Intelligibility-enhancing speech modifications: the Hurricane Challenge

Martin Cooke[1,2]*, Catherine Mayo*[3]*, Cassia Valentini-Botinhao*[3]

[1]Basque Foundation for Science, Bilbao, Spain
[2]Language and Speech Laboratory, University of the Basque Country, Vitoria, Spain
[3]Centre for Speech Technology Research, University of Edinburgh, UK

m.cooke@ikerbasque.org, catherin@inf.ed.ac.uk, C.Valentini-Botinhao@sms.ed.ac.uk

Abstract

Speech output is used extensively, including in situations where correct message reception is threatened by adverse listening conditions. Recently, there has been a growing interest in algorithmic modifications that aim to increase the intelligibility of both natural and synthetic speech when presented in noise. The Hurricane Challenge is the first large-scale open evaluation of algorithms designed to enhance speech intelligibility. Eighteen systems operating on a common data set were subjected to extensive listening tests and compared to unmodified natural and text-to-speech (TTS) baselines. The best-performing systems achieved gains over unmodified natural speech of 4.4 and 5.1 dB in competing speaker and stationary noise respectively, while TTS systems made gains of 5.6 and 5.1 dB over their baseline. Surprisingly, for most conditions the largest gains were observed for noise-independent algorithms, suggesting that performance in this task can be further improved by exploiting information in the masking signal.

Index Terms: intelligibility, speech modification, TTS

1. Introduction

Speech output – whether from mobile phones, public address systems or simply domestic audio devices – is widely used. In many listening contexts the intelligibility of the intended message might be compromised by environmental noise or channel distortion. Problems can be minimised by increasing output intensity or repeating the message, but these approaches are not ideal for either the listener (e.g. discomfort, stress; see [1]) or the output device (e.g. power consumption, failure). A better approach is to seek ways to modify the speech signal to increase intelligibility in noise. The need for more robust speech output is particularly pressing for TTS systems, whose intelligibility in noise falls short of naturally-produced speech [2, 3].

While modification algorithms have been studied for some time in audio [4] and speech technologies [5, 6], recent years have witnessed a renewed interest in tackling what has been termed the 'near-end' speech enhancement problem [7–15]. Consequently it is of interest to compare their performance using shared data and metrics.

The idea of a common evaluation of algorithms was piloted in 2012 within the EU-funded 'Listening Talker' project. That study [16] compared 7 speech modification algorithms against read and Lombard speech and an unmodified TTS system. The best techniques led to substantial gains over baseline. The Hurricane Challenge extends the pilot study to an open international evaluation of algorithms, the results of which are reported here. Further details of the Challenge can be found at http://listening-talker.org/hurricane.

2. The Challenge problem

Entrants to the Challenge (section 3) were provided with a corpus of speech and noise waveforms (section 2.1), as well as optional data resources to construct/adapt a TTS system (section 2.2). Entrants then returned algorithmically-modified or synthetically-generated speech waveforms for the entire corpus. These were subjected to evaluation by listeners (section 4). Entrants had around 6 weeks to prepare their modified signals, and all made a financial contribution to the cost of listening tests.

2.1. Speech and noise corpora

The 'Plain' unmodified natural speech corpus consists of the first 180 sentences of the Harvard corpus [17] read by a male British English speaker. The Harvard corpus contains sentences such as "the salt breeeze came across from the sea" arranged into phonemically-balanced subsets. The Plain corpus was elicited as read speech from a highly-intelligible speaker, and can therefore be considered as intrinsically rather clear (i.e. hyper-articulated).

Entrants also received six sets of noise waveforms for each utterance arising from the combination of two masker types at three signal-to-noise ratios (SNRs). The noise conditions were (i) a *fluctuating* masker, which was competing speech (CS) from a female talker producing read speech scaled to produce utterance-wide SNRs of -7, -14 and -21 dB; and (ii) a *stationary* masker, which was speech-shaped noise (SSN) whose long-term average spectrum matched that of the CS, at SNRs of 1, -4 and -9 dB. Entrants therefore had access to separate speech and masker signals as well as SNRs at which these would be subsequently combined for listener evaluation. Speech was centrally-embedded in the noise with 0.5s lead/lag intervals. Entrants were permitted to modify the overall duration of the speech within these limits (i.e. a maximum total extension of 1s). All materials were provided at a sampling rate of 16 kHz.

