Combining perceptually-motivated spectral shaping with loudness and duration modification for intelligibility enhancement of HMM-based synthetic speech in noise

Cassia Valentini-Botinhao[1], *Junichi Yamagishi*[1,2], *Simon King*[1] *and Yannis Stylianou*[3]

[1]Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK
[2]National Institute of Informatics, Tokyo, Japan
[3] Institute of Computer Science, Foundation of Research and Technology Hellas, Crete, Greece

C.Valentini-Botinhao@sms.ed.ac.uk, jyamagis@inf.ed.ac.uk,
Simon.King@ed.ac.uk, styliano@ics.forth.gr

Abstract

This paper presents our entry to a speech-in-noise intelligibility enhancement evaluation: the Hurricane Challenge. The system consists of a Text-To-Speech voice manipulated through a combination of enhancement strategies, each of which is known to be individually successful: a perceptually-motivated spectral shaper based on the Glimpse Proportion measure, dynamic range compression, and adaptation to Lombard excitation and duration patterns. We achieved substantial intelligibility improvements relative to unmodified synthetic speech: 4.9 dB in competing speaker and 4.1 dB in speech-shaped noise. An analysis conducted across this and other two similar evaluations shows that the spectral shaper and the compressor (both of which are loudness boosters) contribute most under higher SNR conditions, particularly for speech-shaped noise. Duration and excitation Lombard-adapted changes are more beneficial in lower SNR conditions, and for competing speaker noise.
Index Terms: intelligibility of speech in noise, HMM-based speech synthesis, Lombard speech

1. Introduction

Providing speech that is matched to listening conditions is very important, not simply to match the expectations of the listener for appropriate-sounding speech (because listeners expect that speakers will adapt their speech according to conditions), but also to achieve highly intelligible speech in challenging environments. In terms of automated processing, this means speech intelligibility enhancement in which clean speech is modified so the subsequent mixture of speech and noise is more intelligible. This is particularly applicable for Text-To-Speech (TTS), because TTS voices tend to be less intelligible in noise than natural speech, even though they may be equally intelligible in clean conditions. Proposed strategies to enhance speech intelligibility in noise for either natural or TTS voices include: temporal and spectral shapers based on clear speech findings [1–4], noise-dependent strategies based on objective measures of intelligibility [5–9], simulated hyper-articulation [10, 11] and direct use of more intelligible speech recordings – for example speech that has been produced by a natural speaker in noise (Lombard speech) – through adaptation or voice conversion [12, 13]. A recent study evaluated a subset of these strategies on a common database of speech, with the same noises and listeners [14]. The current paper describes our entry for a follow-up larger scale evaluation called the Hurricane Challenge [15].

Voice	Adaptation to Lombard	Modification
TTS	-	-
TTSGP [7]	-	GP
TTSGP-DRC [16]	-	GP+DRC
TTSLGP-DRC	excitation and duration	GP+DRC
TTSLomb [14]	all dimensions	-

Table 1: *Voices evaluated in the Hurricane Challenge [15] alongside voices evaluated in other experiments [14, 16].*

In previous work we found that is possible to obtain larger intelligibility gains by performing spectral modifications, rather than adapting a plain speech TTS spectral model to Lombard data [7]. Moreover we found that dynamic range compression can further boost this gain [16]. Although we obtained substantial gains in speech-shaped noise, our results in the case of a competing speaker noise were not as good. To improve performance, we propose to incorporate duration and excitation changes from Lombard speech, by combining three different modification strategies: spectral changes based on the glimpse proportion measure (GP), dynamic range compression (DRC) and adaptation to Lombard duration and excitation.

In the Section 2 we explain the motivation for using each strategy, followed by details of how we built and modified the TTS voice in Section 3. In Section 4 we present an automatic acoustic analysis of the spectral tilt, duration and loudness values and in Section 5 we show the subjective scores obtained in the Hurricane Challenge and compare how much each strategy contributed to the overall intelligibility.

2. Background

After describing the voices in Table 1, including those built for previous evaluations (TTSGP, TTSGP-DRC and TTSLomb), we explain why we built the voice TTSLGP-DRC which combines three strategies for intelligibility enhancement.

