INTELLIGIBILITY ENHANCEMENT OF SPEECH IN NOISE

Cassia Valentini-Botinhao University of Edinburgh, UK
Junichi Yamagishi University of Edinburgh, UK / National Institute of Informatics, Japan
Simon King University of Edinburgh, UK

1 INTRODUCTION

To maintain communication success, humans change the way they speak and hear according to many factors, like the age, gender, native language and social relationship between talker and listener. Other factors are dictated by how communication takes place, such as environmental factors like an active competing speaker or limitations on the communication channel. As in natural interaction, we expect to communicate with and use synthetic voices that can also adapt to different listening scenarios and keep the level of intelligibility high. Research in speech technology needs to account for this to change the way we transmit, store and artificially generate speech accordingly.

2 NOISE COMPENSATION

We are interested in increasing the intelligibility of speech automatically according to the noise signal. For that, we need to find a method that can predict the effect noise has on intelligibility and find an effective strategy, i.e. which aspects of speech are worth enhancing. Before looking into evaluating different models and strategies we survey the literature on models of hearing and speaking in noise.

2.1 Hearing and speaking in noise

The mechanism that explains why, and to a certain extent how, we are able to understand speech in a mixtures of sound sources concerns several different stages of processing. These mechanisms include: auditory grouping, glimpsing, linguistic adjustments and the regard for spatial and visual cues (Loizou, 2007). Auditory grouping, also known as auditory streaming, is the capacity of a listener to group together different time-frequency regions of speech and associate them to a single audio source (Bregman, 1990). Another important auditory mechanism that aids source separation is so-called glimpsing, which is the ability to extract time-frequency regions where the corrupted speech signal is less masked and therefore less distorted (Cooke, 2003). The linguistic information can also help comprehension by limiting the number of possible guesses for what a word could have been given the lexicon and the context. Spatial release of masking, that is when target and masker sources are located in different regions of the acoustic space, can increase intelligibility by significant amounts due to the additional cues of intensity and time of arrival differences between ears, an advantage of 2 to 7 dB equivalent intensity gain (Hawley et al., 2004). Visual information provided by the speaker can also increase recognition rates. Looking at the lips of the person talking can be crucial in distinguishing phones like /t/ and /p/ which are highly confusable in noise but are produced using different articulators. Such visual cues can give up to 11 dB in equivalent intensity gains (Macleod and Summerfield, 1987).

To increase the success of communication, humans adapt to their immediate context by changing the way they produce speech as well. This adaptation can happen at different levels, that is, at an acoustic level with changes in phonation, place and manner of articulation or at a linguistic level, with changes in words and vocabulary. The increase in vocal effort observed in speech produced in noise is generally called the Lombard effect (Lombard, 1911) and speech produced in noise is known as Lombard speech. Many acoustic changes have been reported for Lombard speech: an increase in intensity, increase in vowel duration, reduction in speaking rate, a shift in the energy distribution of the spectral content from low to middle and high frequency regions which results in flatter spectral tilt, increase in the first formant and in some studies increases in the second formant were also observed, increase in F0 (both the average and the range) (Summers et al., 1988; Junqua, 1993; Hansen, 1996; Garnier et al., 2006; Lu and Cooke, 2008).

2.2 Existing intelligibility enhancement methods

The different mechanisms of speech production in noise show that it is possible to modify speech in such a way that the mixture of speech and noise is more intelligible for the listener without an overall level increase. To emulate such an effect one could for instance modify speech produced in quiet by mimicking the acoustic changes seen in studies of speech produced in noise. Methods under this category include: boosting the consonant-vowel power ratio (an effect usually observed in clear speech) (Niederjohn and Grotelueschen, 1976; Skowronski and Harris, 2006; Yoo et al., 2007), spectral tilt flattening and formant enhancement (McLoughlin and Chance, 1997; Raitio et al., 2011a), manipulation of duration and prosody (Huang et al., 2010), increasing of duration, intensity and F0 of content words (Patel et al., 2006) and both formant and loudness enhancement (Zoril˘a et al., 2012). Because it is not known to what extent the acoustic changes relate to the characteristics of the noise, these types of speech modifications are noise-independent.

