Synthesizing Attitudes in German

Angelika Hönemann[1], *Petra Wagner*[1]

[1] Bielefeld University, Faculty of Linguistics and Literary Studies, Germany

ahoenemann@techfak.uni-bielefeld.de, petra.wagner@uni-bielefeld.de

Abstract

Our study investigates the potential of modeling the synthetic realization of four human attitudes (uncertainty, sincerity, surprise and doubt) based on a set of prosodic and voice quality parameters. A set of acoustic parameters were extracted from a corpus of German expressive speech. A comparison of lexically identical human and synthesized expressive utterances yields mostly positive correlations between the acoustic parameters used for analysis and modeling. In a subjective evaluation, listeners were asked to identify a target attitude in pairs of synthesized utterances. That way, uncertainty was identified in 90%, followed by sincerity (80%), surprise (72%) and doubt (64%).

Index Terms: speech synthesis, expressive speech, computational paralinguistics, prosody, voice quality

1. Introduction

High quality synthetic speech output is an indispensable attribute of any intelligent system such as a virtual agent or robot that uses speech-based communication when interacting with humans. Thus, speech synthesis research has begun to focus on the optimization of speech synthesis to fit the needs of speech-based Human-Machine Interaction (HMI) or dialogue systems [16]. Such interactions go beyond the mere exchange of words and/or factual information. In order to aid interactive grounding, comprehension and floor management, speakers use prosodic means to convey meta-information about the relevance, novelty and importance of what has been said, express information on their cognitive status (e.g. attention by feedback behavior), the ongoing speech planning process (e.g. by hesitations), floor management (e.g. by providing prosodic turn yielding cues) and emotions or attitudes related to the ongoing dialogue situation [15].

The expression of attitudes is a highly relevant factor in social interaction. Unlike emotions, they express the *affectively loaded cognitive appraisal* of a situation (or an object) [4]. We assume that the expression of attitudes can be of a short-timed, transitional nature and is likely to be ubiquitous in everyday communication. Hence, the expression of attitudes may be a crucial factor in HMI, making it more robust, as additional, nonverbal information is transported through the speech channel. E.g., a dialogue system could react to a low reliability of the speech recognition by expressing its subsequent reaction with the attitude of *uncertainty*, thereby implicitly making a confirmation request and critically reducing the number of necessary dialogue turns.

Previous studies have shown that attitudes are expressed through fine-grained adaptations of multiple acoustic prosodic and voice quality related parameters [5, 10]. Therefore, parametric rather than concatenative approaches to speech synthesis are probably suited best for its realization. However, to this day, dialogue systems tend to rely on concatenative approaches to synthesis such as unit selection or slot-and-filler systems, probably due to their high quality and because dialogue systems tend to operate in limited domains. It remains to be shown whether the potential benefits of attitudinal synthesis are strong enough to surpass the quality limitations introduced by parametric synthesis [8].

This paper presents a first feasibility study to explore the possibility of modeling and perceiving –often subtly expressed– attitudes with the help of adaptable parametric synthesis. In the remainder of this paper, we discuss results of an acoustic analysis of attitudinal expression based on German corpus data taken from previous work (section 2, [6]). In section 3, we describe the development of a set of rules for parameter adaptation in synthetic speech for four attitudes (sincerity, uncertainty, doubt, surprise). Section 4 describes the objective evaluation of the resulting attitudinal speech synthesis, section 5 presents the subjective evaluation based on a simple discrimination task. The paper closes with a discussion and a conclusion (section 6).