In [7] we proposed a method for modifying the sequence of Mel cepstral coefficients such that the glimpse proportion measure (GP) [17], an objective measure of intelligibility of speech in noise, increases for each frame, creating the voice TTSGP as seen in Table 1. In [7] we found that the modified Mel cepstral coefficients provide more intelligibility gains than Lombard-adapted coefficients, but that, although significant intelligibility increases were obtained in speech-shaped noise, the gains obtained in the presence of a competing speaker were

	duration (secs.)	mean/range F_0 (Hz)	spectral tilt (dB/oct.)	loudness (sone)
Natural				
plain	2.06	107.1 / 34.60	-2.14	11.43
Lombard	2.32	136.8 / 46.74	-1.83	11.96
TTS				
TTS	1.95	104.5 / 22.45	-2.26	10.96
TTSGP			-1.90	12.43
TTSGP-DRC			-1.45	13.37
TTSLGP-DRC	2.49	145.2 / 42.55	-1.46	13.12
TTSLomb	2.43		-1.71	12.06

Table 2: *Acoustic properties at sentence level averaged across the dataset.*

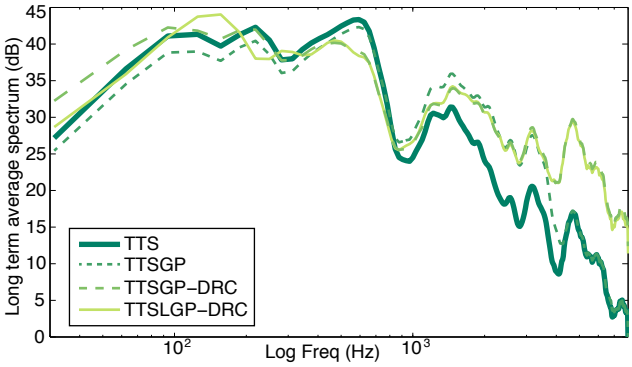

Figure 1: *Long term average spectrum calcualted at a sentence level and averaged across the dataset.*

much smaller. Motivated by our observation that the GP method was mainly boosting vowels and nasals while leaving fricatives and stops relatively untouched, we extended the method in [16] by adding dynamic range compression (DRC) [4], which reallocates energy from higher to lower energetic parts of speech. This voice is referred to here as TTSGP-DRC (the hyphen indicates that GP acts on the generated acoustic parameters before synthesis and DRC acts on the synthesized waveform). This addition did improve our results for both stationary speech-shaped noise and the non-stationary competing speaker noise, yet the gains for the stationary noise condition were still much higher.

Although observed in natural Lombard speech [18–20], reproducing changes in duration and fundamental frequency (F_0) does not necessarily generate significant intelligibility gains [21–23]. In [22] we manipulated the duration and F_0 of a TTS voice and no significant increases in intelligibility were observed in the four noise types tested (car, high frequency, speech-shaped and cafeteria). In [7], however, we saw that quite a significant gain came from using Lombard-adapted fundamental frequency and duration in the competing speaker scenario, even though the noise used for inducing such changes was not matched to the competing speaker masker. We refer to the fully Lombard-adapted (spectral, duration and excitation) voice as TTSLomb. A combined solution for improving results in competing speaker noise while maintaining the gains already achieved in speech-shaped noise, is to apply the GP-based spectral shaper, use Lombard-derived excitation and duration changes (noise-dependent but not matched) through voice adaptation [24], and follow this by DRC: we refer to this combination of strategies here as the TTSLGP-DRC voice.

3. Voice building

To build the voices used in this evaluation we used two different datasets provided by the Hurricane Challenge and recorded by the same British male speaker: normal (plain, read-text) speech data, and Lombard speech (also read-text).

We built two different voices as outlined in Table 1. The voice called simply TTS was created from a high quality average voice model adapted to 2803 sentences of the normal speech data, corresponding to three hours of material. The reason for using adaptation was the lack of phonetic balance in the speech dataset. This voice was also evaluated in [14, 16].