Another strategy is to make direct use of available recordings of Lombard speech data through voice conversion techniques (Langner and Black, 2005) and adaptation techniques (Raitio et al., 2011a; Picart et al., 2013). This requires recordings of Lombard speech data from the speaker whose voice is to be synthesized. Recent work has also been carried out using estimates of the noise context for so-called noise dependent methods. These approaches include modification of the local SNR (Sauert and Vary, 2006; Tang and Cooke, 2010), optimisation of spectral power reallocation based on the SII (Sauert and Vary, 2010, 2011) and a global fixed optimization to maximize the glimpse proportion (GP) (Tang and Cooke, 2012) as well as different strategies for the insertion of small pauses (Tang and Cooke, 2011) and GP-based duration changes (Aubanel and Cooke, 2013). Recently, Taal et al. (2012) presented an optimisation algorithm based on a spectro-temporal perceptual distortion measure and in Petkov et al. (2012) an algorithm based on a statistical model of speech was described.

In speech intelligibility enhancement evaluation (Cooke et al., 2012, 2013) it is common to adopt a sentence-level energy constraint which will also be adopted here: the energy of the modified speech – calculated over the entire sentence – is normalized to be equal to the energy of the unprocessed speech signal. For speech transmission, however, evaluation is more strict: the energy per frame should not be modified. While one might argue against it, per sentence energy normalization allows for a wider range of long-term strategies to be applied such as boosting certain words, phonetic units or regions to the detriment of other parts of speech that can potentially be exploited in applications such as TTS and the reproduction of pre-recorded speech

3 EVALUATION OF INTELLIGIBILITY MEASURES

We present the results of two experiments designed to evaluate objective measures of speech with regards to intelligibility prediction of HMM-generated synthetic speech in additive noise. Experiment I dealt mostly with non-modified synthetic speech and Experiment II with modified synthetic speech. This work was partially published in Valentini-Botinhao et al. (2011b,a).

Table 1 shows the wide variety of objective measures that we evaluated, ranging from conventional spectrum-based measures to recently proposed measures based on rather complex models of the human auditory system. The main findings of this work, concluded from the results presented in Figures 1 and 2, are that model-based measures – notably Dau and GP – have the highest predictive power under diverse listening conditions of varying noise type and speech modification type. We also found that simple modifications at a spectral level can have a significant positive impact on the intelligibility of HMM-generated synthetic speech in noise. By combining a modification strategy that improves intelligibility with an objective measure that accurately predicts the effect of that modification, we will arrived at a first version of what we were aiming for: automatically-controlled speech intelligibility enhancement.

Audio samples used in this evaluation are available at https://wiki.inf.ed.ac.uk/CSTR/Modifications

Acronym	Measure
GP	Glimpse proportion (Cooke, 2006)
Dau	Dau measure (Christiansen et al., 2010)
STOI	Short Term Objective Measure (Taal et al., 2010)
SII	Speech Intelligibility Index (ANSI, 1997)
PESQ	Perceptual Evaluation of Speech Quality (Rix et al., 2001)
FWS	Frequency Weighted SNR (Tribolet et al., 1978)
WSS	Weighted Spectral Slope (Klatt, 1982)
CEP	Cepstral distance (Gray and Markel, 1976)
LSD	Log Spectral Distance (Gray and Markel, 1976)
IS	Itakura Saito distance (Gray and Markel, 1976)
LLR	Log Likelihood Ratio (Gray and Markel, 1976)

Table 1 – List of objective measures evaluated.

