Improving intelligibility in noise of HMM-generated speech via noise-dependent and -independent methods.

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.

IMPROVING INTELLIGIBILITY IN NOISE OF HMM-GENERATED SPEECH VIA NOISE-DEPENDENT AND -INDEPENDENT METHODS

Cassia Valentini-Botinhao[1], Elizabeth Godoy[2], Yannis Stylianou[2], Bastian Sauert[3]
Simon King[1] and Junichi Yamagishi[1]

[1] The Centre for Speech Technology Research, University of Edinburgh, UK
[2] Institute of Computer Science, Foundation of Research and Technology Hellas (FORTH), Crete, Greece
[3] Institute of Communication Systems and Data Processing, RWTH Aachen University, Germany

ABSTRACT

In order to improve the intelligibility of HMM-generated Text-to-Speech (TTS) in noise, this work evaluates several speech enhancement methods, exploring combinations of noise-independent and -dependent approaches as well as algorithms previously developed for natural speech. We evaluate one noise-dependent method proposed for TTS, based on the glimpse proportion measure, and three approaches originally proposed for natural speech - one that estimates the noise and is based on the speech intelligibility index, and two noise-independent methods based on different spectral shaping techniques followed by dynamic range compression. We demonstrate how these methods influence the average spectra for different phone classes. We then present results of a listening experiment with speech-shaped noise and a competing speaker. A few methods made the TTS voice even more intelligible than the natural one. Although noise-dependent methods did not improve gains, the intelligibility differences found in distinct noises motivates such dependency.

Index Terms— Speech intelligibility in noise, HMM-based speech synthesis

1. INTRODUCTION

With growing numbers of applications exploiting speech technologies, real-life situations inevitably arise in which listeners hear speech in noise. Accordingly, there is great interest in speech intelligibility enhancement algorithms for both natural and synthetic speech. Many studies have been conducted in this field, with some approaches motivated by phenomena observed in human speech production: boosting of the consonant-vowel power ratio [1, 2, 3], spectral tilt flattening and formant enhancement [4, 5, 6] and manipulation of duration and prosody [7]. While the above examples are noise-independent, other approaches exploit knowledge of the noise masker. However, since human speech production adaptations to different noise conditions are less known, work that makes use of the noise knowledge is mostly based on the perception of speech in noise: modification of the local signal-to-noise ratio (SNR) [8, 9], optimisation of the spectral audio power reallocation based on the speech intelligibility index [10], spectral optimization based on the glimpse proportion measure [11, 12], a spectro-temporal optimization based on a perceptual distortion measure [13] and the insertion of small pauses [14]. Statistical approaches to speech enhancement using recordings of speech in noise include voice conversion [15] and adaptation [5]. One noise-dependent approach that is of primary focus in the present work is Mel cepstral modification based on the glimpse proportion measure [12].

While TTS voices can be as intelligible as natural speech in clean conditions, intelligibility drops quite rapidly in adverse conditions [16], more strongly motivating the use of intelligibility enhancement methods and potentially requiring knowledge of the noise masker. However, noise-dependent methods, either for natural or TTS voices, have only recently been proposed and it remains relatively unknown to what extend exploiting spectro-temporal characteristics of the masker is useful. To evaluate a range of enhancement algorithms, both noise-dependent and independent, [17] describes a large scale listening experiment with 5 methods for natural speech and 2 for TTS evaluated under the same conditions (speech and noise material). In this evaluation, it was observed that noise-independent spectral shaping with Dynamic Range Compression (SSDRC) [6] provided the best results of the modifications on natural speech while [10], although noise-dependent, did not perform as well. It was also observed that a noise-dependent approach [12] applied to a TTS voice was as intelligible as a Lombard-adapted voice in some stationary noise conditions, but still not as intelligible as natural speech. A significantly large intelligibility gap was also confirmed between TTS and the natural voice in almost all noise conditions.

In this paper, we investigate whether intelligibility enhancement methods originally proposed for natural speech can also improve intelligibility of a TTS voice and help bridge this gap. Furthermore, we seek to discover whether it is possible to improve a noise-independent method [6] and a noise-dependent method [12] by combining them, effectively offering insight on the extent to which noise dependency is required in terms of achieving significant intelligibility gains.

In Section 2 we present the details of the methods we are comparing and in Section 3 we show of how we built the TTS voice used in the evaluation. Section 4 we present the methods spectral gains at a phone level. Section 5 shows the listening experiment design and results followed by conclusions in Section 6.

