Online Lombard-adaptation in incremental speech synthesis

Sebastian Rottschäfer, Hendrik Buschmeier, Herwin van Welbergen, Stefan Kopp

Social Cognitive Systems Group – Faculty of Technology and CITEC
Bielefeld University, PO-Box 10 01 31, 33501 Bielefeld, Germany
{srottschaefer,hbuschme,hvanwelbergen,skopp}@uni-bielefeld.de

Abstract

The 'Lombard effect' consists of various speech adaptation mechanisms human speakers use involuntarily to counter influences that a noisy environment has on their speech intelligibility. These adaptations are highly dependent on the characteristics of the noise and happen rapidly. Modelling the effect for the output side of speech interfaces is therefore difficult: the noise characteristics need to be evaluated continuously and speech synthesis adaptations need to take effect immediately. This paper describes and evaluates an online system consisting of a module that analyses the acoustic environment and a module that adapts the speech parameters of an incremental speech synthesis system in a timely manner. In an evaluation with human listeners the system had a similar effect on intelligibility as had human speakers in offline studies. Furthermore, during noise the Lombard-adapted speech was rated more natural than standard speech.

Index Terms: Speech synthesis, Lombard effect, Speech intelligibility, Incremental processing, Adaptation, Interactive systems

1. Introduction

When people talk in a noisy environment, such as a busy pub, their speech inevitably gets masked by other conversations, music, ambient sounds from the bar, and so on. To make themselves better understood to their conversation partners, speakers constantly try to counteract the effect of masking by adapting their speech, taking the characteristics of the noise into account. For example (see [6] for a comprehensive review), they increase the intensity of their speech to increase the signal to noise ratio; they shift speech energy to higher frequency bands where the ear is most sensitive; or use exaggerated pronunciation of phonemes or words to make them less ambiguous and more recognisable. This behaviour is known as the 'Lombard effect' [11].

Although many of the individual adaptation mechanisms constituting the Lombard effect are known, models of these mechanisms are often underspecified, in the sense that it is not clear how exactly they react to (different) noise conditions. Similarly, the degree to which each of them contributes to speech intelligibility is still under discussion. It is also largely unknown how individual adaptation mechanisms interact.

Lately, researchers started to use synthesis-based approaches to study how Lombard-models influence intelligibility [6, 7]. Advanced models even work with non-stationary background noise and change the synthesis parameters in parallel to changes of the noise. Most of these models only work 'offline', i.e., they need to know the background noise before synthesis.

This paper presents and evaluates a first system for 'online' Lombard speech synthesis. It analyses the auditory environment continuously and updates the Lombard model parameters accordingly. These are then used in an incremental speech synthesis system (INPRO_iSS, [2]), which immediately incorporates the model parameters into its synthesis process. In contrast to offline-models of Lombard speech, such an online system is suitable for use in interactive systems – for example conversational agents in spoken dialogue systems – that operate in continuously changing and hard to predict acoustic environments, such as for example in cars or in industrial settings. Constructing online models of Lombard speech poses new and interesting challenges – for speech science as well as for research on speech synthesis – that are not obvious from offline models.

2. Background

2.1. Lombard speech

For some adaptation mechanisms in Lombard speech, research already provides good findings of the behaviour of speech parameters in changing noise conditions. In general, however, the exact behaviour of many speech parameters, especially their interaction, is not yet understood very well. For the most obvious adaptation parameter, speech intensity, it is known that human speakers increase the level of their voice linearly (simultaneously emphasising higher frequencies), but only half as much, with the perceived noise level [10]. This relatively small increase is sufficient to compensate for the noise as multiple perception channels are used during speaking and listening. This parameter is known to be most effective (it straightforwardly increases the signal-to-noise ratio) and well understood.

