"	**342.5; < 0.001**	56.5; 0.63
"Relativ heiss"	**308.5; < 0.001**	**123; 0.04**
"Eher kalt"	**351; < 0.001**	79.5; 0.14

The results for the intelligibility of (un)certain utterances are shown in Fig. 3 and Tab. 3. Subjects ranked both the *certain* and the *uncertain* versions of "ziemlich kühl" and "eher kalt" with a median of 4. In both wordings, there was no significant difference between these two data series ("ziemlich kühl": V = 56.5; p = 0.63, "eher kalt": V = 79.5; p = 0.14).

Further on, the median value for "relativ heiss" in a *certain* way of speaking was 3, whereas the *uncertain* version had a lower median of 2. The statistical analysis resulted in a significant difference between the judgments (V = 123; p = 0.04).

When regarding the absolute ranking values for intelligibility, it became obvious that the wording "relativ heiss" was less intelligible than the other two wordings. Statistical testing showed that the *certain* version of "relativ heiss" was rated significantly less intelligible than the *certain* versions of the other wordings ("relativ heiss" vs. "ziemlich kühl": V = 398.5; p < 0.001, "relativ heiss" vs. "eher kalt": V = 514; p < 0.001). The intelligibility of the *uncertain* version of "relativ heiss" was also rated significantly lower than that of other two wordings ("relativ heiss" vs. "ziemlich kühl": V = 465; p < 0.001, "relativ heiss" vs. "eher kalt": V = 547; p < 0.001).

In summary, as the results of the perception of certainty indicate, subjects could clearly distinguish between the intended *certain* and *uncertain* utterances in all wordings.

Furthermore, intelligibility was only very weakly influenced by the intended certainty and uncertainty, respectively.

3.1.4. Discussion

The relatively low ranking of the intelligibility of "relativ heiss" might come from the fact that some of the phones used in this utterance seemed to be hard to understand (presumably the /r/ of "relativ"), because they presented some technical problems during the speech generation process. Therefore, only the wordings "ziemlich kühl" and "eher kalt" will be considered in experiment 2.

Since experiment 1 is meant to be an initial investigation and therefore focuses on the perception of intonation and delay, the question remains open which role *fillers* play in perceiving articulatory speech synthesis. This leads to the setup of a second experiment in which the stimuli cover *more* acoustic aspects of uncertainty by defining different degrees of uncertainty to get more detailed results.

3.2. Experiment 2

3.2.1.Goal

The goal of experiment 2 is to determine if there is a ranking regarding the impact of the different cues signaling uncertainty. Thus, subjects are tested to find out to what extent different combinations of acoustic cues affect the perception of uncertainty.

3.2.2. Method

The same method was applied as in experiment 1. Subjects were 34 seminar students[5] (23 females, 11 males, average age of 23 years), tested within three group experiments, each one having a different random order of stimuli. After listening to one example dialog, subjects were presented 10 test dialogs[6]. For each answer of the expert system, they were asked to evaluate the certainty on a 5-point Likert scale with 1 meaning *uncertain* and 5 meaning *certain*.

The results were again analyzed using the Wilcoxon Signed Rank test. Like in experiment 1, the ratings of the stimuli were compared pairwise to test if there were significant differences in rating the intended certain utterances and uncertain ones. However, now there were three different levels of uncertainty. The null hypothesis (H_0) was as follows: There is no dependency between the rating of the utterances as *certain* and *uncertain,* respectively, and their particular level of certainty. The alternative hypothesis (H_1) was: The rating of the utterances as *certain* and *uncertain,* respectively,

[5] Subjects were different from those of experiment 1.
[6] Including the two filler items described in Sec. 2.3.

depends on their intended certainty and level of uncertainty, respectively. The level of significance was 5 %.

3.2.3. Results

The results for the perception of "ziemlich kühl" and also of "eher kalt" are displayed in Fig. 4 and Tab. 5. The certain version of "ziemlich kühl" was rated with a median of 4 as most certain compared to the uncertain versions. The comparison of the data series between *certain* and *uncertain 1* (median = 3) was statistically significant ($V = 162$; $p < 0.01$). In a similar way, *uncertain 2*, compared to *certain*, achieved a median value of 3 in a statistically highly significant way ($V = 251$; $p < 0.001$). *Uncertain 3* was rated lowest with a median of 2. The difference between *certain* and *uncertain 3* was highly significant ($V = 519$; $p < 0.001$). The graph also shows that the ratings for "ziemlich kühl" were very similiar with 3.5 and 3 for *uncertain 1* and *uncertain 2*. The statistical analysis showed no significant difference ($V = 64.5$; $p = 0.11$). In contrast, the difference between *uncertain 1* (with a median of 3) and *uncertain 3* was highly significant ($V = 504.5$; $p < 0.001$), as well as the one between *uncertain 2* and *uncertain 3* ($V = 369.5$; $p < 0.001$).