2. INTELLIGIBILITY ENHANCEMENT METHODS

We evaluate a total of seven TTS voice styles, as shown in Table 1: two noise-independent (SS-DRC [6] and SSE-DRC), two noise-dependent (TTSGP [12] and OptSII [10]) and two method combinations (TTSGP-DRC and TTSGP-SS-DRC). The TTSGP method is applied directly to the generated spectral parameters, all other methods work as a post processing of the waveform generated by the TTS model (represented by the addition of the acronym TTS-). The following describes each of the methods in more detail. Audio samples can be found at: https://wiki.inf.ed.ac.uk/CSTR/TtsHc.

SS-DRC [6] performs spectral shaping (SS) followed by dy-

Style	Approach	ND
TTS	unmodified TTS	no
TTS-SS-DRC	spectral shaping (SS) followed by dynamic range compression (DRC) [6] applied to TTS	no
TTS-SSE-DRC	extended version of SS (SSE) followed by DRC applied to TTS	no
TTS-OptSII	SII optimisation [10] applied to TTS	yes
TTSGP	Glimpse-optimised TTS [12]	yes
TTSGP-DRC	TTSGP followed by DRC	yes
TTSGP-SS-DRC	TTSGP followed by SS-DRC	yes

Table 1. Speech styles tested. ND: noise dependency.

namic range compression (DRC). Spectral shaping consists of two cascaded subsystems which are adaptive to the probability of voicing: (i) an adaptive sharpening where the formant information is enhanced, and (ii) an adaptive pre-emphasis filter. A third fixed spectral shaping is used to prevent attenuation of high frequencies in the speech signal during the signal reproduction. The output of the spectral shaping system is then input to the DRC, inspired by compression strategies used in sound recording and reproduction, audio broadcasting as well in amplification techniques in hearing aids [18].

The extended spectral shaping (SSE) is carried out on all voiced frames and consists of three components: (i) a fixed filter to increase the spectral energy gain in certain frequency bands, (ii) peak enhancement via cepstral liftering and (iii) slight formant shifting via frequency warping. First, the fixed filter is bi-modal, with the most gain (12 dB) between 1-4 kHz, mimicking the spectral gains observed in Lombard speech [19], and the secondary mode has approximately half of the maximal gain and is concentrated between 5.5-7.5 kHz. Second, the peak-enhancement follows the peak-weighted cepstral lifter (α=0.85) presented in [20] for enhancement in speech recognition. Third, the frequency warping shifts the first and second formants moderately (less than 100 Hz), on average upwards in frequency. The frequency warping function is constant and derived from observations on the expanded vowel space of two speakers in a separate clear speech corpus involving the Harvard sentences.

In the OptSII method [10, 21] the audio power of the speech signal is spectrally reallocated with respect to the speech intelligibility index (SII) [22]. A recursive closed-form optimisation scheme calculates, for each time frame, the spectral weights in 21 Bark-scaled subbands which maximise the SII, given the current disturbance spectrum levels, with the additional constraint of an unchanged short-term audio power of the speech signal. Opposed to [10] and the OptSII style used in [17], in this evaluation a moving average noise estimator is used, which is also able to track CS noise.

To create the TTSGP style voice a Mel cepstral coefficient modification method [12] was applied to the spectral parameters generated by TTS models. The TTSGP method is based on the glimpse proportion measure for speech in noise [23] and it works on each individual time frame by modifying the first two Mel cepstral coefficients (excluding the log-energy coefficient) in order to maximise intelligibility of speech in noise, as given by an approximated version of the glimpse proportion measure, with the constraint that the energy in the frame should remained fixed. This implies that there is no reallocation of energy across time frames, only within frequency regions. Both convergence and distortion (10% relative increase in the Euclidian distance between the auditory representation of original and modified speech) are used as stopping criteria.

3. VOICE BUILDING

We build the synthetic voice using the statistical and parametric HMM-based TTS framework [24]. To train the TTS models, we used a dataset recorded by a British male speaker of plain read speech. The unmodified baseline voice TTS was created from a high quality average voice model adapted to 2803 sentences of this plain speech database (three hours of material). The data was sampled at 48 kHz. We extracted the following features: 59 Mel cepstral coefficients, Mel scale F0 and 25 aperiodicity energy bands extracted using STRAIGHT [25]. To model these features, we used a hidden semi-Markov model. The observation vectors for the spectral and excitation parameters contained static, delta and delta-delta values, with one stream for the spectrum, three streams for F0 and one for the band-limited aperiodicity. We applied the global variance method [26] to compensate for the over smoothing effect caused by the statistical nature of the acoustical modelling.