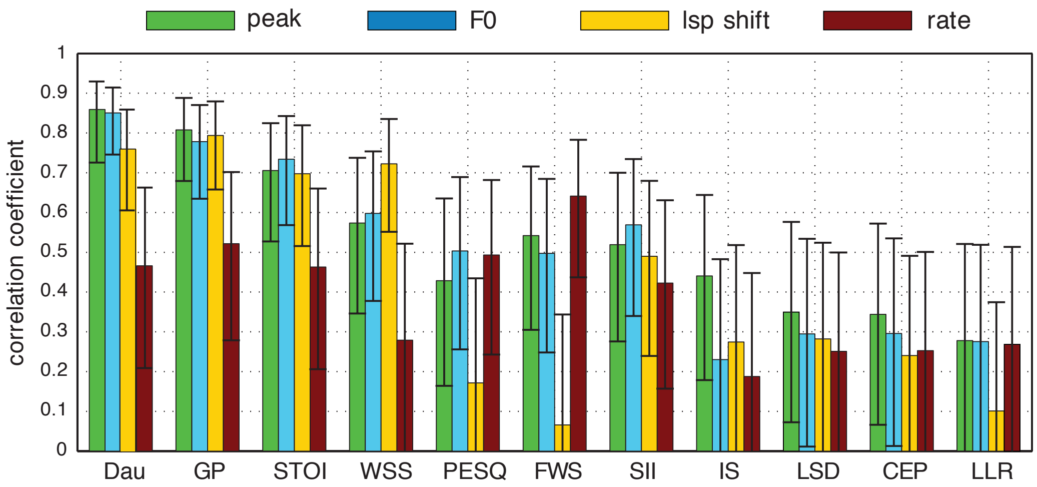

Figure 1 – Correlation coefficient between objective and listeners subjective scores broken down by speech enhancement method.

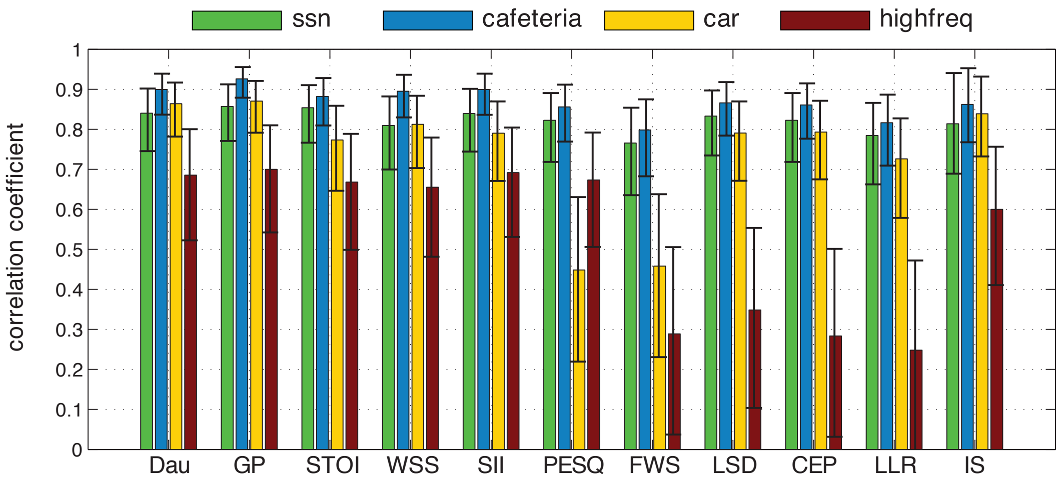

Figure 2 – Correlation coefficient between objective and listeners subjective scores broken down by noise type.

4 PROPOSED INTELLIGIBILITY ENHANCEMENT METHODS

We observed in the previous experiments that the Glimpse Proportion (GP) measure for speech intelligibility in noise (Cooke, 2006) has a high correlation coefficient with subjective intelligibility scores for HMM-generated synthetic speech whose spectral envelope has been modified (LSP-shift modification). Moreover, modifications in the spectral envelope domain can achieve quite high intelligibility gains. Now, we propose a method that can be applied at synthesis time, and does not require any information about the spectral envelope of natural speech to achieve distortion control. In this new method, we maximize the GP alone given the synthetic speech to be reproduced and the background noise. The maximization of the GP without any constraint will generate glimpses across all spectral envelope generating audible distortions.

