Adaptive Speech Synthesis in a Cognitive Robotic Service Apartment: An Overview and First Steps Towards Voice Selection

Angelika Hönemann and Petra Wagner

Fakultät für Linguistik und Literaturwissenschaft, CITEC, Universität Bielefeld

ahoenemann@techfak.uni-bielefeld.de

Abstract: The Cognitive Robotic Service Apartment is both a realistic apartment and a laboratory environment in which the one or several user(s) interact with various manifestations of an intelligent agent e.g. a talking head. We expect that across various situational settings in the apartment, different specifications and adaptations of the synthetic voice will become necessary. Some of the dynamic adaptations will depend on physical factors e.g. ambient noise affecting speech intelligibility others on interpersonal factors e.g. familiarity and even others on the manifestation of the artificial agent itself e.g. the agent's voice, perceived gender, age and competence. It is the overall aim of our ongoing project to build a voice for a dynamically speech synthesis adaptation across various typical interaction scenarios and agent manifestations (robot, virtual agent). In the final implementation, the voice adaptation will be realized incrementally, i.e. the adaptation will be effected while talking. The adaptive synthesis module will be extended the existed incremental speech process system InproTK that is part of the cognitive architecture of the apartment. In order to determine an ideal set of adaptive parameters, a series of experiments is currently being planned and carried out. The paper will present our general methodology and describes our first study to find suitable synthesis voices for the virtual agent or humanoid robot used in the Cognitive Robotic Service Apartment.

1. Modeling Adaptive Speech Synthesis in the CSRA

Modeling the interaction between humans and machines remains a major speech technological challenge. This affects not only the interfaces between the different interacting system components (ASR, NLU, dialogue model, NLG, TTS) but each component individually. Our present project focuses on the improvement of a speech synthesis component in an interactive system in general, and on the situation-specific adaptation and modification of the synthetic speech output in particular. Such adaptations of the voice, driven by communicative purposes, are natural in humans and necessary in machines mimicking human speech communication. The present paper outlines how dynamically adaptive synthetic speech is realized in an ongoing research project as part of a complex interaction environment called the Cognitive Service Robotic Apartment (CSRA).

1.1 Situation-specific adaption in human speech production

An everyday example for situation-specific human speech adaptation has become famous as the Lombard Effect: Quite often dialogues between humans take place in noisy environments (outdoors in the presence of traffic noise; indoors with background music or with several people engaged in chatting simultaneously, e.g. in a pub). These conditions impede the intelligibility of the spoken content caused both by limited transmission quality and by the speakers' limited ability to self-monitor their voices. E. Lombard was the first researcher who discovered adaptation processes in speech produced under noisy conditions and his findings initiated a lot of subsequent research in this area [9]. His main observation was that self-monitoring is the regulator between speech production and perception and that lacking self-monitoring leads to an involuntary adaptation process to the environmental conditions, i.e. it leads to *Lombard Speech*. Many studies investigated the Lombard Effect from a medical or psychological perspective, but more recently, it has been investigated also from an acoustic, phonetic, linguistic and speech technological perspective. These studies could show that compared to speech in a quiet environment, Lombard Speech exhibits a decreased speaking rate, an increased fundamental frequency (F0) and range, a shift of intensity from low to high frequency, an increased vowel duration and a shift of F1 and F2 [6, 10]. However, the identified differences depend both on the speaker and the amount and type of ambient noise [7]. Lombard Speech adaptations occur spontaneously, immediately and unintentionally and thus have a different cause than phonetically similar, but intended adaptations such as the kind of speech addressed at an inattentive listener, a distant listener, a bad ASR, a listener with hearing problems, or a listener unaware of a potential danger. So far, very little is known about Lombard Speech occurring under real-life communicative conditions as it has mostly been investigated in monologue reading tasks. Still, it can be safely assumed that human-human communication certainly profits from Lombard Speech as its adaptations serve to improve intelligibility [6, 11]. Therefore, despite the fact that we cannot know precisely whether *intended* adaptations made for the cause of an improved intelligibility resembles Lombard Speech in all its facets, we make this simplified assumption in our ongoing study.