2. Data Analysis

The present analyses are based on a previously collected corpus of paralinguistic German speech [6]. The whole corpus consists of productions of two short utterances (*Marie tanzte*, *Eine Banane)* produced in 16 different attitudes by 20 native German speakers (11f., 9m). The full corpus contains a total of 640 utterance recordings. All utterances were force-aligned on phone level and SAMPA-transcribed using the Munich AUtomatic Segmentation system MAUS [12]. For each utterance, a set of acoustic parameters related to paralinguistic expression (F0, intensity, duration, jitter, shimmer) was extracted with [3]. The analyses showed that two of the selected attitudes are prototypically realized with a rising, "interrogative" contour (doubt, surprise), while two others tend to follow a falling, "declarative" contour (uncertainty, sincerity). These four attitudes were selected for further analysis and synthesis modeling. In total 80 stimuli (10 speakers * 4 attitudes * 2 utterances) were analyzed. Cross-speaker averages of these analyses are presented in Table 1 (human speakers).

3. Adaptable Synthesis

In order to realize the adaptation of synthetic speech according to the results of the acoustic analysis, a version of the MaryTTS system [14] embedded into the incremental speech processing system InproTK was used [1]. InproTK offers a 'just in time' modification of the speech parameters during the synthesis, thus, it can react immediately to dynamically

changing situations during an ongoing discourse, e.g. those that require an attitudinal reaction. InproTK provides modules to realize modifications of the synthesis output [2], but these are limited to the HMM synthesis offered by MaryTTS.

3.1. Acoustic parameter matching

We used the attitude specific mean values across human productions as input for calculating phone durations, fundamental frequency (F0), intensity as well as voice quality (VQ) parameters such as jitter and shimmer, since these acoustic parameters have been shown to be crucial for the perception of different attitudes [5, 9, 10].

This initialization is performed on phone level, while distinguishing between the phone classes of *vowels (V). consonants (C)* and *long vowels (LV)*. Additionally, a random factor ranging from zero to the standard deviation of each feature (RF) was added to the mean of each respective feature. This random factor simulates the measured speaker-specific variations in the resulting synthesized productions across the various attitudinal states. For the initialization, we defined the position of each phone and computed the percentage (PF) of the overall mean of an acoustic feature either for a phone at the first, middle and last position in a word or utterance or of a stressed phone. This allows for marking of stressed positions and stress related lengthening.

MaryTTS was used to generate the relevant acoustic parameters (F0, intensity, duration, phone duration) used for common synthesis. The parameters were then adapted by equations 1-5, based on the attitude-specific means of the various speech parameters for each phone (i), based on human productions [6]. Furthermore, interdependencies between the various acoustic parameters – especially between F0 and intensity – were derived from our empirical analyses. From these, we derived a set of heuristic rules used in the synthesis modeling: The exact rules are described in the equations below. Additionally, correspondences between F0 and intensity were modeled by adding the attitude-specific variability of intensity on F0 and vice versa. This leads to an increase of F0 or intensity based on attitude-specific variability in the corresponding acoustic domain.

Phoneme duration (Dur) The duration generated by MaryTTS (gDur) is shifted by a factor based on the sum of the duration derived from the human analysis (setDur) and the random factor of the duration (RF) multiplied by the position of the phone (PF(i)) divided by 100.

$$Dur_{i=1}^{n} += \frac{(setDur + RF_{dur})\ PF_i}{100}\ gDur_i \quad (1)$$

Intensity (Int) The intensity is based on the sum of the intensity derived from the human analysis (setInt) and the RF(s) of the intensity. The sum is multiplied by the phone's PF and added to the intensity.

$$Int_{i=1}^{n} += (setInt + RF_{int})\ PF_i \quad (2)$$

Fundamental Frequency (F0) The F0 is the sum of the F0 derived from the human analysis (setF0) and the RF of F0. The sum is multiplied by the phone's PF and added to the F0.

$$F0_{i=1}^{n} += (setF0 + RF_{F0})\ PF_i \quad (3)$$

Jitter (Jit) The jitter is the sum of the jitter derived from the human analysis (setJit) and the RF of the jitter. The sum is multiplied by the phone's PF and added to the jitter.

$$Jit_{i=1}^{n} += (setJit + RF_{Jit})\ PF_i \quad (4)$$

Shimmer (Shim) The shimmer is the sum of the shimmer derived from the human analysis (setShim) and the RF of the shimmer. The sum is multiplied by the PF of the phone and added to the shimmer.

$$Shim_{i=1}^{n} += (setShim + RF_{Shim})\ PF_i \quad (5)$$

3.2. Adaptation process

The general adaptation process is shown in Figure 1. It starts with the initialization of the *AdaptableSynthesisModule.* This module implements each phone as the *SysSegmentIU* class of the utterance and determines its position in a word and utterance. Furthermore the utterance mode is assigned.