As a more advanced countermeasure, speakers increase the fundamental frequency of their voice when noise is present [3]. This adaptation is done in a way that is sensitive to the exact noise context. Human speakers temporarily shift their fundamental frequency to a detected minimum in the noise spectrum [8]. Furthermore, the variation in fundamental frequency increases up to about 0.8 tones. It is, however, not yet clear how the increase in variation interacts with changes in noise [8].

Another adaptation observed in Lombard speech is speech rate [21]. As many factors like phone duration and count and duration of stops influence the rate of speaking, there is no simple and universal measure for change in speech rate [15]. In general, it is known that the duration of specific sentences increases by around 30% in hyper-articulated speech [15], which is very similar to Lombard speech [19]. Which noise level or type of noise produces which changes in speech rate is, again, not yet well understood. Further speech parameters, such as formants or consonant–vowel ratio, are also adapted in Lombard speech, but will not be modelled in this work. A comprehensive survey of known adaptation mechanisms used in Lombard speech is provided by Cooke and colleagues [6].

2.2. Adaptive speech synthesis

Research has embraced the possibility to use synthetic speech to model and evaluate Lombard-adaptation mechanisms. With

synthetic speech it is possible to control both the model-based manipulation and the underlying speech itself, making the process reproducible and objective in the sense that certain noise characteristics always result in the same adaptation.

Synthesis-based Lombard speech usually results from specific Lombard-adapted voice-models that are obtained by deriving them from a corpus of human Lombard speech or by modifying parameters of a non-Lombard adapted synthetic voice such that it exhibits Lombard-like properties afterwards. Valentini-Botinhao and colleagues, for example, do the latter by modifying the Mel cepstral coefficients by optimising the distribution of spectral energy according to a 'glimpse proportion' measure [23]. Systems that use such specific voice-models for Lombard speech have the limitation that they are not optimised for different types of noise and are therefore suboptimal for systems used in changing noise conditions.

A more flexible system for such conditions is C2H, which implements a feedback loop that optimises vowel and consonant space given the perceived environmental noise. This system is not specifically generating Lombard speech, but is meant as a computational model of hypo- and hyper-articulated speech [14].

Both systems took part in the 'Hurricane challenge' [7], which aims at making synthetic speech as intelligible as possible given a predefined noise sequence. While the challenge provides an objective intelligibility rating for participating systems, it does not measure their reactivity to noise, nor the human-likeness of the resulting speech.

Online adaptation of synthetic speech has also been used in different applications. Ström and Seneff, for example, have implemented a turn-keeping strategy which increases loudness using a pre-emphasis filter that specifically boosts the higher frequencies of speech [20].

2.3. Incremental speech synthesis

Standard speech synthesis systems take text (e.g., a sentence) as input, encode it into a speech signal, and then return the audio signal to be delivered as output [22]. It is not possible to change the speech output once it is encoded. This is problematic in interactive applications, where speech output may need to be altered quickly depending on factors that are not under the system's control and thus hard to predict (e.g., user actions or noise in the environment), or when it is not yet possible to fully specify the input text (e.g., when the environment is changing).

A solution to this problem is an 'incremental' approach to speech synthesis. In incremental processing, a system is "triggered into activity by a minimal amount of its characteristic input and produces characteristic output as soon as a minimal amount of output is available" [9, p. 70]. An incremental speech synthesis system thus needs to be able to take partial input (e.g., the first few words of a sentence) and to produce initial audio frames (rather than the complete audio signal) as output. While delivering the audio frames already encoded, the input to an incremental speech synthesis system may be extended, and not yet delivered parts may be revised or re-synthesised with changed parameters.

The system presented in this paper builds upon the incremental HMM-based speech synthesis system INPRO_iSS [2], an incremental version of MARYTTS [18]. INPRO_iSS reduces lookahead context on each internal level of processing as much as possible (with minimal impact on synthesis quality) and changes processing to produce only as much output as is needed on the next lower level for a 'just-in-time' delivery of speech [1]. Changes to a unit (e.g., a word, a phoneme) of synthesis can thus be requested until shortly before its actual delivery.