Voice TTSLGP-DRC was based on voice TTS but the models for duration and excitation were further adapted using 780 sentences of Lombard speech data, corresponding to 53 minutes of recorded material. Again, the reason for using adaptation was lack of phonetic balance. The two first Mel cepstral coefficients were modified using the method proposed in [7] and we applied a dynamic range compressor (DRC) [4] to the synthesized waveform.

To train the acoustic modes we extracted from the speech database sampled at 48 kHz the following parameters: 59 Mel cepstral coefficients with $\alpha = 0.77$, Mel scale F_0, and 25 aperiodicity energy bands extracted using STRAIGHT [25]. As an acoustic model we used a hidden semi-Markov model; one stream for the spectrum, three streams for the logF0 and one for the band-limited aperiodicity. The spectral and excitation observation vectors contained static, delta and delta-delta values. To overcome the over smoothing effect caused by the statistical modelling we applied the Global Variance method [26]. The labels used to train and generate the test sentences were created using the pronunciation lexicon combilex [27].

According to the rules of the Hurricane Challenge, each sentence can not be longer than its corresponding noise file, as provided by the challenge, which is around one second longer than the corresponding natural speech signal. To keep within this rule we had to restrict the duration of the generated sentences, because otherwise they would have been on average 0.69 seconds longer than the natural speech, with a significant number of sentences more than one second longer than natural speech. We decided to restrict the duration of each generated sentence to be no more than 300 ms longer than the corresponding natural speech, to allow 300 ms leading / 200 ms lagging noise signal before/after in the stimuli presented to listeners. To achieve that, we forced the overall duration of the sentence to be within this rule (only if necessary) [28]. Because changing this parameter does not actually guarantee a sufficiently reduced duration, we then iteratively decrease the duration (in steps of 100 ms) until it was within the required limits. In the final stimuli, the average duration difference (compared to natural speech) was 0.45 secs, with only once sentence above the 0.5 limit (0.53 secs). Audio samples of our entries are available at `https://wiki.inf.ed.ac.uk/CSTR/HcExternal`.

4. Acoustic analysis

To give more insights into the results, we provide in Table 2 and Fig. 1, a sentence-level acoustic analysis of duration, fundamental frequency F_0 (mean and range), spectral tilt and loudness (measure using the ISO procedure). To measure loudness we used the ISO-532B method [29], the F_0 range was calculated as the difference between the 80th and 20th percentiles and the

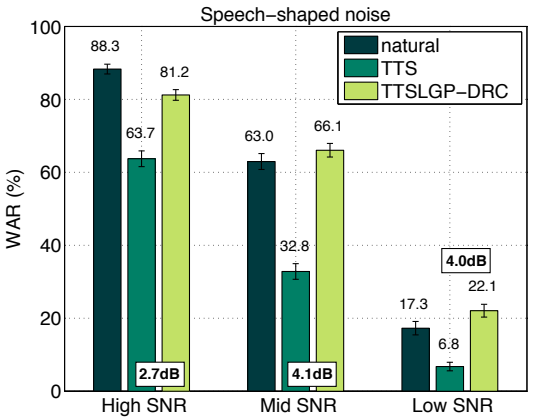

Figure 2: *Hurricane Challenge results for speech-shaped noise.*

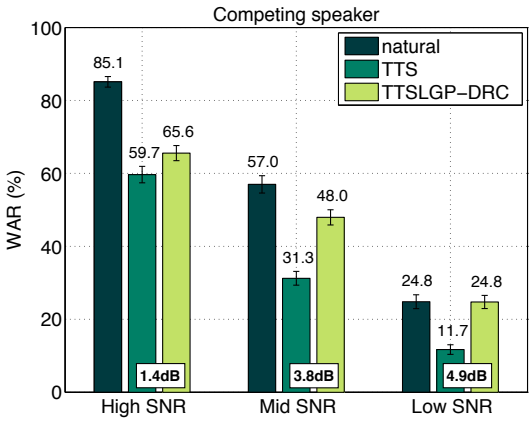

Figure 3: *Hurricane Challenge results for competing speaker.*

spectral tilt was measured as the slope of the linear regression of the long term average spectrum on a one-third octave band scale. These values, presented in Table 2, are first calculated per sentence and then averaged across the 180 sentences that were used in the listening test. Fig. 1 shows the long term average spectrum calculated per sentence and averaged across sentences, for some of the TTS voices.