The method operates on the Mel cepstral coefficients (parameters that describe the spectrum envelope of speech) generated by acoustic models which were trained only on natural read speech collected in quiet conditions, of the type normally used to build text-to-speech systems. The method updates the Mel cepstral coefficients iteratively via gradient descent such that the glimpse proportion increases, without changing the overall energy. We observed that sentences generated with such modified Mel cepstral coefficients have a boost in frequencies between 1-4 kHz and that this boost is highly dependent on the phonetic units: vowels and nasals are more enhanced than fricatives and stops. To evaluate the method, we built eight different voices from normal read-text speech data from a male speaker. Intelligibility results with a speech-shaped noise masker, presented in Figure 3, show that the modified voice (N-M2) is as intelligible as a synthetic voice trained with plain speech then adapted to Lombard speech (L). When mixed with a competing talker the gains are more modest. This work was partially published in (Valentini-Botinhao et al., 2012c, 2013c).

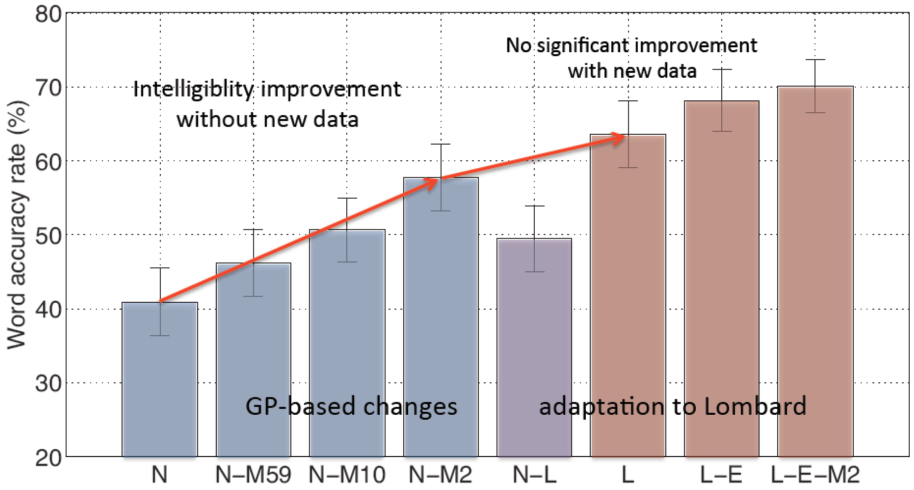

Figure 3 – Word accuracy rates obtained with subjective listening tests of synthetic speech in speech-shaped noise.

4.1 Combining and comparing with other methods

In this section, we present the result of one large scale listening experiments comparing the modification proposed in the previous section, referred now as GP, to other intelligibility enhancement methods applied to the same TTS baseline. Additionally, we evaluate a series of method combinations of GP with the following noise-independent methods: dynamic range compression (DRC) and adaptation to Lombard excitation and duration HMM models (L). The results presented were obtained as part of a wider listening experiment, with entries for modifications applied to natural speech as well, described in Cooke et al. (2012) as well as the follow-up evaluation called the Hurricane Challenge (Cooke et al., 2013) with an even wider number of entries. Our entries in evaluation were published in (Valentini-Botinhao et al., 2013d).

Figure 4 shows the long term average spectrum of the baseline unmodified TTS and the modified synthetic entries: TTSGP (with GP modification only), TTSGP-DRC (with DRC as well) and TTSLGP-DRC (L modification is further combined). We can see a tendency for GP and DRC to flatten the spectral tilt. Changes in the spectral tilt observed when DRC is applied are a consequence of the temporal modifications this method performs. DRC boosts regions of the speech waveform that are of a lower level and these regions correspond to the higher frequency components of the speech spectrum, i.e. a flatter spectrum. We see the boosting effect that GP has around the formant frequency range, and the boosting that DRC gives to higher frequencies. F0 and its range (within a sentence) is increased in voices built with Lombard excitation.