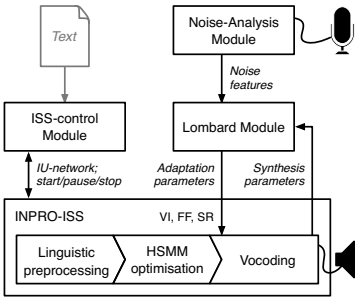

Figure 1: Schematic architecture of the online Lombard speech synthesis system. Based on a real-time analysis of the characteristics of the acoustic environment, the Lombard Module adapts the synthesis parameters in the vocoder of the incremental speech synthesis system INPRO_iSS while the system is speaking.

The way incremental processing is realised in INPRO_iSS, via the 'IU'-framework for incremental processing [17], makes it highly suitable for modelling online Lombard-adaptation. Exchanging a word for another – a high-level adaptation – is possible until shortly before the word is delivered. Changing the pronunciation of a phoneme (e.g., by lengthening) is possible until the phoneme is delivered. Low-level adaptations, such as changes to speech intensity, are even possible immediately.

INPRO_iSS realises changes in speech rate by simply skipping or repeating vocoding parameters (for shortening or lengthening, respectively). Speech intensity changes are realised by boosting the speech energy and simultaneously emphasising higher frequencies, via post-processing of the vocoding parameters, which results in a decrease of spectral tilt. Changes are only possible in certain limits as the synthetic voices are not constructed with highly flexible adaptation in mind.

3. System

The system for online Lombard speech synthesis consists of three main modules: the noise-analysis module, the Lombard module, and the incremental speech synthesis system INPRO_iSS (see Figure 1). Additionally, the speech synthesis is controlled by the ISS-control module that provides incremental input (text in form of incremental unit networks, cf. [5]), and starts, pauses, and stops speech output. The noise-analysis module and the Lombard module run concurrently to the speech synthesis. This way, the Lombard module can continuously adapt synthesis parameters according to the acoustic environment and the implemented model of Lombard speech. The following sections describe the noise-analysis module and Lombard model/module.

3.1. Noise-analysis module

The noise-analysis module is connected to a microphone that records with a sampling rate of 44.1 kHz. It characterises the noise context in terms of (1) the sound pressure level (SPL) and (2) the frequency with minimum amplitude. SPL is calculated via the root mean square method. To find the global minimum amplitude in the spectrum the Fourier transform for the limited frequency bandwidth that is in reach of the female human voice (175–275 Hz; [8]) is computed (we use a female synthetic voice). To smooth these features, frames with the size of 2^{14} (i.e., 16k) samples (0.37 s), with an overlap of 75%, are used. This allows for a higher update frequency without losing time resolution of the Fourier transform. Furthermore, to enhance the accuracy, each frame is multiplied with the Hann function. Both features are passed on to the Lombard model.

3.2. Lombard model and module

The Lombard module takes the features that characterise the noise context, computes adjustments for speech parameters according to the Lombard-model, and passes the adjusted parameters on to the incremental speech synthesis module INPRO_iSS. To avoid abrupt changes in speech adaptation, the noise level feature is smoothed again. For SPL this is done by using the mean $\text{SPL}_{\text{noise}}(t) = \text{Mean}(\text{SPL}_t, \ldots, \text{SPL}_{t-9})$ of the last 10 values instead of the value SPL_t of the current frame only. As a result, the system needs around one second to fully accommodate a change in the noise level. In the following, the adjustment of speech parameters used in our system is explained in detail.

Voice intensity and spectral emphasis In general, the increase in voice intensity and shift of spectral energy only depends on the sound pressure level of the noise [10]. The ability to raise the voice intensity in INPRO_iSS, however, is limited to a maximum increase $\Delta I_{\max}$ of about 7–9 dB (see Section 2.3). Following Lane [10], the intensity gain at each point of time in our model is

$$I(t) = \frac{\text{SPL}_{\text{noise}}(t) - \text{SPL}_{\text{base}}}{2\Delta I_{\max}},$$

where SPL_{base} is the base sound pressure of the environment (set to 46 dB) and $\Delta I_{\max}$ is the maximum increase in sound pressure the system can counter (assumed to be 7.5 dB). Tests showed that the sound pressure level of the voice in INPRO_iSS increases roughly linearly with its 'energy' parameter (which ranges from 0 to 190) in the frame post-processor of the vocoder and is thus set to $I(t) \cdot 190$.