We see a tendency that GP and DRC increase, implicitly or not, the loudness of speech and flatten spectral tilt, even more so than the Lombard natural voice. We see the boosting effect that GP has around the formant frequency range, and the boosting that DRC gives to higher frequencies. F_0 and its range (within a sentence) is increased in the case of Lombard excitation.

5. Subjective intelligibility results

We now present the subjective intelligibility scores obtained in the Hurricane Challenge and then compare the gains relative to natural speech obtained by the voices displayed in Table 1, with a discussion about the effectiveness of each modification.

5.1. Hurricane Challenge

In total 175 native English speakers participated in the listening test. Each participant transcribed 9 different stimuli per entry. The word accuracy rate was scored as the average across each listeners' individual scores for a particular voice at a particular noise condition, so the standard errors reflect listeners deviation rather than sentence material. The noisy conditions were two maskers added at three different SNR levels, referred here as High Mid and Low: speech-shaped noise (1 , -4 and -9 dB) and competing speaker (-7, -14 and -21 dB). More details in [15].

Figures 2 and 3 show word accuracy rates (WAR) and standard errors for the synthetic voices TTS and TTSLGP-DRC and the natural plain speech entry, mixed with speech-shaped noise and competing speaker respectively. In all noise conditions the gap between natural and TTS is smaller with the TTSLGP-DRC voice, particularly for the lower SNR cases in both noise types.

Another way of presenting these results is through gains in dB that a voice provides relative to another – the so-called equivalent intensity change (EIC) [14]. Relative to TTS, the proposed voice obtained from highest to lowest SNR gains of 2.7, 4.1 and 4.0 dB in speech-shaped noise and 1.4, 3.8 and 4.9 dB when mixed with a competing speaker. The gains across the noises are more comparable under this scale.

5.2. Comparing modifications across different listening tests

To be able to understand the contribution of each component (GP, DRC and Lombard excitation & duration) we compared the WAR changes relative to natural speech that each voice described in Table 1 obtained. We show these results expressed in WAR changes and EIC (dB) in Fig. 4.

The listening tests reported in [14] and [16] had the same set-up (sentence material, noise types and SNRs, stimuli presentation) and same scoring rules. To compare results across them we calculate the gains that each modification obtained relative to the WAR results that the natural speech entry obtained in that particular test. Similar to the Hurricane Challenge methodology, we present the WAR change averaged across the gains obtained by each voice for each listener, and again this means that the standard error measures the variability across listeners. The number of points that define the standard error is defined by the number of participants: 139 in [14], 88 in [16] and 175 in the Hurricane Challenge evaluation [15]. As the TTS entry was present in all three experiments, we show the WAR change for that system averaged across all participants (402 points).

When comparing the voices TTS, TTSGP, TTSGP-DRC and TTSLGP-DRC we can see the gain that each component adds. This addition depends on the noise type and SNR, meaning that some components are more important in one condition than another. In speech-shaped noise, as shown in the top part of Fig.4, GP and DRC contribute most. Duration and excitation changes start contributing only at quite low SNRs. The picture is different for competing speaker in the lower part of Fig.4, where GP and DRC gains are quite modest (apart from the significant gain observed in the highest SNR condition for DRC – where the masker is more an energetic masker than an informational one) and the Lombard-based changes contribute most for the Mid and Low SNR conditions, where 'filling the gaps' in time/frequency is more beneficial than being louder (as seen in Table 2).

Comparing TTSLGP-DRC and TTSLomb we can see the additional gain that GP and DRC provide over adapting spectral parameters as well as duration and excitation parameters, particularly for the mid and low SNR conditions of SSN and for the low SNR of competing speaker.