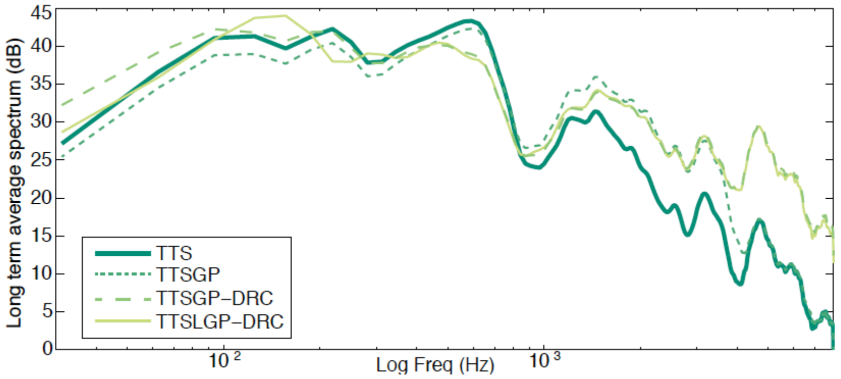

Figure 4 – Long term average spectrum calculated at a sentence level and averaged across the dataset.

Figure 5 shows the intelligibility gap between the synthetic voices and natural speech in terms of word accuracy changes. We can see that for most SNRs and modifications synthetic speech in noise is less intelligible than natural speech in noise, i.e. the gap is negative. When comparing the entries TTS, TTSGP, TTSGP-DRC and TTSLGP-DRC we can see the gain that each component adds. This addition depends on SNR, meaning that some components are more important in one condition than another. In speech-shaped noise middle SNR (SNR=-4dB) condition GP and DRC contribute most. Duration and excitation changes start contributing only at quite low SNRs. The entry TTSLGP-DRC gave a relative gain of 4.2dB in comparison to the unmodified baseline TTS.

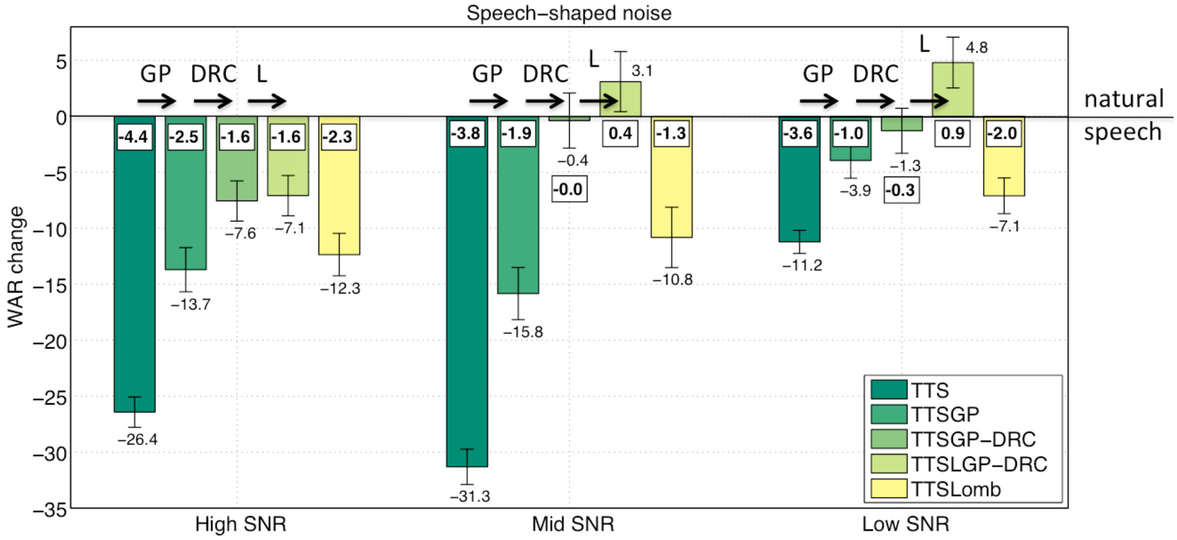

Figure 5 – Word accuracy rate (WAR) changed and dB gains relative to natural speech (values inside boxes) for synthetic speech presented in speech-shaped noise.