Fundamental frequency The long-term progression of the fundamental frequency is adapted according to the findings of Garnier and Henrich [8], i.e., the fundamental frequency raises with the voice intensity. Since the synthetic speech starts to sound unnatural above a fundamental frequency of 210 Hz, our system is limited to a maximum f_0-shift ($\Delta f_{0\max}$) of 20 Hz. This is much less than f_0-shifts observed in human Lombard speech (up to 100/200 Hz for male/female speakers, [8]).

Garnier and Henrich also found that humans shift the fundamental frequency towards minima in the noise spectrum – if present and reachable by the voice [8]. The presence of such a minimum is determined by comparing the variance of the last 50 minima $\mathbf{f}_{min} = [\text{argmin}_f(A_t(f)), \ldots, \text{argmin}_f(A_{t-49}(f))]$ with a threshold θ (set to 25 Hz in our system), where $A_t(f)$ is the amplitude of frequency f in the spectrum of frame t. Combining these two cases, the shift of the fundamental frequency $\Delta f_0(t)$ is thus modelled as

$$\Delta f_0(t) = \begin{cases} \text{Mean}(\mathbf{f}_{min}) & \text{if } \text{Var}(\mathbf{f}_{min}) < \theta, \\ \Delta f_{0\max} \cdot I(t) & \text{otherwise.} \end{cases}$$

In addition to the long-term progression, short-term variation in fundamental frequency is a second characteristic that is modelled in our system. Based on the finding that the standard deviation of the fundamental frequency increases in noisy conditions [8], the difference between the fundamental frequency and the frequency of each voiced frame is multiplied with a factor $\alpha(t)$ that only depends on the intensity gain $I(t)$:

$$\alpha(t) = 1 + \alpha_{\max} \cdot I(t)$$

Here, $\alpha_{\max}$ is the factor which increases the deviation of the fundamental frequency by 0.8 tones, a value measured in human Lombard speech [8]. The fundamental frequency parameter is modified in the frame post-processor of INPRO_iSS's vocoder.

Speech rate The Lombard model adjusts the speech rate linearly with changing noise levels. As the synthetic speech starts to sound unnatural below a speech rate of 85%, a maximum decrease of 0.15 – again less than reported for human Lombard speech in the literature [12, 15] – was set. This is due to the fact that the synthetic voice does not natively support lengthening and that the model of how speech rate influences phoneme and pause durations in INPRO_iSS (see Section 2.3) does not result in more frequent and longer pauses typical in human Lombard speech [15]. Further, more complex phenomena such as influences on glottal stops, consonant–vowel ratio, or phoneme reduction are not modelled in the system, yet. In our model, the adjusted speech rate at each point of time t is set in INPRO_iSS' vocoder to be

$$\text{SR}(t) = 1 - 0.15 \cdot I(t).$$

4. Evaluation

To evaluate the speech intelligibility and naturalness of the online Lombard speech synthesis and the different adaptation models presented above, a listening study was conducted. In a within-subject experiment, participants listened to segments of a German short story [16] synthesised in real-time with different degrees of Lombard-adaptation in non-stationary noise conditions. After each segment of the story, participants assessed how natural the adaptations to the changing conditions were and how intelligible the synthetic voice was. In contrast to the objective Lombard speech intelligibility evaluation procedure used, e.g., in the Hurricane challenge [7], a subjective rating measure is used here. Although this is potentially less accurate, it enables an evaluation of longer stimuli that allow significant changes of background noise to occur within one stimulus.