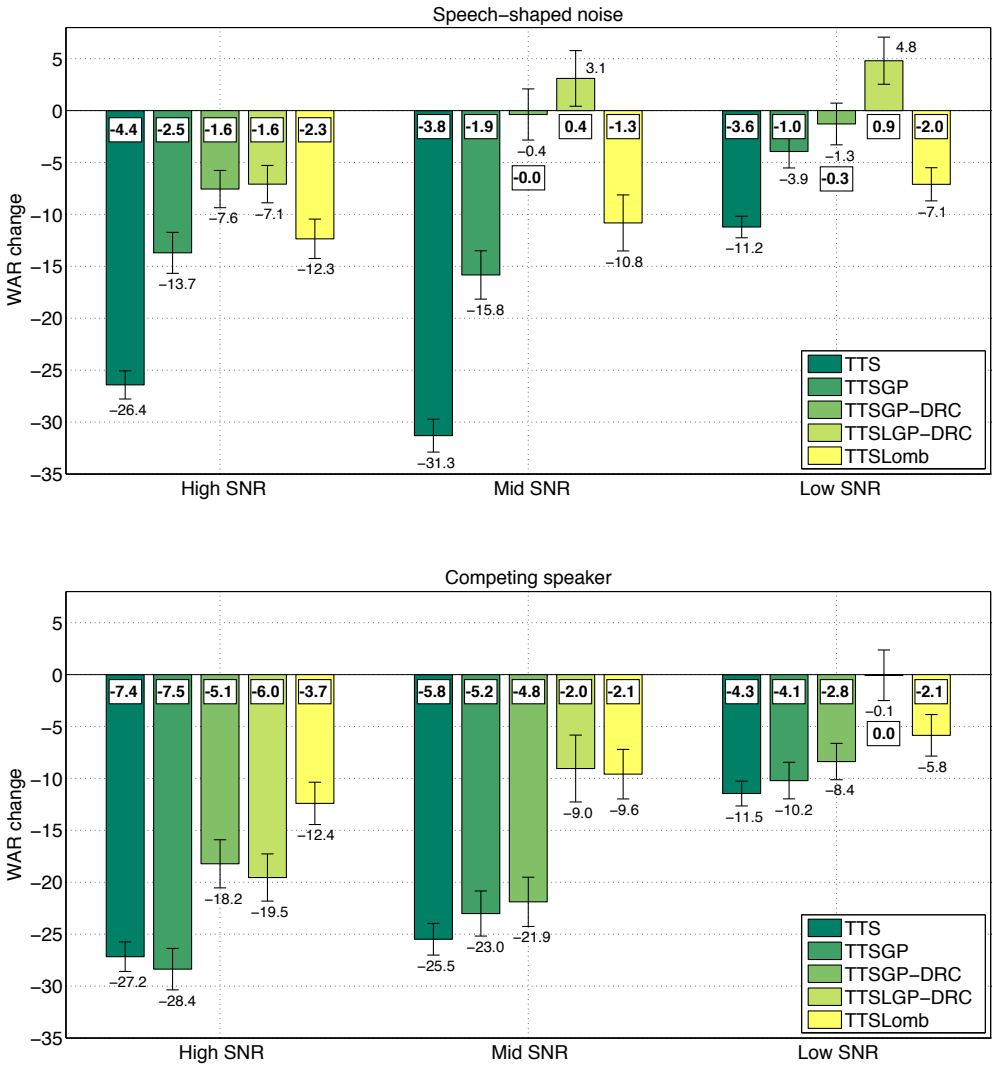

Figure 4: *WAR change and EIC in dB (value inside boxes) relative to unmodified natural speech for speech-shaped noise (top) and competing speaker (bottom). The results for TTSGP and TTSLomb were obtained from [14] and TTSGP-DRC from [16].*

6. Conclusions

By combining duration and excitation changes with other techniques, we have managed to increase intelligibility for competing speaker noise to a comparable level as already obtained for a stationary noise. Although the Lombard changes were induced by a third masker, adaptation to Lombard duration and excitation models contributed to gains not only in the competing speaker but also for the stationary masker at the lowest SNR. This approach however still entails the use of recorded Lombard speech of that particular speaker. We would like to investigate whether similar intelligibility gains can be obtained by applying cross-speaker adaptation of duration and excitation. Additionally we would like to observe its effect on speaker similarity and speech quality.

7. Acknowledgements

The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements 213850 (SCALE) and 256230 (LISTA), and from EPSRC grants EP/I031022/1 (NST) and EP/J002526/1 (CAF).