1.2 Adaptive interaction in the CSRA

Our project's interaction architecture is a Cognitive Robotic Service Apartment (CSRA). Unlike typical speech synthesis evaluations, this setting enables us to evaluate our adaptation strategies both under real-life and laboratory conditions. The former is possible as the human-apartment interactions are monitored permanently and across a wide range of everyday "university lab" situations such as demo tours, meetings or lunchtime chats in course of which individuals or groups interact with the interactive components both verbally and non-verbally. The verbal interactions will use different manifestations of intelligent agents such as a humanoid robot, a virtual agent or a disembodied apartment voice. Therefore, the agent's interaction strategy should suit various settings (information, service, interaction with a group, interaction with an individual, formal/informal settings) and their concrete manifestations (background music, quiet environment, attentive/inattentive user). We assume that the perceived interaction quality is at least to some extent influenced by the agents' overall voice quality and design as these factors are associated with characteristics such as perceived competence, trustworthiness, dominance, anxiety, reliability or credibility. Therefore, in a first step, a set of suitable voices and designs working across various types of artificial agents and situations needs to be determined. It is possible that the suitability of a voice is to some extent situation dependent, e.g. it might be more important to have a "competent" sounding voice in a formal situation where the agent explains something, while a "friendly, warm" voice might be more important in an informal situation where the agent welcomes the user.

1.3 Modeling adaptation in synthetic speech

In contrast to speech recognition systems [10, 8], the realization of Lombard Speech or similar types of environmental adaptation in synthetic speech synthesis is hitherto not well understood. This comes somewhat as a surprise as such adaptations can be expected to improve both intelligibility and perceived naturalness. Two potential adaptive strategies can be identified. One approach is the generation of an artificial voice trained with a different speaking style, e.g. a Lombard Speech corpus recorded in a noisy environment. Those methods produce speaker-dependent synthetic voices and require a large amount of training data [15]. Another strategy lies in the modification of an existing 'neutrally speaking' voice. Such adaptations are achieved via the modification of extracted speech parameters such as F0, energy or spectral characteristic and a subsequent re-synthesis. One advantage of this solution is that no new training data are required. More importantly, such an adaptation can be performed dynamically, speedy and incrementally, without the need to switch to a different “voice”. Such a dynamic, incremental type of adaptation to the situational needs models the automaticity of the Lombard Effect in humans (cf. above) and may therefore significantly contribute to the perceived naturalness of the resulting interaction, as has been previously shown for other aspects of verbal interaction in human-machine dialogues (cf. below). In order to objectively assess the intelligibility of the synthetic speech thus modified, several solutions were proposed in the literature, mainly based on human auditory system modeling (Glimpse Proportion, Dau model) and relying on the signal-to-noise ratio (SNR) [4].

1.4 Adaptation as part of incremental speech processing

In order to realize the situation-specific adaptations, the synthetic speech is realized within the speech-processing tool InproTK as part of the cognitive architecture of the CSRA apartment [1]. It includes a speech recognition module and a speech synthesis module and manages the speech input and output for the human-machine communication together with the dialog management tool Pamini [9]. Our speech adaptation module is based on the speech synthesis module using a modified version of the MaryTTS synthesizer [13]. This modification of the internal data structures was necessary to support the incremental processes offered by InproTK:

Incremental speech processing means that the system can react just-in-time to situational changes in speech, e.g. disfluencies, interruptions or other environmental changes both on the side of speech recognition and speech synthesis. This is reached by a step-by-step bottom-up process. Each utterance is split into chunks (**I**ncremental **U**nits), which can be phonemes, words or an entire phrase before handled. For any type of adaptation, this functionality is highly suitable because it allows prosodic changes of speech such as the intensity or loudness in course of the synthesis process. Many conventional text-to-speech systems are based on the sequential processing of utterances. That means, before a next sentence is processed, the previous sentence is synthesized completely. Such a traditional architecture allows adaptation only on a full incoming utterance, but not in course of an ongoing synthesis on its parts. An incremental architecture allows for more flexibility. For example, an incremental Lombard adaptation may continuously modify the synthetic speech output in the presence of steadily increasing background noise.

InproTK already includes first extensions of voice adaptation handling prosodic changes such as pitch, duration, loudness and spectral shifts of a MaryTTS HMM voice. These manipulations are carried out on the phone level. Furthermore, a demonstrator exists which provides the possibility to manipulate pitch, duration and loudness of a complex sentence during the synthesis process [2].