4.1. Setup

Participants sat in front of a computer that ran the online Lombard speech synthesis system, controlled the experiment, and displayed the evaluation questionnaires. Non-stationary background noise was played from two speakers (M-Audio AV40 Studiophile) set up behind the computer and recorded with a microphone (Samson Go Mic USB) that fed into the noise-analysis module. Speech, synthesised in the German female MARYTTS-voice 'bits1-hsmm' with default settings, was played to the participants via 'open-back' headphones (Turtle Beach PX22). This setup made it possible for participants to hear both the voice as well as the background noise, without creating a feedback loop between the speech output and the noise-analysis module.

A multi-babble noise from a bar [13] was used as the background noise and its loudness level was systematically changed between approximately 46 dB and 60 dB during play-back. Periods of low and high intensity were between 3–7 s long (sampled from a uniform random distribution) with 0.5 s linear transitions.

The short story was segmented into 32 parts, each between 20–30 s long (with the exception of the final segment, which had a length of approximately 50 s). For each participant, each segment was assigned to one of the following four synthesis conditions, reflecting different degrees of Lombard-adaptation:

NA The speech parameters were not adapted with changing noise conditions

VI Voice intensity and spectral emphasis was adapted

FF+SR Fundamental frequency and speech rate was adapted

VI+FF+SR Voice intensity and spectral emphasis, fundamental frequency, and speech rate was adapted

Assignment to conditions was random, but balanced such that for every set of four participants each segment is paired with all conditions and each condition occurs with the same frequency.

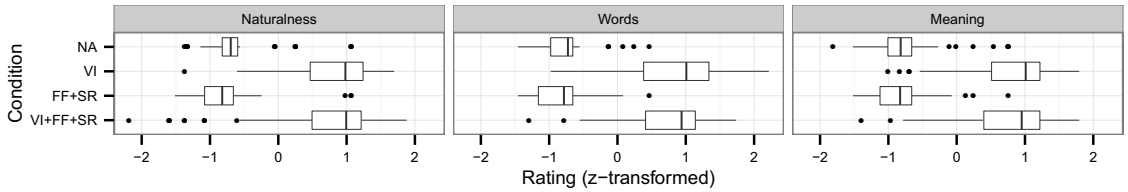

Figure 2: Boxplots of participants' ratings of the items naturalness, understood words, and understood meaning by synthesis conditions (NA: no adaptation; VI: adaptation of voice intensity and spectral emphasis; FF: adaptation of fundamental frequency; SR: adaptation of speech rate). Scores are z-transformed by participant and item, higher scores are better.

4.2. Procedure

Participants read a written description of the study, were able to ask questions, and, based on this information, consented to participate. They were then seated in front of the computer, put on the headphones and listened to the synthetic voice providing more detailed instructions. Participants were told to regard this voice as the baseline for later assessments.

After instructions were given, the study was self-guided. Participants started each segment of the short story with a mouse-click. The background noise began playing with the start of the speech output and ended together with the segment. After each segment participants assessed the synthetic speech by rating a single item on naturalness and two items on intelligibility on a 7-point Likert-scale displayed on screen:

Naturalness How do you rate the naturalness of the synthetic voice in increasing noise conditions? (unnatural—natural)

Words I understood every word (none—every)

Meaning I understood the meaning (not at all—fully)

They then proceeded to the next segment and were debriefed after the final segment.

4.3. Results and discussion

Ten students and employees from Bielefeld University volunteered to participate in the evaluation study. The data of two of them had to be discarded due to technical difficulties with the system (as this was noticed immediately, the balance over the conditions could be maintained nevertheless). All participants were native speakers of German and reported normal hearing capabilities. Participants received a sweet for their participation.