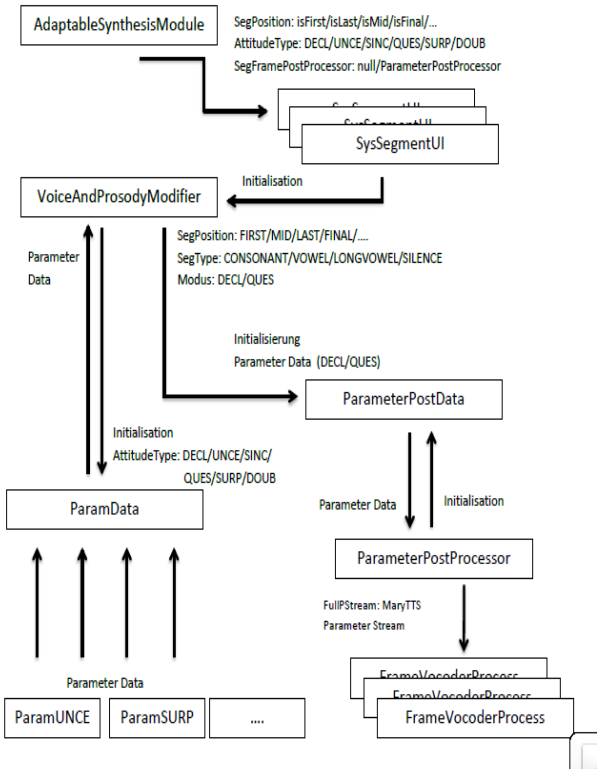

Figure 1: *Schematic diagram of synthesis process*

Each phone holds a class *VoiceAndProsodyModifier* including the equations for computing each feature value and sets the speech parameters for the current phone accordingly. This class receives the values for a specific attitude from the parameter class such as *ParamUNCE* for the uncertainty values or *ParamSURP* for the surprise values. All parameters are defined in these classes. Finally the *PostParameterData* container holds all relevant values for post processing. During post processing, the parameters of the MaryTTS HMM model for the common synthesis are adapted before the actual vocoding process starts. F0, intensity as well as the spectral information are computed for each frame. Each frame of a MaryTTS voice has a period of 5ms (200/sec).

Jitter and shimmer describe irregularities of the F0 (jitter) and the energy (shimmer) in the voice. The irregularity of F0 (cf. eq. 3) and intensity (cf eq. 2) are computed following the procedure in [9]: For each frame (i) we calculated a factor using the mean jitter derived from the human data (cf. eq. 4) as a multiplier to compute three sine waves, which are then added to each F0 value (cf. eq. 6, 7).

Table 1: *Means and standard deviation of the speech parameter for synthesized and spoken attitudes (across speaker and utterance) for five males (M) and five females (F)*