1.5 First steps towards voice selection

As a first step towards realizing synthetic speech within the CSRA, the general suitability of a set of synthetic voices for typical interaction situations is evaluated. As it is planned to carry out the situation-specific adaptations incrementally, only parametric synthesis voices can be used. To this end, three voices (two male, one female) already integrated in the MaryTTS synthesis architecture are part of this assessment. In addition, a newly created female voice is tested. The voice suitability is tested for two typical interaction scenarios (greeting, information) under both formal and informal conditions. The suitability is inferred from the perceived competence and warmth of the agent. Finally, we want to find out the best matches between existing voices and agent type (robot, virtual agent). The experiment and its main results are presented below.

2 Experimental Settings

We carried out a preliminary study in order to evaluate the general suitability of four different synthetic HMM-voices (2m, 2f) for the purpose of being used for the verbal interaction between humans and a virtual agent or humanoid robot within typical apartment interactions. 8 native German participants (age range 30-55 years, 4m/4f) took part in the experimental study. The evaluation software was implemented in JAVA and the experiment was carried out on individual personal computers in the participants' homes or quiet offices. The participants used stereo headsets to listen to the experimental stimuli.

2.1 Stimulus Design

The stimuli were built using a 5x2x4 design (5 prototypical interactions, 2 degrees of interaction formality, 4 voices). Each interaction comprised some kind of greeting and information presented by the synthetic voice, but were not varied systematically. Thus, the variation in interaction is not analyzed further, i.e. it is not used as a factor in the following statistical analyses. All interactions constitute prototypical situations for an embodied agent in our apartment, e.g. the agent greets the user after s/he enters the apartment and then provides the user with some daily news. These prototypical interactions were differentiated further in their degree of formality, either formal (FormUtt) or informal (InformUtt). Level of formality was differentiated mainly by varying the form of address: second person singular was used for the informal address ("duzen"), third person plural was used for the formal address ("siezen"). In German, informal "duzen" and formal "siezen" signal different degrees of familiarity and social distance. In addition to varying this type of address, lexical expressions were introduced to vary the degree of formality, e.g. "Hallo" rather than "Angenehmen Tag" as a less formal form of greeting. That way, formal and an informal utterances were created for each interaction. Each utterance was then synthesized with the help of the MaryTTS synthesizer. Three German MaryTTS HMM voices (male, male, female) and an additional female HMM voice created by ourselves were used. In the following, one prototypical interaction situation (Int1) with their respective informal (InformalUtt1) and formal (FormUtt1) utterances are presented:

> *Int1: Du betrittst dein Apartment und dein intelligenter Agent begrüßt dich und sagt, dass du es dir im Wohnzimmer bequem machen und dir von den Getränken auf dem Tisch nehmen kannst. Nachdem du im Wohnzimmer Platz genommen hast, erzählt der Agent von den technischen Neuigkeiten im Apartment.*
>
> *(engl. You enter your apartment and are greeted by your intelligent agent. The agent tells you that you can make yourself comfortable in the living room and that you can help yourself with the drinks prepared for you on the table. After you have taken place in the living room, the agent informs you about the newly installed technological equipment in the apartment.)*
>
> *InformUtt1: "Hallo. Mach es dir doch im Wohnzimmer bequem. Du kannst dir von den Getränken auf dem Tisch nehmen. Wie du siehst, sind hier ein paar technische Erneuerungen installiert worden. In deinem Apartment sind jetzt fünf Kameras und fünf Mikrofone für audiovisuelle Aufnahmen, zusätzlich sind Bewegungssensoren angebracht worden, die deine Position lokalisieren können."*