2.2. TTS

In addition to the speech and noise waveforms outlined above, those entrants wishing to submit a TTS entry had available two natural speech datasets (spoken by the same speaker who produced the Plain material) and associated orthographic transcriptions. One consists of about 3 hours of additional unmodified natural speech for three different reading materials: 2023 newspaper style sentences, 300 sentences containing words from the modified rhyme test [18] inserted in the carrier sentence 'Now we will say *word* again', and the remaining 540 Harvard sentences not used in the evaluation. The second dataset consists of just under 1 hour of Lombard speech from the same speaker who produced the Plain corpus, recorded with speech modu-

lated noise from a male speaker [19] played at 84 dBA over headphones. This dataset consists of the same reading material as the Plain set with the exception of the newspaper sentences. Both datasets were sampled at 96 kHz.

3. Challenge entries

Each entry has a short name which is used in the results presentation. Many entries will be reported in full papers at Interspeech 2013 using the same identifiers.

AdaptDRC: AdaptDRC aims at enhancing speech content at high frequencies as well as boosting low energy speech content in conditions of low predicted intelligibility. It applies a time- and frequency-dependent dynamic range compression (DRC) and frequency-shaping (FS) in octave bands. The amount of DRC and FS is controlled by an estimate of the Speech Intelligibility Index (SII).

F_0-shift: F_0 is shifted per-utterance to maximise an objective intelligibility metric based on energetic masking. Predicted intelligibility is typically highest for large downward shifts in F_0, whose effect is to increase the number of resolved harmonic components in an auditory-scaled speech representation.

GCRetime: Local speech rate is modified to minimise overlap with a known fluctuating masker. Continuous time-scale factors are derived from an optimisation procedure applied to the energetic masking relations of the speech and noise mixture [20] supplemented by the identification of potentially most informative speech regions [21]. Intelligibility gains come from energetic masking release, particularly in the time domain.

IWFEMD: Intrinsic mode functions (IMF) of empirical mode decomposition (EMD) [22] representing speech are modified based on an inverse Wiener filter. Without reducing the time-frequency resolution, the enhancement process for voiced speech is performed on low frequency IMFs, in which harmonic components are detected. For the unvoiced consonants, this enhancement process is performed on the high frequency IMFs.

on/offset: Speech components such as bursts and vocalic onsets/offsets are selected using an extrapolation-based detector [23] and amplified with a variable gain. The additional power used for amplification is taken from the strong voiced components. The main goal was to subjectively evaluate this basic time domain speech modification method without considering modification of the spectral information.

OptimalSII: A linear time-invariant filter is designed which redistributes speech energy over frequency to maximise the SII. Using a nonlinear approximation to the SII, a closed-form solution could be found to the power-constrained optimisation problem [24]. Note that this is not the same as the OptSII system described in [16].

phoneLLabso: A recogniser trained on WSJ0 [25] provides phone segmentation information and associates signal frames with acoustic models. Phone energy, normalised by duration, is equalised for all phones in the sentence. The log-likelihood (LL) of noisy frame-based features is maximised for each phone in isolation, conditioned on the correct acoustic model, for a set of band-gain adjustment coefficients under an energy-preservation constraint. A noise PSD estimator from past observations enables the computation of noisy features.

phoneLLdscr: This entry builds on phoneLLabso, augmenting the objective measure with the difference of the measure in phoneLLabso, and the log of the sum of likelihoods conditioned on alternative acoustic models [26]. To reduce complexity, the context (phone neighbours) is assumed known and only a subset of all alternative models is considered based on the proximity of their LL scores to that obtained by the correct model.

RESSYSMOD: Perceptually-significant features of the excitation and vocal tract system are modified to increase the perceived loudness of speech. Impulse-like excitation around glottal closure instants and sharpness of formants are major contributors to perceived loudness. Modifications sharpen these two features according to the level of degradation.

SBM: A spectrum binary mask for the clean speech is calculated by comparing the short-time Fourier spectra of speech and noise. At each frequency point, the SBM is set to 1 if the speech spectrum amplitude is larger than the noise, otherwise 0. The processed speech spectrum is obtained by multiplying the SBM with the original clean spectrum. The modified speech is re-synthesized by inverse Fourier transform/overlap-add.