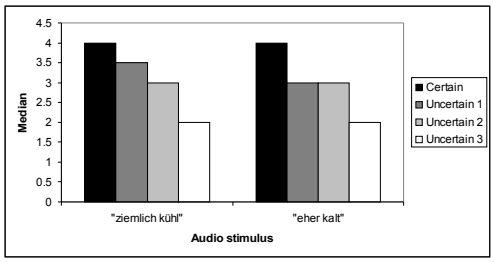

Figure 4: Medians for perceiving certainty of two wordings in one *certain* and three *uncertain* ways of speaking (experiment 2).

Table 5: Results of the pairwise comparisons of *certain* (C) vs. different types of *uncertain* utterances (U1,2,3) in experiment 2. Significant differences are marked in bold face.

Wording	Levels of certainty compared	*V* value; *p* value
"Ziemlich kühl"	**C vs. U1**	**162; < 0.01**
	C vs. U2	**251; < 0.001**
	C vs. U3	**519; < 0.001**
	U1 vs. U2	64.5; 0.11
	U1 vs. U3	**504.5; < 0.001**
	U2 vs. U3	**369.5; < 0.001**
"Eher kalt"	**C vs. U1**	**268; < 0.001**
	C vs. U2	**210; < 0.001**
	C vs. U3	**528; < 0.001**
	U1 vs. U2	70.5; 0.32
	U1 vs. U3	**397.5; < 0.001**
	U2 vs. U3	**319; < 0.001**

"Eher kalt" in a *certain* version of speaking was judged with a median of 4, whereas the median of *uncertain 1* was lower with a median of 3. The difference between the two data series was highly significant ($V = 268$; $p < 0.001$). The lower ranking of *uncertain 2* (median = 3) in comparison with the one for *certain* was also statistically significant ($V = 210$, $p < 0.001$). Furthermore, *uncertain 3* obtained a much lower median with 2 than *certain*. The statistical analysis resulted in a highly significant difference ($V = 528$; $p < 0.001$).

Additionally, the analysis showed that the rankings for *uncertain 1* differed not significantly from those for *uncertain 2* ($V = 70.5$; $p = 0.32$): the median value was 3 each time. In contrast to that, the judgments for *uncertain 1* and those for *uncertain 3* yielded a highly significant difference ($V = 397.5$; $p < 0.001$). Along these lines, the rankings for *uncertain 2* also differed in a highly significant way from those for *uncertain 3* ($V = 319$; $p < 0.001$).

In summing up, the results indicate that each of the intended *uncertain* versions (of all levels) were clearly perceived as being more uncertain than the *certain* versions for both wordings.

Within the set of uncertain stimuli for each wording, *uncertain 3* was judged significantly less certain than the other two levels of certainty.

However, in both of the wordings, there was no significant difference in evaluating the degree of certainty of *uncertain 1* vs. *uncertain 2*.

3.2.4. Discussion

First of all, the results of experiment 2 generally confirm the ones of experiment 1, in that intended certain utterances can be clearly distinguished from uncertain ones. While in experiment 1 there was only one level of uncertainty, conveyed by high intonation and delay, our more fine-grained analysis in experiment 2 showed more detailed results. Even if uncertainty is signaled only by high intonation, this is sufficient to be perceived as *uncertain.* The role of delay and fillers exclusively cannot be inferred from our data due to the design of our set of stimuli. It can only be said that, firstly, delay as an additional acoustic cue to high intonation does not yield a higher degree of perceived uncertainty. Secondly, our data suggest that the combination of fillers, delay, and high intonation have the strongest effect on the perception of uncertainty. However, from our data it is not clear how far this strongest effect is *purely* due to fillers.

4. Conclusions

The experiments presented in this paper present a first step towards the modeling of certainty and different degrees of uncertainty with the means of articulatory speech synthesis. Previous studies identified acoustical cues such as intonation, delay, and fillers in human-human dialog that convey uncertainty. Our study focused on the role of these cues in human-machine interaction.

The experiments brought to light that intonation by itself does contribute to the perception of uncertainty in articulatory speech synthesis in our test data. This is also true for the combination of all three cues. Further experiments are necessary, though, to determine how far the perception of uncertainty is purely influenced by fillers and delay, respectively. In contrast to previous studies, our data might suggest that delay by itself does not contribute to a stronger

perception of uncertainty. One should take into account, though, that listeners are presumably less sensitive to delays in our context since they expect from a machine that the response time is not as quick as from a human being. Future work could also consider the set of problems which are linked with the judgment of a machine's meta-cognitive state.

It can be well argued that the choice of the wordings (e.g. "ziemlich" as an adverb denoting vagueness) could also convey different levels of certainty in themselves. When further investigating *para*linguistic features conveying uncertainty, it seems useful to choose lexically more neutral wordings.