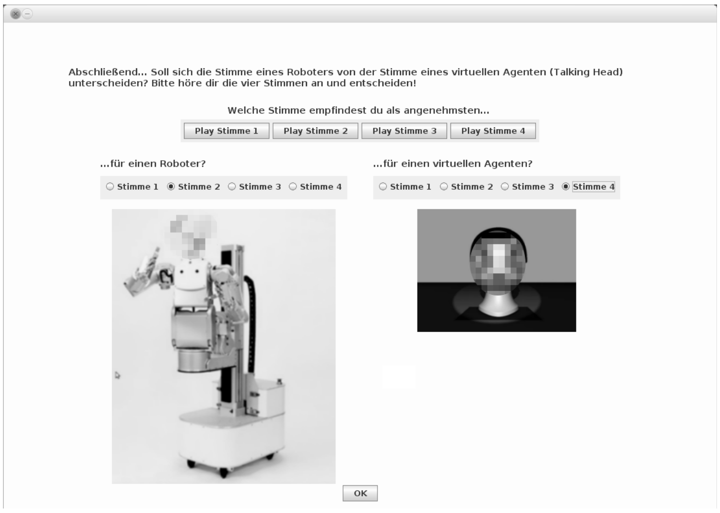
FormUtt1: “Angenehmen Tag. Machen Sie es sich doch im Wohnzimmer bequem. Sie können von den Getränken auf dem Tisch nehmen. Wie Sie sehen sind hier ein paar technische Erneuerungen installiert worden. In Ihrem Apartment sind jetzt fünf Kameras und fünf Mikrofone für audiovisuelle Aufnahmen, zusätzlich sind Bewegungssensoren angebracht worden, die Ihre Position lokalisieren können.”

2.2 Procedure

Prior to the experiment, each participant was informed about the experimental procedure by the experimenter. The experiment was comprised of two parts: In the first part, the participants were asked to judge the four voices with respect to their perceived competence and warmth across all four interactions. In a second part, the participants were asked to choose their preferred, or rather, the most pleasant voice for both a virtual agent and a humanoid robot acting as an interactive agent in an apartment. In total 44 stimuli were presented, 40 for in the first part and four in the second part of the experiment.

The experiment consisted of the following steps: (1) A prototypical situation was presented to be read by the subjects. (2) The five interaction specific recordings were presented one after the other and evaluated. That is, per interaction, the participants listened to a trial of 4 voices in a row. The participants were allowed to listen to each stimulus only once. (3) After listening, each stimulus was directly judged without delay for its degree of a) competence and b) warmth. When all voices were thus rated for one interaction, the next interaction scenario was introduced. (4) After all interaction-specific stimuli had been judged, the second part of the experiment started. Here, participants listened to each artificial voice again but the utterance they listened to did not contain any formality or interaction specific features. After listening to these four stimuli, the participants were asked to choose most suitable voice for both the artificial agent and the humanoid robot. (5) The experiment ended.

Both the order of the presented situations and the order of stimuli presented within each situation were randomized. The participants rated their impressions of warmth and competence by using a magnitude scaling technique. That is, participants were completely free in their assignment of the perceived strength. Instead of a predefined scale, they could choose any numeric value for their impression. That way, each subject could use his/her individual range and adapt their internal scale in course of the experiment according to their impressions.



Abschließend... Soll sich die Stimme eines Roboters von der Stimme eines virtuellen Agenten (Talking Head) unterscheiden? Bitte höre dir die vier Stimmen an und entscheiden!

Welche Stimme empfindest du als angenehmsten...

...für einen Roboter?

○ Stimme 1 ● Stimme 2 ○ Stimme 3 ○ Stimme 4

...für einen virtuellen Agenten?

○ Stimme 1 ○ Stimme 2 ○ Stimme 3 ● Stimme 4

Figure 1: Screenshot of the second part of the experiment where the participants chose the preferred voice for the virtual agent and humanoid robot currently used in the CSRA project.

Figure 1 illustrates the screen shown to the participants in this final decision process. To get a better impression of the agents' dimensions and generic design features, images of the virtual agent and the humanoid robot currently used in the CSRA project are shown. Their faces were distorted, as the participants should not be influenced by their specific facial designs – these are currently evaluated and developed further in a simultaneous study.

3 Results and Discussion

All participants carried out the complete experiment, yielding 640 valid judgments from the first part for the experiment and 16 "voice votes" from the second part of the experiment. Because of the participants-specific unbounded judgment scale, the computed means of each participant's answers from the five trails for each question (degree of warmth, degree of competence) and for each utterance (formal and informal) were z-score normalized.

The results of the voices' judgment across participants and speaking style reveal that the participants favor two voices, the MaryTTS *bits1-hsmm* voice (female) and the MaryTTS *bits3-hsmm* voice (male). Figure 2 shows the mean distribution of the ratings for the four voices with respect to perceived 'warmth' and 'competence'. A One-Factorial ANOVA revealed significant differences between voices both for both perceived warmth (F=19.23, p=4.565e-05, p<0.001) and competence (F=33.66, p=2.40e-07, p<0.001).