			Duration (ms)		F0 (Hz)		Intensity (dB)		Jitter (%)		Shimmer (%)	
		attitudes	mean	sd	mean	Sd	mean	sd	mean	sd	mean	sd
Human Speaker	M	sincerity	84	41	103.49	23.74	73.12	6.11	2.75	1.83	14.79	7.70
		uncertainty	102	58	96.65	20.94	70.32	6.53	3.00	2.43	12.99	8.89
		doubt	112	77	111.98	50.11	71.23	6.15	3.19	2.13	16.35	9.64
		surprise	115	75	127.10	58.53	72.84	6.43	3.59	2.86	13.14	6.32
	F	sincerity	94	61	188.75	38.11	72.60	8.10	2.91	2.35	13.88	7.59
		uncertainty	107	83	192.75	43.68	70.88	7.48	2.75	2.56	12.76	7.69
		doubt	112	71	196.61	65.75	70.39	6.51	3.53	3.08	15.40	10.20
		surprise	111	71	216.06	74.74	71.06	7.35	3.85	3.23	13.60	6.84
Synth. Adp. MaryTTS	M	sincerity	149	118	111.13	14.71	54.81	8.60	1.25	0.79	6.81	2.71
		uncertainty	160	160	110.21	14.97	54.75	9.29	1.47	2.15	7.24	5.47
		soubt	133	107	118.12	14.60	55.59	7.83	1.51	0.99	8.00	4.79
		surprise	121	85	117.84	14.48	55.10	8.58	1.43	0.88	7.42	3.78
	F	sincerity	149	120	150.72	21.95	63.67	8.05	1.27	1.15	8.20	5.32
		uncertainty	168	141	150.49	21.69	63.01	8.43	1.25	1.12	7.97	4.77
		doubt	189	163	172.78	15.63	64.97	8.57	1.42	1.85	7.52	5.73
		surprise	148	114	173.43	15.80	65.32	8.53	1.42	2.35	7.26	6.72

We used the same process for the intensity adaptation. A multiplier is computed using the shimmer yielded from the human data (cf. eq. 5, 9) to calculate the sine waves added to each energy value (cf. eq. 10). Finally each F0 and energy value within a frame is subtracted from the current mean of the phone to ensure smooth transitions (cf. eq. 8, 11).

$$FJ_{i=1}^{n} = setJit\ 100\ \pi(\frac{i}{200}) \quad (6)$$

$$F0_{i=1}^{n} += \sin(12.7FJ) + \sin(7.1FJ) + \sin(4.7FJ) \quad (7)$$

$$F0_{i=1}^{n} = \emptyset F0 - F0_i \quad (8)$$

$$FS_{i=1}^{n} = setShim\ 100\ \pi(\frac{i}{200}) \quad (9)$$

$$Eng_{i=1}^{n} += \sin(12.7FS) + \sin(7.1FS) + \sin(4.7FS) \quad (10)$$

$$Eng_{i=1}^{n} = \emptyset Eng - Eng_i \quad (11)$$

The vocoding process produces an audio stream on a frame-by frame-basis until the whole utterance is finished or the vocoding process is interrupted. The audio stream can be heard immediately, i.e. adaptation is simultaneous to the voice output.

4. Objective Evaluation

To compare the result of the attitudinal synthesis adaptation with the human productions, we synthesized a set of utterances directly comparable with the human data used in the analysis (5f, 5m, simulated by the random factor). We then extracted the identical speech parameters from the synthetic productions as in the analysis of the human productions. Table 1 shows the means and standard deviations of the two utterances produced by 10 human speakers and their synthetic counterparts. As there was no significant difference between individual utterances, means were calculated across utterances.

In most cases, the synthesized acoustic parameters for males and females are smaller than their corresponding human parameters. An exception to this is duration, i.e. synthetic speech tends to be slower. The following differences between the analyzed acoustic parameters for males (M) and females (F) can be observed: $\Delta\ Dur_M$=37.5, $\Delta\ Dur_F$=57.5, $\Delta\ F0_M$=9.15, $\Delta\ F0_F$=36.7, $\Delta\ Int_M$ =16.8, $\Delta\ Int_F$=6.9, $\Delta\ Jitter_M$=1.7, $\Delta\ Jitter_F$=1.9, $\Delta\ Shimmer_M$=6.9, $\Delta\ Shimmer_F$=6.2.

In order to get an estimate of the similarity between human and synthesized attitudes, we calculated correlations between two versions. For each acoustic parameter correlations are based on the mean values of each phone for both utterances (cf. Table 2). The tests yield high positive correlations for the majority of parameters, but a few marginal or even negative correlations in a few cases (displayed in red) indicate less fitting synthetic realizations.