It would be interesting to run a cross-technique evaluation, since so far our cues only covered the purely acoustic domain – a domain that other kinds of synthesis could also cover.

As there is much evidence in the literature that prosody is not only conveyed by the acoustical channel but also by the visual one (e.g. [22]), we are planning to test unimodal and bimodal stimuli for several levels of certainty, finally making use of the three-dimensional vocal tract provided by the articulatory synthesizer.

5. Acknowledgements

We would like to thank Bernhard Schröder and Jürgen Trouvain for helpful comments. We are also grateful to Wolfgang Hess, Petra Wagner, Bernhard Fisseni, Stefan Breuer, Caja Thimm, Ulrich Schade, and Jürgen Trouvain for helping us with the experimental setup. Our thanks go, last but not least, to Peter Birkholz for giving us the chance to use the articulatory speech synthesis system, and to our speaker Tobias Ebel.

6. References

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

[2] Ang, J., Dhillon, R., Krupski, A., Shribers, E., Stolcke, A. (2002). "Prosody-based automatic detection of annoyance and frustration in human-computer dialog". In *Proceedings of ICSLP*, vol. 3, 2037-2040.

[3] Lee, M. and Narayanan, S. (2005). "Towards detecting emotions in spoken dialogs". In *IEEE Transactions on Speech and Audio Processing, 13* (2), 293-303.

[4] Litman, D. and Forbes-Riley, K. (2004). "Predicting Students Emotions in Computer-Human Tutoring dialogs". In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL),* 351-358, Barcelona, Spain.

[5] Ai, H. (2006). "Position paper". In *Online Proceedings of the Second Annual Young Researchers Roundtable on Spoken Dialogue Systems*. Pittsburg, PA. http://people.csail.mit.edu/alexgru/yrrsds/proceedings/yrrsds_proceedings06.pdf

[6] Burkhardt, F. (2001). *Simulation emotionaler Sprechweise mit Sprachsyntheseverfahren*. PhD Thesis, University of Berlin. Shaker Verlag.

[7] Schröder, M. (2004). *Speech and Emotion Research: An overview of research frameworks and dimensional approach to emotional speech synthesis*. PhD thesis, PHONUS 7, Research Report of the Institute of Phonetics, Saarland University.

[8] Ekman, P. (1972). "Universals and cultural differences in facial expressions of emotion". In Cole, J. (ed.), *Nebraska Symposium on Motivation 1971*, vol. 19, 207-283. Lincoln, NE: University of Nebraska Press.

[9] Schröder, M. and Trouvain, J. (2003). "The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching". In *International Journal of Speech Technology, 6*, 365-377.

[10] Wundt, W. (1896). *Grundriss der Psychologie*. Leipzig: Verlag von Wilhelm Engelmann.

[11] Sagisaka, Y. (1988). "Speech synthesis by rule using an optimal selection of non-uniform synthesis units". In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 679- 682.

[12] Black, A., Campbell, N. (1995). "Optimising Selection of Units from Speech Database for Concatenative Synthesis". In *Proceedings of Eurospeech, Vol 1*, Madrid, 581-584.

[13] Campbell, N., Black, A. (1996). "Prosody and the Selection of Source Units for Concatenative Synthesis". In Santen, J. van, Sproat, R., Olive, J., Hirschberg, J. (eds), *Progress in speech synthesis*, 279-282, Springer Verlag.

[14] Beskow, J. (2003). *Talking Heads – Models and Applications for Multimodal Speech Synthesis.* Doctoral Dissertation, KTH, Stockholm, Sweden.

[15] Lasarcyk, E. and Trouvain, J. (to appear). "Imitating conversational laughter with an articulatory speech synthesizer." To appear in *Proceedings of the Interdisciplinary Workshop on the Phonetics of Laughter*, Saarbrücken, August 2007.

[16] Lasarcyk, E. (to appear). "Investigating Larynx Height With An Articulatory Speech Synthesizer". To appear in *Proceedings of the 16th ICPhS*, Saarbrücken, August 2007.

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

[18] 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.

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

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

[21] Swerts, M., Krahmer, E., Barkhuysen, P. & van de Laar, L. (2003). "Audiovisual cues to uncertainty". In: *Proceedings of ISCA workshop on error handling in spoken dialog systems,* Chateau-d'Oex, Switzerland, August/September 2003.

[22] Massaro, D.W. (1998). "Perceiving Talking Faces: *From speech perception to a behavioral principle*. Cambridge, Massachusetts: MIT Press.

[23] Beskow, J. (2003). *Talking Heads – Models and Applications for Multimodal Speech Synthesis.* Doctoral Dissertation, KTH, Stockholm, Sweden.