To put these results in context we present in Figure 6 the gain in dB of different methods obtained from a large scale evaluation, the Hurricane Challenge (Cooke et al., 2013), that included the TTSLGP-DRC and other TTS based modifications (in blue) as well as modifications made to natural speech (in green). We can see that for both competing speaker (vertical axis) and speech-shaped noise our entry improved upon the baseline, obtaining among the best results of the TTS entries. The gains compared to the TTS unmodified baseline (blue dotted line) were also comparable to some of the best natural speech entries compared to the natural speech baseline (the green dotted line). For explanation of other methods refer to (Cooke et al., 2013).

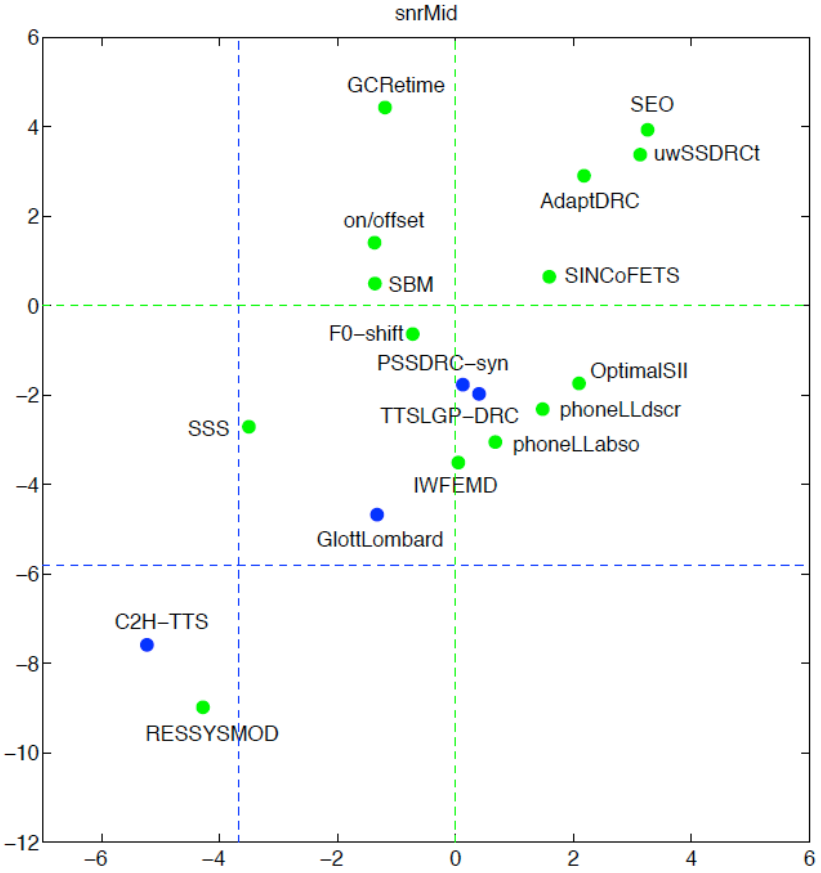

Figure 6 – Hurricane challenge (Cooke et al., 2013)results for SNR Mid: gains in dB relative to Plain (dotted green lines) and TTS baselines (dotted blue lines) for the speech shaped noise (horizontal axis) and competing speaker (vertical axis). Green: natural speech entries; blue: TTS entries.