Table 2: *Correlation coefficient for females (F) and males (M) for duration. F0, intensity, jitter and shimmer*

		Dur	F0	int	jitter	shim
M	sincerity	.74	.34	.78	-	.30
	uncertainty	.60	-	.71	.68	-.41
	doubt	.55	.75	.76	.37	-.09
	surprise	.68	.74	.79	.72	.41
F	sincerity	.47	.59	.74	.21	.19
	uncertainty	.32	-.55	.82	.44	-.07
	doubt	.86	.51	.79	.33	.27
	surprise	.83	.54	.75	-.12	.63

5. Subjective Evaluation

As the acoustic identification of attitudes is a difficult task even in human speech [6, 7, 11], a simple identification task was set up to assess the potential suitability of our approach. The evaluation was carried out with ten native German participants (5m, 5f). Each participant was asked to identify a (textually represented) target attitude out of a pair of two synthetic utterances representing different attitudes. A major discriminating feature of attitudes appears to be the global F0 contour (rising "interrogative": *doubt/surprise*; falling "declarative": *uncertainty, insecurity*). To exclude this all too obvious feature and to ensure that listeners need to take into account more subtle cues, only "interrogative" or "declarative" attitudes were compared with each other, i.e. "doubt vs. surprise" and "uncertainty vs. sincerity". Participants were allowed to listen each stimulus repeatedly. The utterance *Diese Banane ist gebogen* (engl. *This banana is*

bent) was synthesized with a male and a female voice for each of the four target attitudes and in five variations, using the random factor implemented in the synthesis strategy (see section 3). The variations simulate individual speaking styles. In total, our evaluation set contained 40 stimuli for identification, which were presented to listeners in 20 pairs. The stimulus pairs were assigned randomly within "interrogative" and "declarative" attitudes. The target attitude to be identified for each pair was selected randomly as well.

5.1. Results

The test yielded 50 identifications for each target attitude. Figure 2 shows the identification score (%) for each target attitude across subjects. Each participant identified the target attitudes better than chance level of 50%. Declarative attitudes were more convincing than interrogative ones. The best identification was found for uncertainty (90%, 45 of 50), the worst for doubt 64% (32 of 50), sincerity is identified in 80% (40 of 50) and surprise in 72% (36 of 50) of the cases.

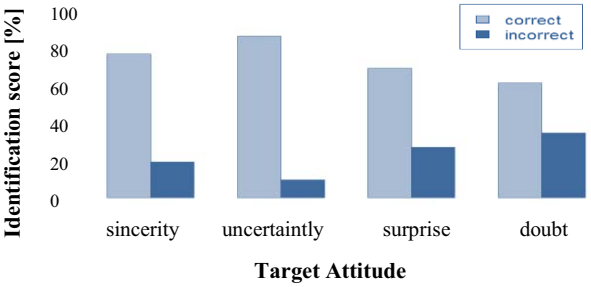

Figure 2: *Correct (lightblue) and incorrect (darkblue) identification score [%] of the synthesized attitudes across the subject*

6. Discussion and Conclusion

The current paper engaged in the parametric synthesis of four attitudinal states in German. Human recordings of attitudes have provided the empirical base for our synthesis strategy. We chose a rule-based approach because it offers a simple environment to identify and optimize the relevant parameters for an attitudinal synthesis and allows for a straightforward phonetic interpretation. The current work is a preliminary step for the later development of a model-based synthesis.

The usage of the unmodified results of the human analysis leads to a satisfactory simulation of the attitudinal states despite the obvious limitations of the HMM synthesis. The objective evaluation found that acoustic prosodic and voice quality parameters resemble those of the human originals. Simulating individual characteristics by introducing a random factor proved a successful approach.