4. SPECTRAL GAIN ANALYSIS

As all methods modify the speech spectrum, while maintaining prosody and duration, we present here acoustic analysis based only in terms of spectrum gains. We calculated these gains at a phone level and grouped the results in phone classes: vowels, nasals, approximants, stops and fricatives as seen in Fig.1. To obtain these gains, we calculated the phone periodogram by extracting a 512 points discrete Fourier transform calculated at 20 ms hamming window at every 10 ms and averaged across the time frames within the phone boundaries. The gain is then the difference of the phone periodogram in dB for a certain method and the periodogram for the unmodified TTS speech. For the noise-dependent methods, this was calculated for speech-shaped noise (SSN) in the mid SNR condition: results will differ for other noise types and levels.

For the majority of the methods, the average spectral gains can be interpreted as a sort of correction filter that re-allocates spectral energy and remains largely constant across phones for stationary maskers like the SSN. For TTS-SS-DRC, the gain curve shape is determined primarily by the SS fixed filter (seen for all phones), but the effective scale of the gains is affected by the DRC. As we can see in Fig.1, the SS fixed-filter has a very wide-flat gain between 1-4 kHz and a gradual rolloff with increasing frequency. The gain curves in TTSGP, TTS-SSE-DRC and TTS-OptSII, on the other hand, are generally bi-modal. Note that the shape of the fixed-filters or gain curves is most apparent with the voiced phones, particularly with vowels. As observed in [21] at low SNR, OptSII shows a bandpass characteristic, at mid SNR the general spectral shape of the speech signal tends to follow the shape of the noise, and at higher SNR the spectral gains are quiet low.

When comparing TTSGP and TTSGP-DRC, we can clearly see the effect DRC has: gain reduction especially on vowels and increased gain on stop and fricatives, while also determining an upward-sloping linear-like gain curve shape on these last phones. That is, DRC is re-allocating energy of frames in such a way as to increase loudness of the unvoiced parts of speech.

Looking at the gain curves for the TTSGP-SS-DRC method, we can see that the fixed-filter shape of SS dominates, but the GP gain curve is apparent in the roundness of the first mode in the voices (first three categories). More importantly, the scale of the gain is compounded by combining the GP-SS as seen from the gain obtained on the voiced segments.