8. References

[1] R. J. Niederjohn and J. H. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 277–282, Aug. 1976.

[2] M. D. Skowronski and J. G. Harris, "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments," *Speech Comm.*, vol. 48, no. 5, pp. 549–558, 2006.

[3] S. D. Yoo, J. R. Boston, A. El-Jaroudi, C.-C. Li, J. D. Durrant, K. Kovacyk, and S. Shaiman, "Speech signal modification to increase intelligibility in noisy environments," *J. Acoust. Soc. Am.*, vol. 122, no. 2, pp. 1138–1149, Aug. 2007.

[4] T. C. Zorilă, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proc. Interspeech*, Portland, USA, 2012.

[5] B. Sauert and P. Vary, "Near end listening enhancement considering thermal limit of mobile phone loudspeakers," in *Proc. Conf. on Elektronische Sprachsignalverarbeitung (ESSV)*, vol. 61, Aachen, Germany, Sep. 2011, pp. 333–340.

[6] Y. Tang and M. Cooke, "Optimised spectral weightings for noise-dependent speech intelligibility enhancement," in *Proc. Interspeech*, Portland, USA, September 2012.

[7] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Mel cepstral coefficient modification based on the Glimpse Proportion measure for improving the intelligibility of HMM-generated synthetic speech in noise," in *Proc. Interspeech*, Portland, USA, September 2012.

[8] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure," in *Proc. ICASSP*, Mar. 2012, pp. 4061–4064.

[9] P. Petkov, G. Henter, and W. B. Kleijn, "Maximizing phoneme recognition accuracy for enhanced speech intelligibility in noise," *IEEE Trans. on Audio, Speech and Language Processing*, 2013.

[10] B. Picart, T. Drugman, and T. Dutoit, "Continuous control of the degree of articulation in hmm based speech synthesis," in *Proc. Interspeech*, Florence, Italy, 2011.

[11] M. Nicolao, J. Latorre, and R. K. Moore, "C2H A computational model of H&H-based phonetic contrast in synthetic speech," in *Proc. Interspeech*, Portland, USA, September 2012.

[12] B. Langner and A. W. Black, "Improving the understandability of speech synthesis by modeling speech in noise," in *Proc. ICASSP*, vol. 1, 18-23, 2005, pp. 265–268.

[13] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Analysis of HMM-based lombard speech synthesis," in *Proc. Interspeech*, Florence, Italy, August 2011.

[14] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Comm.*, vol. 55, pp. 572–585, 2012.

[15] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane Challenge," in *Proc. Interspeech*, Lyon, France, (submitted) 2013.

[16] C. Valentini-Botinhao, E. Godoy, Y. Stylianou, B. Sauert, S. King, and J. Yamagishi, "Improving intelligibility in noise of HMM-generated speech via noise-dependent and -independent methods." in *Proc. ICASSP*, Vancouver, Canada, May 2013.

[17] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[18] J. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Am.*, vol. 93, no. 1, pp. 510–524, 1993.

[19] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise," *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 3261–3275, 2008.

[20] V. Hazan and R. Baker, "Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions," *J. Acoust. Soc. Am.*, vol. 130, no. 4, pp. 2139–2152, 2011.

[21] Y. Lu and M. Cooke, "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Comm.*, vol. 51, no. 12, pp. 1253–1262, 2009.

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

[23] J. Villegas, M. Cooke, and C. Mayo, "The role of durational changes in the Lombard speech advantage," in *Proc. LISTA Workshop*, Edinburgh, UK, 2012.

[24] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, no. 1, pp. 66 –83, Jan. 2009.

[25] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Comm.*, vol. 27, pp. 187–207, 1999.

[26] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.

[27] K. Richmond, R. Clark, and S. Fitt, "On generating Combilex pronunciations via morphological analysis," in *Proc. Interspeech*, Makuhari, Japan, September 2010, pp. 1974–1977.

[28] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis," in *Proc. ICSLP*, Sydney, Australia, Dec. 1998, pp. 29–32.

[29] ISO 532, "Acoustics - method for calculating loudness level," ISO, Geneva, Switzerland, 1975.