SEO: Spectral energy is optimised retaining and emphasising acoustical features important for speech perception. Three processing methods (flattening spectral tilt, enhancement of spectrum contrast and retaining harmonics components in the low frequency region) are combined. The processing is performed with fixed parameters determined by consideration of the energy balance of the three processed parts.

SINCoFETS: This system combines different noise-dependent and independent algorithms. Non-uniform time-scaling is used to slow down the speech and redistribute the available time between the vowels and consonants (cf [27]). Dynamic range compression is applied to decrease amplitude differences between vowels and consonants. Finally, if severely degraded SNR levels are detected, the system applies psycho-acoustic based adaptive equalisation to improve intelligibility robustness against the detected noise (cf [10]).

SSS: Steady-state portions of speech (syllable nuclei) are detected from spectral transitions and their amplitudes are suppressed, given their lesser importance for speech perception and their greater energy compared with transient portions (syllable onset and coda) [28]. Since SSS suppresses steady-state portions and hence relatively enhances transient portions when compared with an unprocessed signal at the same SNR, it is expected to increase speech intelligibility in noise.

uwSSDRCt: This entry incorporates additional spectral and time domain modifications into the Spectral Shaping and Dynamic Range Compression method [15]. (i) Speech is uniformly time stretched within the constraints of the Challenge in order to increase signal redundancy; (ii) a frequency warping approach to vowel space expansion is incorporated into the SS; (iii) scaling to enhance the transient regions of speech is applied in the time-domain along with DRC.

TTS: The TTS entry was a voice built by adapting a high quality average voice model to the Plain dataset provided. The training and adaptation data had a sampling rate of 48 kHz. To train and adapt speech the following were extracted: 59 Mel cepstral coefficients with $\alpha = 0.77$, Mel scale F_0, and 25 aperiodicity energy band. See [16] for more details.

TTSLGP-DRC: The excitation and duration parameters of the voice 'TTS' were adapted to the Lombard dataset provided in order to mimic a speaker's Lombard duration and F_0 changes. To enhance the spectral envelope a noise-dependent optimisation based on the glimpse proportion measure was performed [29]. Finally, DRC was applied on the generated waveform to boost the lower level regions of speech.

C2H-TTS: This entry is a HMM-based TTS system inspired by the C2H model of Hyper- and Hypo-articulated speech production [12, 30]. Transformations on synthetic speech aim to control phonetic contrast by increasing/reducing

the acoustic distance between what are hypothesised to be low-energy attractors for both human and synthetic speech. In this instance, the system was applied to achieve the maximum degree of hyper-articulated speech, i.e. maximum phonetic contrast.

GlottLombard: A TTS voice trained from modal speech [31] was transformed to a Lombard voice by modifying glottal pulse shape, spectral tilt, harmonic-to-noise ratio and F_0. The modifications were applied in unsupervised fashion based on a few utterances of Lombard speech from the target speaker. In addition, DRC and formant-sharpening were applied to increase noise robustness.

PSSDRC-syn: HMM synthesis plus noise-independent modifications at vocoder level: (1) amplification of the 1-4 kHz band; (2) postfiltering with a voicing probability dependent factor; (3) F_0 increment by factor 1.2; (4) standard deviation of log-F_0 multiplied by factor 1.5; (5) uniform lengthening of the signal up to 120%; (6) DRC applied to the energy contour.

4. Listener evaluation

The subjective intelligibility of the 20 entries was measured in 6 noise conditions using a total of 21600 stimuli (20 entries x 180 sentences x 2 maskers x 3 SNRs) divided into blocks of 30. Within each block, entries were mixed such that by listening to 6 blocks (=180 sentences) a single participant would hear 9 sentences from each entry. A balanced design assigned listeners to blocks such that (i) each listener heard one block of 30 sentences in each of the six noise conditions, (ii) no listener heard the same sentence twice, and (iii) each noise condition was heard by the same number of listeners.

Young adult listeners (predominantly 19-27 years old) were recruited via the University of Edinburgh Student and Graduate Employment service. Listeners were required to be native English talkers, to report no history of speech and/or language disorders and to pass an audiological screening; 175 listeners met these criteria. All were paid for their participation.