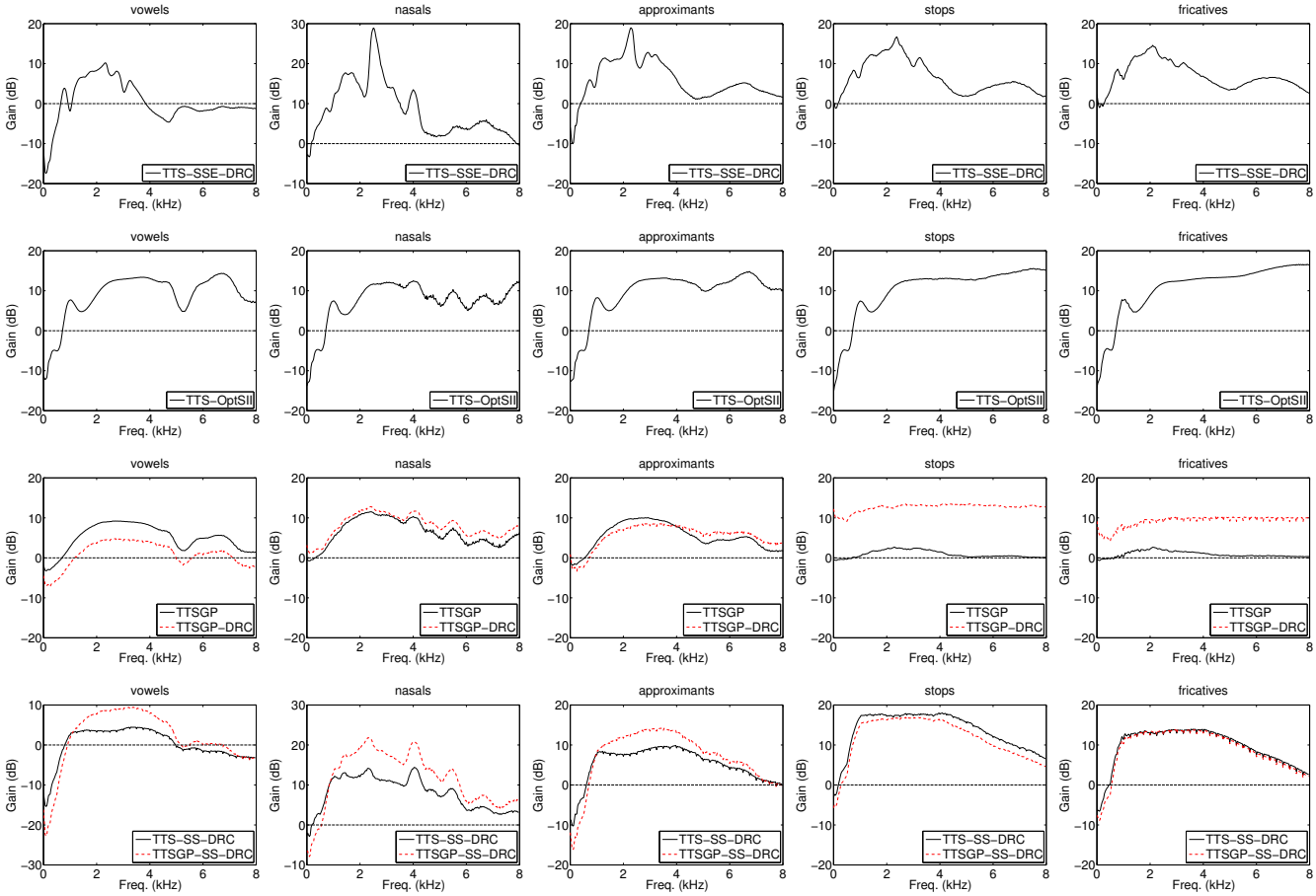


Fig. 1. Spectral gains (dB) obtained by each style over the unmodified baseline TTS.

5. INTELLIGIBILITY SCORES

We evaluated the seven different TTS styles displayed in Table 1 as well as natural speech in a large listening experiment. We present intelligibility scores in terms of word accuracy rates (WAR) in % and equivalent intensity change (EIC) in dB.

5.1. Listening experiment

We added the seven synthetic styles shown in Table 1 and the natural speech to speech-shaped noise (SSN) and a female competing speaker (CS) at 3 different signal to noise ratios ('Low SNR', 'Mid SNR', and 'High SNR'), estimated in pilot tests to be -9, -4 and $+1$ dB for SSN, and -21, -14 and -7 dB for CS as used in [17]. In order to obtain listening scores for word accuracy, we performed a listening test with 88 native English speakers. Each participant heard a set of 180 Harvard sentences, groups of 4 participants heard 15 different sentences for each listening condition (style/noise/SNR).

5.2. Results and discussions

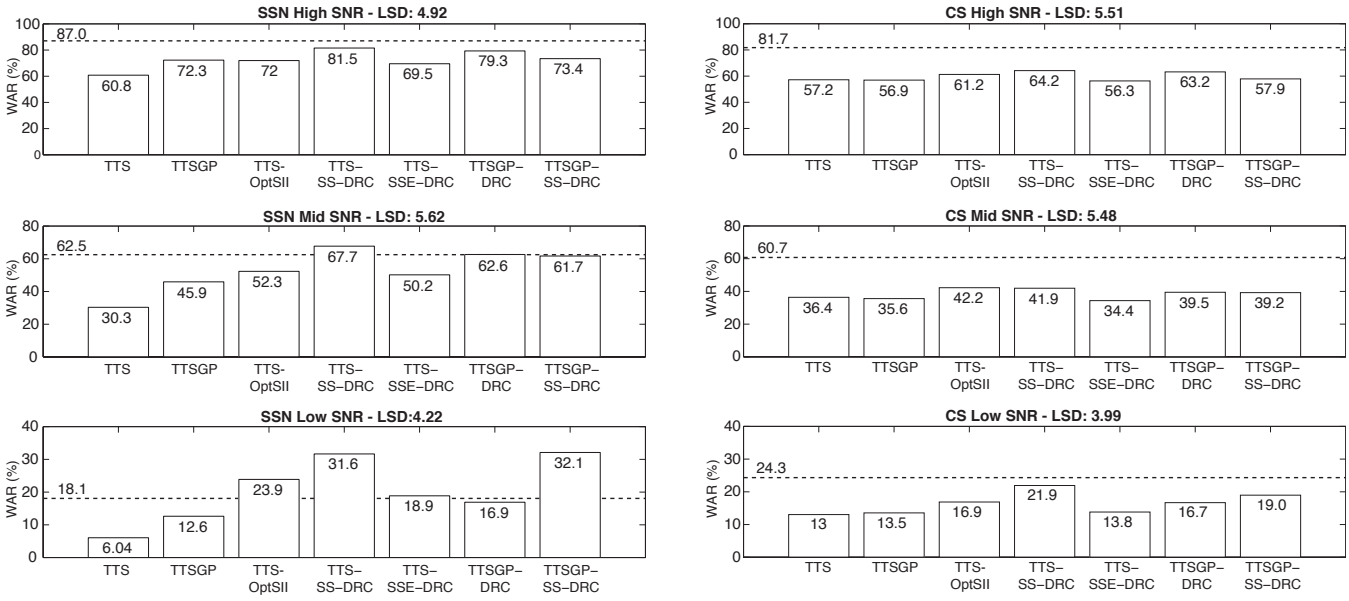
Fig.2 shows the WAR calculated across all sentences for each style in each listening condition, only content words were counted. The dashed line corresponds to the WAR obtained by the natural speech in that condition. The results are organized by noise type and level.