The results of the subjective evaluation revealed that attitudes produced with a "declarative" contour were identified better than those with an "interrogative" contour. For now, we assume that the reason for this lies in the comparative proximity of *surprise* and *doubt* in function, form and their position in affective space [13]: *Surprise* and *doubt* share a rather high emotional activation, which has been shown to increase both F0 and intensity, while the declarative attitudes are more dissimilar: Uncertainty has a negative valence, while sincerity is considered as neutral. Furthermore, our results corroborate findings on the perception of attitudes expressed by human speakers, which have likewise shown that *doubt* and *surprise* can be less reliably identified in the absence of additional visual cues, i.e. facial expression [6, 7]. We therefore conclude that a less ambiguous synthesis of attitudinal behaviour needs to follow a multimodal approach.

7. Acknowledgements

This research was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

8. References

[1] Baumann. T.. & Schlangen. D. The InproTK 2012 Release. In Proceedings of NAACL-HLT. 2012

[2] Baumann T. Schlangen D. INPRO iSS: A Component for Just-In-Time Incremental Speech Synthesis. In: Proceedings of the ACL 2012 System Demonstrations. ACL: 103–108.. 2012

[3] Boersma, P., Weenink, D:: Praat: doing phonetics by computer [Computer program]. Version 5.3.51, retrieved 2 June 2013 from http://www.praat.org/, 2013

[4] Fazio, R.H & M.A. Olson. Attitudes: Foundations, Functions, and Consequences. M.A. Hogg & J. Cooper. The SAGE Handbook of Social Psychology (pp. 139-160), London: Sage, 2003

[5] Gobl C., Chasaide A.N., The role of voice quality in communicating emotion, mood and attitude. Speech Communication 40, p. 189–212, 2003

[6] Hönemann. A.. Mixdorff. H.. Rilliard. A.. Social Attitudes - Recordings and Evaluation of an audio-visual Corpus in German. 7th Forum Acusticum. Krakow. Polen. 2014

[7] Hönemann, A., Rilliard A., Mixdorff, H., Classification of Auditory-Visual Attitudes in German, FAAVSP - The 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing, Vienna, Austria 2015

[8] Huang, X., Acero, A. & H. Hsiao-Wuen. Spoken Language Processing. A guide to Theory, Algorithm, and System Development. Upper Saddle River, New Jersey: Prentice Hall.

[9] Klatt. D. & Klatt. L. Anaysis. synthesis. and perception of voice quality variations among female and male talkers. J. Acoust. Soc. America 87(2), 820-857, 1990

[10] Mixdorff. H.. Hönemann. A.. Rilliard. A. Acoustic-prosodic Analysis of Attitudinal Expressions in German. Proceeding of Interspeech 2015. Dresden. Germany. Page 1294, 2015

[11] Ricci B., Pio E., Luisa B., Paolo M., Roberto C., Pierluigi G., Expression and communication of doubt/uncertainty through facial expression. Journal of Theories and Research in Education, Ricerche di Pedagogia e Didattica, pp. 159-177, 2016.

[12] Schiel F, A statistical model for predicting pronunciation.. In: Proc. of the International Conference on Phonetic Sciences, Glasgow, United Kingdom, Paper 195, 2015

[13] Schauenburg, G., Ambrasat, J., Schröder, T., von Scheve, C., & Conrad, M. Emotional connotations of words related to authority and community. *Behavior Research Methods, 47*, 720-735, 2015

[14] Schröder. M. & Trouvain. J. The German Text-to-Speech Synthesis System MARY: A Tool for Research. Development and Teaching. International Journal of Speech Technology. 6. pp. 365-377. 2003

[15] Wagner, P. (What is) the contribution of phonetics to contemporary speech synthesis (?) D. Mehnert et al. (eds.). Systemtheorie, Signalverarbeitung, Sprachtechnologie. Studientexte zur Sprachkommunikation, Band 68 (pp.75–81.), TUD Press, Dresden, 2013.

[16] Ward, N. The challenge of modeling dialog dynamics. Workshop on Modeling Human Communication Dynamics at the 24th Annual Conference on Neural Information Processing Systems, Whistler, British Columbia, 2010.