Modified speech entries were combined with maskers at each SNR, computed over the region where the speech was present (entrants who modified the original speech duration also provided endpoint markers for the modified speech). Stimuli were normalised to have the same root-mean-square level and presented to participants in dedicated, sound-attenuated listening booths at the University of Edinburgh using Beyerdynamic DT770 headphones. Listeners were given two short practice sessions, one per masker type, presented at 0 dB SNR for SSN and -3 dB for CS, using Plain speech Harvard sentences from outside the sets used for the main test. Stimuli were presented once only, and listeners could not change the output level. Custom-built MATLAB software controlled the presentation of stimuli and collection of responses. Participants were instructed to type what they had heard rather than attempt to reconstruct the whole sentence. The subsequent stimulus was presented automatically after the entry of a response. Null responses were not permitted: listeners typed 'X' for those sentences when no words were intelligible. The listening test was completed on average in 40-45 minutes.

Responses were scored in terms of number of words correctly identified. Short words ('a', 'the', 'in', 'to', 'on', 'is', 'and', 'of', 'for', 'at') were not scored. Prior to scoring, both reference sentence lists and listener responses were edited to remove punctuation. A custom dictionary was employed to match common response alternatives (e.g. 'sideshow' vs 'side show', '50' vs 'fifty').

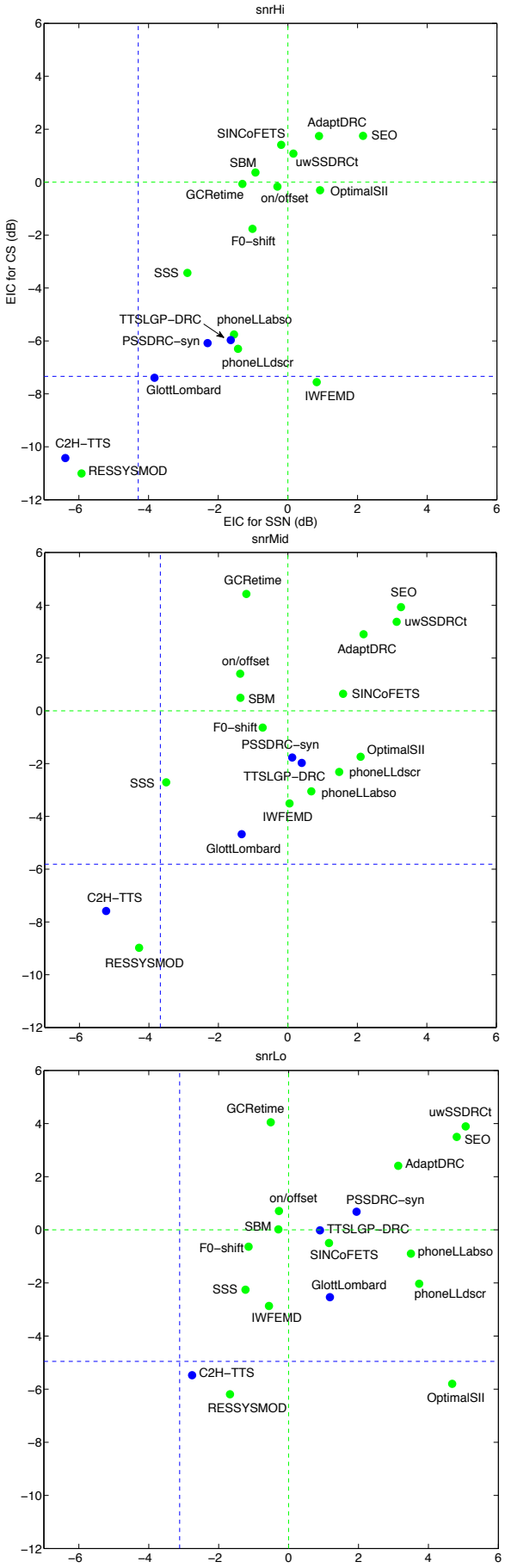

Figure 1: *EICs in dB re Plain/TTS baselines (dotted lines) for the SSN and CS maskers. Green: natural speech; blue: TTS.*