Overall, we can see that the most effective method is the TTS-SS-DRC. It seems that the width and the gain of the primary mode is very important as TTSGP, TTS-SSE-DRC and TTS-OptSII suffer slightly as a result of too narrow or curved gains indicating that, for example, the gains around 1 kHz should be higher. The secondary mode does not seem to benefit intelligibility as much.

All methods except the TTSGP perform some sort of high frequency boosting which enhances voiced segments. This significantly aids intelligibility in the SSN and CS conditions as these noises have stronger low frequency components. We can clearly see this intelligibility gain by comparing the results of TTSGP-DRC and TTSGP: DRC improves TTSGP performance in all noisy conditions, particularly for the SSN Mid SNR condition. TTSGP-DRC and TTS-OptSII obtained similar performance: in SSN TTSGP-DRC performs better in the mid and high conditions and no significant differences appeared in CS. At lower SNRs larger gain at higher frequencies (observed for TTS-SS-DRC and TTS-OptSII) seems to be more beneficial most likely due to the masker.

Applying SS-DRC to a TTSGP style voice did not improve intelligibility as we see that TTSGP-SS-DRC either obtained worst or similar WARs than TTS-SS-DRC. The compounded gain of SS-GP seen in the acoustic analysis is most beneficial at Low SNR (specifically for SSN). Otherwise, it seems excessive and TTSGP-DRC or TTS-SS-DRC are sufficient.

A few methods were as intelligible as natural speech in SSN



(a) Speech-shaped noise (SSN). (b) Competing speaker (CS).

Fig. 2. Word accuracy rate (WAR) obtained with the listening evaluation. The dashed line corresponds to the WAR obtained by natural speech in that condition. LSD is Fisher's least significant difference.

Mid SNR and Low SNR condition. TTS-OptSII, TTS-SS-DRC and TTSGP-SS-DRC were significantly more intelligible than natural speech in SSN Low SNR. For the CS case TTS-SS-DRC was as intelligible as natural speech at Low SNR. Natural speech in CS for all SNRs was significantly more intelligible than the TTS styles. The differences among the methods is attenuated in CS, that is the gains obtained by the noise-independent method TTS-SS-DRC were attenuated.

These results can be converted as equivalent intensity changes (EIC) relative to normal natural speech in a dB scale as proposed in [17]. Calculating this changes with respect to natural speech we found that TTS-SS-DRC was 2.0 dB higher than natural speech in SSN Low SNR, this gain is lower for the Mid SNR condition: 0.7 dB. The higher gain was obtained by TTSGP-SS-DRC: 2.25 dB in SSN Low SNR. The gap between modified TTS and natural speech is larger for CS, for Mid and High SNR conditions most methods were at least 4 dB away from natural speech while for the Low SNR condition TTS-SS-DRC decreased the gap to -0.7 dB.

As the speech material and the noise conditions were the same as the ones used in [17] we can directly compare SS-DRC results (OptSII noise estimation here is different than the system used in [17]). SS-DRC applied to a TTS voice improves TTS performance by 4.5 dB for SSN and 2.1 dB for CS (averaged across all SNRs) compared to performance gains of 4.2 dB and 3.1 dB for natural speech. Slight increase in performance 0.3 dB for TTS SSN but a decrease for CS of 1.0 dB.