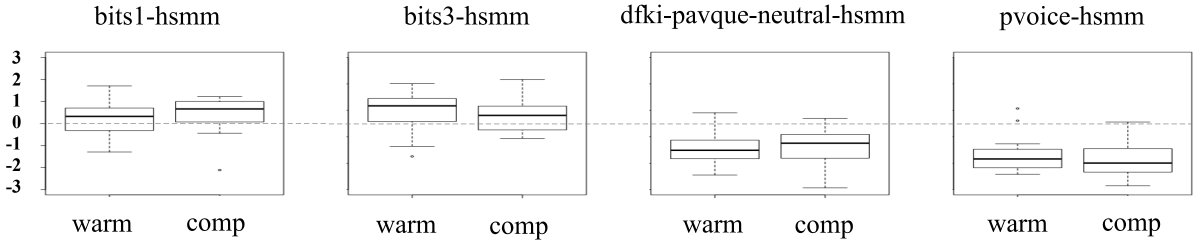


Figure 2: Mean judgments of the four voices across the participants and across the degree of formality

Figure 2 also indicates that the perceptions of warmth and competence are not independent. This impression is corroborated by a linear regression analysis on the judgments for warmth and competence yielding a linear relationship (R^2=0.51, p<0.0001).

A Two-Factorial ANOVA yields significant differences of the factor voice but not for formality in relation to the judgments for warmth (F=18.95, p=4.328e-11, p<0.001) and competence (F=34.44, p=1.941e-07, p<0.001). Thus, verbally expressed degree of formality did not evoke any effect on perceptual warmth or competence expressed by the voice characteristics.

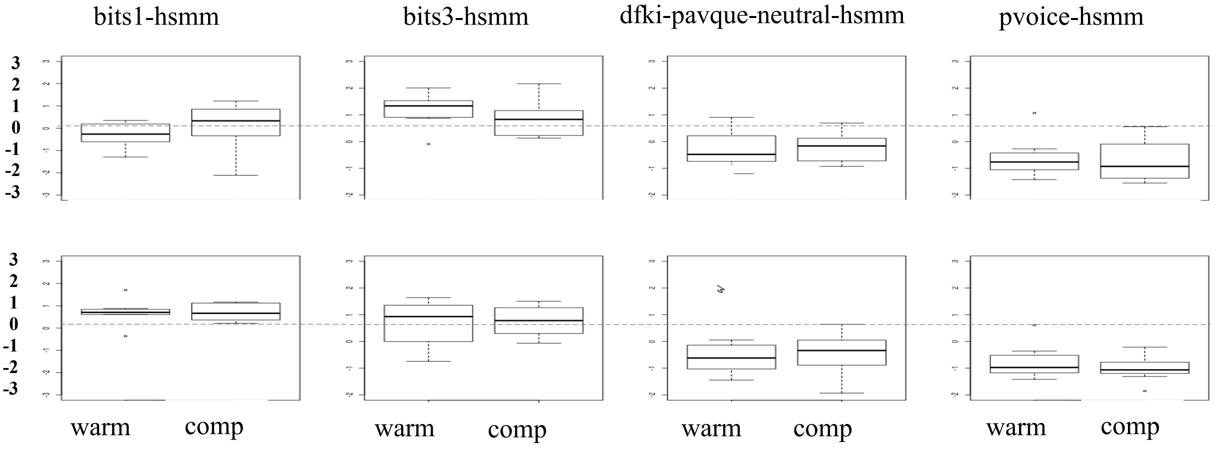


Figure 3: Mean judgments for the four voices in the 'formal' (above) and 'informal' (below) condition across participants

Figure 3 summarizes the results of the relation between the degree of formality and perceived voice quality. Due to the missing statistical effect of formality, the representations of the foregoing Figures 2 and 3 look very much alike. The *bits1-hsmm* and the *bits3-hsmm* voices are perceived as both more warm and most competent than the *dfki-pavque-neutral-hsmm* and the *pvoice-hsmm* voice. Despite the missing statistical effect, the *bits1-hsmm* voice was rated as 'warmer' in the informal setting. A similar effect can be found for the *bits3-hsmm* voice, but here the formal condition led to a higher judgment of 'warmth'. A negative influence of the formality condition can be seen for the *pvoice-hsmm* voice. More participants perceived this voice as less competent in the informal condition. This mixed picture indicates that – despite a missing statistical effect– the influence of formality might be to some degree depending on the voice and probably needs further investigation.