Audio samples can be found at:
https://wiki.inf.ed.ac.uk/CSTR/TtsHc
https://wiki.inf.ed.ac.uk/CSTR/HcExternal

5 CONCLUSIONS

We set out from the idea that objective measures can be used to increase the intelligibility of synthetic speech in noise. We thought to automatically modify speech according to the environmental noise, in much the same way as humans control their speech. It transpired that not all measures can reliably predict intelligibility of synthetic speech in noise. Those that did work were based on models of the internal processing that takes place in the human auditory system. With this information our next step was to modify synthetic speech to improve intelligibility as defined by the scores from one of these measures. We observed in listening tests that spectral envelope modifications based on the glimpse proportion measure significantly increased intelligibility in stationary noise conditions, particularly if combined with a noise-independent strategy like dynamic range compression. To achieve similar gains in the competing speaker condition further changes to the excitation signal and duration, based on Lombard speech, were most effective.

Possible extensions that can follow the work include further analysis into the quality of enhanced synthetic speech. While speech intelligibility increases, naturalness and quality can be compromised, especially if the modified speech is heard in clean conditions. The most intelligible voice created in this work is a combination of different enhancing strategies: GP-based modification, dynamic range compression and Lombard adapted duration and excitation. This voice obtained up to 4.2 dB of equivalent intensity gain, however it required additional recordings of Lombard speech of the speaker for which we built the voice. It would be interesting to investigate whether similar intelligibility gains could be obtained by applying cross-speaker adaptation of duration and excitation while maintaining quality and speaker similarity.

6 ACKNOWLEDGMENTS

The research was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements 213850 (SCALE), 256230 (LISTA) and 287678 (Simple4All), and from EPSRC grants EP/I031022/1 (NST) and EP/J002526/1 (CAF).

7 REFERENCES

1. Loizou, P. C. (2007). Speech Enhancement: Theory and Practice (Signal Processing and Communications). CRC Press, Boca Raton, USA, 1 edition.
2. Bregman, A. (1990). Auditory scene analysis. MIT Press, Cambridge, USA.
3. Cooke, M. (2003). Glimpsing speech. Journal of Phonetics, 31:579 – 584.
4. Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. JASA., 115(2):833–843.
5. Macleod, A. and Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. British Journal of Audiology, 21(2):131–141.
6. Lombard, E. (1911). Le signe d'´el´evation de la voix [the sign of the elevation of the voice]. Annales des maladies de l'oreille et du larynx, 37:101–119.
7. Summers, W., Pisoni, D., Bernacki, R., Pedlow, R., and Stokes, M. (1988). Effects of noise on speech production: Acoustic and perceptual analysis. JASA., 84:917–928.
8. Junqua, J. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers. JASA., 93(1):510–524.
9. Hansen, J. (1996). Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. Sp. Com., 20:151 –173.
10. Garnier, M., Bailly, L., Dohen, M.,Welby, P., and Loevenbruck, H. (2006). An acoustic and articulatory study of Lombard speech: global effects on the utterance. In Proc. ICSLP, pp. 2246–2249, Pittsburgh, USA
11. Lu, Y. and Cooke, M. (2008). Speech production modifications produced by competing talkers, babble, and stationary noise. JASA., 124(5):3261–3275.
12. Niederjohn, R. J. and Grotelueschen, J. H. (1976). The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression. IEEE TASSPrans. on Acoustics, Speech and Signal Processing, 24(4):277– 282.
13. Skowronski, M. D. and Harris, J. G. (2006). Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments. Sp. Com., 48(5):549–558.
14. Yoo, S. D., Boston, J. R., El-Jaroudi, A., Li, C.-C., Durrant, J. D., Kovacyk, K., and Shaiman, S. (2007). Speech signal modification to increase intelligibility in noisy environments. JASA., 122(2):1138–1149.
15. McLoughlin, I. and Chance, R. (1997). LSP-based speech modification for intelligibility enhancement. In Proc. Digital Signal Processing, volume 2, pp. 591–594, Greece.