6. CONCLUSIONS

In this paper we presented results of a spectral analysis and a listening experiment evaluating speech intelligibility enhancement methods applied to a HMM-generated Text-To-Speech (TTS) voice. We evaluated two noise-independent approaches proposed for natural speech and based on spectral shapers followed by a dynamic range compressor (TTS-SS-DRC and TTS-SSE-DRC) and two noise-dependent methods both based on speech intelligibility objective measures: the glimpse proportion (TTSGP) and the speech intelligibility measure (TTS-OptSII). We mixed these four voices, two method combinations (TTSGP-DRC and TTSGP-SS-DRC) and natural speech with speech-shaped noise (SSN) and competing speaker (CS) maskers. Although the methods share similar spectral gain shapes the absolute gains and its modal nature are quite different. The most effective strategy in SSN was TTS-SS-DRC, a noise-independent unimodal spectral gain combined with DRC. We also observed that some styles were more intelligible than natural speech, which shows how effective these methods can be when applied to a synthetic voice. In the CS scenario, all methods performed significantly worse than natural unmodified speech, except for the lower SNR condition where TTS-SS-DRC obtained similar gains to natural speech. Compared to its performance with natural speech, SS-DRC originally proposed in that scenario had similar results for TTS in SSN and a slight drop in performance for CS.

While a noise-independent approach sufficiently increased intelligibility in a stationary masker, performance dropped in the case of the competing speaker: the noise type significantly influences intelligibility. This motivates further exploration into tailoring methods to specific spectro-temporal characteristics of the noise masker.

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 and 256230 (SCALE and LISTA). The authors would like to thank Martin Cooke and Yan Tang for providing scripts and psychometric data.

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, 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. 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, 2007.

[4] I.V. McLoughlin and R.J. Chance, "LSP-based speech modification for intelligibility enhancement," in *Proc. Digital Signal Processing*, Santorini, Greece, July 1997, vol. 2, pp. 591–594.

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

[6] 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, Sept. 2012.

[7] D.-Y. Huang, S. Rahardja, and E.P. Ong, "Lombard effect mimicking," in *Proc. SSW7*, Kyoto, Japan, Sept. 2010, pp. 258–263.

[8] B. Sauert and P. Vary, "Near end listening enhancement: Speech intelligibility improvement in noisy environments," in *Proc. ICASSP*, Toulouse, France, May 2006, pp. 493–496.

[9] Y. Tang and M. Cooke, "Energy reallocation strategies for speech enhancement in known noise conditions," in *Proc. Interspeech*, Dallas, USA, 2010, pp. 1636–1639.

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

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

[12] 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, Sept. 2012.

[13] 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*, Kyoto, Japan, 2012, pp. 4061–4064.

[14] Y. Tang and M. Cooke, "Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints," in *Proc. Interspeech*, Florence, Italy, Aug. 2011, pp. 345–348.

[15] B. Langner and A.W. Black, "Improving the understandability of speech synthesis by modeling speech in noise," in *Proc. ICASSP*, Philadelphia, USA, March 2005, vol. 1, pp. 265–268.

[16] S. King and V. Karaiskos, "The Blizzard Challenge 2010," in *Proc. Blizzard Challenge Workshop*, Kyoto, Japan, Sept. 2010.

[17] 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. in revision, 2012.

[18] B.A. Blesser, "Audio dynamic range compression for minimum perceived distortion," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, no. 1, 1969.

[19] E. Godoy and Y. Stylianou, "Unsupervised acoustic analyses of normal and Lombard speech, with spectral envelope transformation to improve intelligibility," in *Proc. Interspeech*, Portland, USA, Sept. 2012.

[20] H.K. Kim and H.S. Lee, "Spectral peak-weighted liftering of cepstral coefficients for speech recognition," *IEICE Trans. Inf. Syst.*, vol. 83, no. 7, pp. 1540–1549, 2000.

[21] B. Sauert and P. Vary, "Near-end listening enhancement in the presence of bandpass noises," in *Proc. of ITG-Fachtagung Sprachkommunikation*, Sept. 2012, vol. 10.

[22] ANSI, "ANSI S3.5-1997 Methods for the calculation of the speech intelligibility index," 1997.

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

[24] H. Zen, K. Tokuda, and A.W. Black, "Statistical parametric speech synthesis," *Speech Comm.*, vol. 51, no. 11, pp. 1039–1064, 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.