The ratings of each participant for each of the three items was z-transformed to allow for easier comparison between participants. The results are visualised in Figure 2. As can be seen, the VI condition (voice intensity and spectral emphasis based Lombard-adaptation) as well as the VI+FF+SR condition (voice intensity and spectral emphasis, fundamental frequency and speech rate based Lombard-adaptation) were consistently rated higher (i.e., higher perceived naturalness and speech intelligibility) than both the NA condition (with the normal, non-adapted voice) and the FF+SR condition.

The adaptation of fundamental frequency and speech rate, however, had no effect on naturalness or intelligibility: The FF+SR condition is rated similar to the NA condition (albeit with a larger interquartile range and longer whiskers, which could be interpreted as uncertainty of participants). The VI+FF+SR condition is also rated similarly to the VI condition (here the increase in intensity seems to have masked any perceivable differences).

The result that an increase in speech intensity entails an increase in intelligibility matches the finding from Garnier and Henrich's study of human Lombard speech [8]. Effects of fundamental frequency and speech rate adaptation on speech intelligibility could not be found in our evaluation.

One reason for this may be that the adaptation of fundamental frequency, and speech rate found in human Lombard speech could not be fully modelled with the speech synthesis system used. The dynamics of the Lombard-models had to be limited, because the HMM-voice starts to sound unnatural when parameters are set to more extreme values.

However, the non-result for these parameters came not fully unexpected. Although adjustment to fundamental frequency and speech rate is found in human Lombard speech, Garnier and Henrich note that these adaptations are subtle and less prominent than increases in speech intensity [8]. Similarly, Bradlow and colleagues found that fundamental frequency and speech rate in general do not have a large impact on speech intelligibility [4].

5. Conclusion and outlook

This paper presented a first system for online Lombard speech synthesis which adjusts synthesis parameters in order to maintain a high level of intelligibility in changing noise conditions. Such a system has useful applications in interactive systems – for example spoken dialogue systems – that are used in noisy environments. The system was evaluated with different models of Lombard speech in a listening experiment and it was found that online adaptation of voice intensity and spectral emphasis increased speech intelligibility and was rated to be more natural in noisy conditions than normal speech. A more subtle and advanced Lombard-adaptation model, however, did not have an effect on intelligibility and perceived naturalness.

While building the system, two general problems were encountered: (1) Current synthetic voices are less dynamic than the human voice and thus limited in adaptability. It is therefore not possible to model all research findings of human Lombard speech with 'off-the-shelf' synthetic voices and speech synthesis systems. (2) As experimental studies on the human Lombard effect usually examine a limited number of noise conditions, empirical models are often underspecified for online Lombard-adaptation in speech synthesis which needs continuous models of how speech parameters are to be adapted. To make significant improvements to online Lombard speech synthesis for interactive systems, further research, both on the engineering and on the empirical side, is needed. Nevertheless, the limited models employed here may already be a useful improvement to speech interfaces used in noisy environments.

The source code of the software (Noise and Lombard modules) as well as supplementary material on the evaluation study is available upon request: `http://purl.org/net/lombard`

Acknowledgements – This research is supported by the Deutsche Forschungsgemeinschaft (DFG) at the Center of Excellence in 'Cognitive Interaction Technology' (CITEC) and by the German Federal Ministry of Education and Research (BMBF) within the 'Leading-Edge Cluster Competition'. We thank Timo Baumann for answering our questions on INPRO_iSS [2], for making its source available, and for commenting on a draft of this text.

6. References

[1] T. Baumann and D. Schlangen, "Evaluating prosodic processing for incremental speech synthesis," in *Proceedings of* INTERSPEECH, Portland, OR, USA, 2012, pp. 438–441.

[2] T. Baumann and D. Schlangen, "INPRO_iSS: A component for just-in-time incremental speech synthesis," in *Proceedings of the ACL 2012 System Demonstrations*, Jeju Island, South Korea, 2012, pp. 103–108.

[3] H. Bořil and P. Pollák, "Design and collection of Czech Lombard speech database," in *Proceedings of* INTERSPEECH, Lisbon, Portugal, 2005, pp. 1577–1580.