16. Raitio, T., Suni, A., Vainio, M., and Alku, P. (2011a). Analysis of HMM-based Lombard speech synthesis. In Interspeech, pp. 2781 – 2784, Florence, Italy.
17. Huang, D.-Y., Rahardja, S., and Ong, E. P. (2010). Lombard effect mimicking. In Proc. SSW, pp. 258–263, Kyoto, Japan.
18. Patel, R., Everett, M., and Sadikov, E. (2006). Loudmouth: Modifying text-to-speech synthesis in noise. In SIGACCESS Conf. on Computers and Accessibility, pp. 227–228, New York.
19. Zorila, T. C., Kandia, V., and Stylianou, Y. (2012). Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression. In Interspeech, Portland.
20. Langner, B. and Black, A. W. (2005). Improving the understandability of speech synthesis by modeling speech in noise. In Proc. ICASSP, volume 1, pp. 265–268, Philadelphia.
21. Picart, B., Drugman, T., and Dutoit, T. (2011). Continuous control of the degree of articulation in HMM based speech synthesis. In Interspeech, pp. 1797 – 1800, Florence, Italy.
22. Sauert, B. and Vary, P. (2006). Near end listening enhancement: Speech intelligibility improvement in noisy environments. In Proc. ICASSP, pp. 493–496, Toulouse, France.
23. Tang, Y. and Cooke, M. (2010). Energy reallocation strategies for speech enhancement in known noise conditions. In Interspeech, pp. 1636–1639, Makuhari, Japan.
24. Sauert, B. and Vary, P. (2010). Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement. In Proc. Fachtagung Sprachkommunikation, volume 9, Bochum, Germany.
25. Sauert, B. and Vary, P. (2011). Near end listening enhancement considering thermal limit of mobile phone loudspeakers. In Proc. Conf. on Elektronische Sprachsignalverarbeitung, volume 61, pp. 333–340, Aachen, Germany.
26. Tang, Y. and Cooke, M. (2012). Optimised spectral weightings for noise-dependent speech intelligibility enhancement. In Interspeech, Portland, USA.
27. Tang, Y. and Cooke, M. (2011). Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints. In Interspeech, pp. 345–348, Florence, Italy.
28. Aubanel, V. and Cooke, M. (2013). Information-preserving temporal reallocation of speech in the presence of fluctuating maskers. In Interspeech, Lyon, France.
29. Taal, C. H., Hendriks, R. C., and Heusdens, R. (2012). A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure. In Proc. ICASSP, pp. 4061–4064, Kyoto, Japan.
30. Petkov, P., Kleijn, B., and Henter, G. (2012). Enhancing subjective speech intelligibility using a statistical model of speech. In Interspeech, Portland, USA.
31. Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., and Tang, Y. (2012). Evaluating the intelligibility benefit of speech modifications in known noise conditions. Sp. Com., 55:572–585.
32. Cooke, M., Mayo, C., and Valentini-Botinhao, C. (2013). Intelligibility-enhancing speech modifications: the Hurricane Challenge. In Interspeech, Lyon, France.
33. Valentini-Botinhao, C., Yamagishi, J., and King, S. (2011b). Evaluation of objective measures for intelligibility prediction of HMM-based synthetic speech in noise. In Proc. ICASSP, pp. 5112–5114, Prague, Czech Republic.
34. Valentini-Botinhao, C., Yamagishi, J., and King, S. (2011a). Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise? In Interspeech, pp. 1837 – 1840, Florence, Italy.
35. Valentini-Botinhao, C., Yamagishi, J., and King, S. (2012c). Mel cepstral coefficient modification based on the Glimpse Proportion measure for improving the intelligibility of HMM-generated synthetic speech in noise. In Interspeech, Portland, USA.
36. Valentini-Botinhao, C., Yamagishi, J., King, S., and Maia, R. (2013c). Intelligibility enhancement of HMM-generated speech in additive noise by modifying mel cepstral coefficients to increase the glimpse proportion. Computer Speech and Language.
37. Valentini-Botinhao, C., Yamagishi, J., King, S., and Stylianou, Y. (2013d). Combining perceptually-motivated spectral shaping with loudness and duration modification for intelligibility enhancement of HMM-based synthetic speech in noise. In Interspeech, Lyon, France.