The aim of the second part of the experiment concerned the general question which voice is most suitable for a robot and/or a virtual agent. Table 1 lists the participants' selections.

Voice/Agent	Humanoid Robot	Virtual Agent
bits1-hsmm	0	1
bits3-hsmm	3	3
dfki-pavque-neutral-hsmm	3	2
pvoice-hsmm	2	2

Table 1: Distribution of the participants' choices of a most pleasant voice for a humanoid robot and a virtual agent

The evaluation did not yield any clear preference and does not allow for a statistical analysis due to the low number of participants. Still, an interest point is that it contradicts the results of the first part of the experiment where the voices *dfki-pavque-neutral-hsmm* and *pvoice-hsmm* were perceived as both less 'warm' and less 'competent'. Nevertheless, nine of the 16 choices favor these voices. Just as surprising is, that the *bits1-hsmm* voice could neither convince as suitable for the robot nor for the virtual agent despite it being perceived as 'warm' and 'competent' in the first part. Still, the *bits3-hsmm* remains to be the most convincing one for both parts of the experiment. Even if we only have a few data points, they point out that the suitability of a voice for a multimodal application cannot be interpreted without its visual characteristics. Besides, the results show that the quality features of 'warmth' and 'competence' might not be the most reliable indicators of a voice's suitability for a speech technological application.

4 Conclusion

Our paper illustrated the ongoing project on building a dynamically adaptive speech synthesis module that enables the interaction between a virtual agent or a humanoid robot with a human interlocutor in a real-life apartment context. A first preliminary analysis identified a preferred synthetic voice for this endeavor, but also showed that synthesis quality heavily interacts with the agent's appearance: Context-free tests do not appear to be an adequate means to select a suitable agent voice. The degree of interaction formality did not substantially influence the perceived synthesis quality. Lastly, it needs to be said that our very preliminary tests need to be continued with further participants to consolidate our findings.

References

[1] Baumann T, Schlangen D (2012), The InproTK 2012 release, In: Proceedings of the NAACL-HLT Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data (SDCTD 2012). ACL: 29–32.

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

[3] Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., & Tang, Y. (2013). Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Communication*, *55*(4), 572-585.

[4] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Can objective measures predict the intelligibility of modified HMMbased synthetic speech in noise?," in Proc. Interspeech, Florence, Italy, August 2011.

[5] Damjan Vlaj and Zdravko Kacic (2011). The Influence of Lombard Effect on Speech Recognition, Speech Technologies, Prof. Ivo Ipsic (Ed.), ISBN: 978-953-307-996-7, InTech, DOI: 10.5772/17520.

[6] Folk L., Schiel F., The Lombard Effect in Spontaneous Dialog Speech. INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011

[7] Garnier, M. & Henrich, N. (2013). Speaking in noise: how does the lombard effect improve acoustic contrasts between speech and ambient noise? Computer Speech & Language, 28, 580–597

[8] Hirsch H. G., Pearce D., 2000 The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions, Proceedings of the ISCA ITRW ASR'00, Paris, France.

[9] Lombard, E. (1911). Le signe de l'elevation de la voix, Annals maladiers oreille, Larynx, Nez, Pharynx, Vol. 37, pp. 101-119.

[10] Lu, Y., and Cooke, M. P. 2008. "Speech production modifications produced by competing talkers, babble and stationary noise," J. Acoust. Soc. Am. 124, 3261–3275

[11] Lau, Priscilla. "The lombard effect as a communicative phenomenon." *UC Berkeley Report* (2008).

[12] Peltason J., Wrede B., "Pamini: A framework for assembling mixed-initiative human-robot interaction from generic interaction patterns" in Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 229–232, The University of Tokyo, September 24-25, 2010.

[13] Schröder M., Trouvain J. (2003). The German text-to-speech synthesis system MARY: a tool for research, development and teaching. International Journal of Speech Technology, 6, 365–377

[14] Suni A., Karhila R., Raitio T., Kurimo M., Vainio M., and Alku P., "Lombard modified text-to-speech synthesis for improved intelligibility: Submission for the Hurricane challenge 2013, " in Proc. Interspeech, 2013, pp. 3562-3566.

[15] Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J., 2009a. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans. Audio Speech Lang. Process. 17, 66–83