[4] A. R. Bradlow, G. M. Torretta, and D. B. Pisoni, "Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics," *Speech Communication*, vol. 20, pp. 255–272, 1996. DOI:10/dczbbk

[5] H. Buschmeier, T. Baumann, B. Dosch, S. Kopp, and D. Schlangen, "Combining incremental language generation and incremental speech synthesis for adaptive information presentation," in *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Seoul, South Korea, 2012, pp. 295–303.

[6] M. Cooke, S. King, M. Garnier, and V. Aubanel, "The listening talker: A review of human and algorithmic context-induced modifications of speech," *Computer Speech & Language*, vol. 28, pp. 543–571, 2014. DOI:10.1016/j.csl.2013.08.003

[7] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane Challenge," in *Proceedings of* INTERSPEECH, Lyon, France, 2013, pp. 3552–3556.

[8] M. Garnier and N. Henrich, "Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise?" *Computer Speech & Language*, vol. 28, pp. 580–597, 2013. DOI:10.1016/j.csl.2013.07.005

[9] M. Guhe, *Incremental Conceptualization for Language Production*. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 2007.

[10] H. Lane, B. Tranel, and C. Sisson, "Regulation of voice communication by sensory dynamics," *Journal of the Acoustical Society of America*, vol. 47, pp. 618–624, 1970. DOI:10.1121/1.1911937

[11] É. Lombard, "Le signe de l'élévation de la voix," *Annales des Maladies de l'Oreille, du Larynx du Nez et du Pharynx*, vol. 37, pp. 101–119, 1911.

[12] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise," *The Journal of the Acoustical Society of America*, vol. 124, pp. 3261–3275, 2008. DOI:10.1121/1.2990705

[13] McKinneySound. (2012) "Ambience, Bar, Large, Busy 001". Sound file. Retrieved 2014-08-18 from http://www.freesfx.co.uk/download/?id=3562.

[14] M. Nicolao, J. Latorre, and R. K. Moore, "C2H: A computational model of H&H-based phonetic contrast in synthetic speech," in *Proceedings of* INTERSPEECH, Portland, OR, USA, 2012, pp. 987–990.

[15] B. Picart, T. Drugman, and T. Dutoit, "Analysis and synthesis of hypo- and hyperarticulated speech," in *Proceedings of the 7th ISCA Tutorial and Research Workshop on Speech Synthesis*, Kyoto, Japan, 2010, pp. 270–275.

[16] A. Rietsch. (2005) "Alice – mit Sicherheit tot". Short story. Retrieved 2014-08-06 from http://www.e-stories.de/view-kurzgeschichten.phtml?12881.

[17] D. Schlangen and G. Skantze, "A general, abstract model of incremental dialogue processing," *Dialogue & Discourse*, vol. 2, pp. 83–111, 2011. DOI:10.5087/dad.2011.105

[18] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: A tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, pp. 365–377, 2003. DOI:10.1023/A:1025708916924

[19] M. D. Skowronski and J. G. Harris, "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments," *Speech Communication*, vol. 48, pp. 549–558, 2005. DOI:10.1016/j.specom.2005.09.003

[20] N. Ström and S. Seneff, "Intelligent barge-in in conversational systems," in *Proceedings of the 6th International Conference on Spoken Language Processing*, vol. 2, Beijing, China, 2000, pp. 652–655.

[21] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *Journal of the Acoustical Society of America*, vol. 84, pp. 917–928, 1988. DOI:10.1121/1.396660

[22] P. Taylor, *Text-to-Speech Synthesis*. Cambridge, UK: Cambridge University Press, 2009. DOI:10/dx9zw2

[23] C. Valentini-Botinhao, J. Yamagishi, S. King, and R. Maia, "Intelligibility enhancement of HMM-generated speech in additive noise by modifying Mel cepstral coefficients to increase the glimpse proportion," *Computer Speech & Language*, vol. 28, pp. 665–686, 2014. DOI:10.1016/j.csl.2013.06.001