	Noise Dependent?	Duration modified?	gains in CS			gains in SSN		
			snrHi	snrMid	snrLo	snrHi	snrMid	snrLo
re. Plain			85.1	57.0	24.8	88.3	63.0	17.3
AdaptDRC	yes	no	1.7 (4)	2.9 (12)	2.4 (9)	0.9 (3)	2.2 (15)	3.1 (20)
F_0-Shift	yes	no	-1.8 (-5)	-0.6 (-3)	-0.6 (-2)	-1.0 (-4)	-0.7 (-6)	-1.1 (-5)
GCRetime	yes	yes	-0.1 (0)	**4.4 (18)**	**4.0 (16)**	-1.3 (-5)	-1.2 (-10)	-0.5 (-2)
IWFEMD	yes	no	-7.6 (-26)	-3.5 (-16)	-2.9 (-9)	0.8 (3)	0.0 (0)	-0.6 (-3)
on/offset	no	no	-0.2 (0)	1.4 (6)	0.7 (3)	-0.3 (-1)	-1.4 (-11)	-0.3 (-1)
OptimalSII	yes	no	-0.3 (-1)	-1.7 (-8)	-5.8 (-15)	0.9 (3)	2.1 (15)	4.7 (33)
phoneLLabso	yes	yes	-5.8 (-19)	-3.1 (-14)	-0.9 (-3)	-1.5 (-7)	0.7 (5)	3.5 (23)
phoneLLdscr	yes	yes	-6.3 (-21)	-2.3 (-11)	-2.0 (-6)	-1.4 (-6)	1.5 (11)	3.7 (25)
RESSYSMOD	no	no	-11.0 (-42)	-9.0 (-37)	-6.2 (-15)	-5.9 (-38)	-4.3 (-35)	-1.7 (-7)
SBM	yes	no	0.4 (1)	0.5 (2)	0.0 (0)	-0.9 (-4)	-1.4 (-11)	-0.3 (-1)
SEO	no	no	**1.7 (4)**	3.9 (16)	3.5 (14)	**2.2 (6)**	**3.3 (21)**	4.8 (34)
SINCoFETS	yes	yes	1.4 (3)	0.6 (3)	-0.5 (-2)	-0.2 (-1)	1.6 (12)	1.2 (6)
SSS	no	no	-3.4 (-10)	-2.7 (-12)	-2.3 (-7)	-2.9 (-14)	-3.5 (-29)	-1.2 (-5)
uwSSDRCt	no	yes	1.1 (2)	3.4 (14)	3.9 (16)	0.2 (1)	3.1 (20)	**5.1 (37)**
TTS	no	yes	-7.3 (-25)	-5.8 (-26)	-5.0 (-13)	-4.3 (-25)	-3.7 (-30)	-3.1 (-11)
re. TTS			59.7	31.3	11.7	63.7	32.8	6.8
TTSLGP-DRC	yes	yes	**1.4 (6)**	3.8 (17)	4.9 (13)	**2.7 (17)**	**4.1 (33)**	4.0 (15)
C2H-TTS	yes	yes	-3.1 (-14)	-1.8 (-7)	-0.5 (-1)	-2.1 (-17)	-1.6 (-10)	0.4 (1)
GlottLombard	no	yes	-0.1 (0)	1.1 (5)	2.4 (5)	0.5 (4)	2.3 (19)	4.3 (17)
PSSDRC-syn	no	yes	1.3 (5)	**4.0 (18)**	**5.6 (16)**	2.0 (14)	3.8 (31)	**5.1 (22)**
		Fisher LSD	2.0 (4.7)	1.2 (5.5)	1.3 (4.6)	1.2 (4.2)	0.7 (5.2)	1.0 (4.8)

Table 1: *Changes relative to Plain and TTS baselines for Hurricane 2013 entries, expressed as EICs in dB, with percentage points changes in keyword scores in parentheses. Entries with the largest gains for each noise type/SNR combination are highlighted. The keywords correct scores expressed in absolute percentages for the Plain and TTS baselines are also provided as well as Fisher's LSD values.*

Intelligibility gains/losses for each entry over the appropriate Plain or TTS baseline are shown in Figure 1. Gains are expressed as equivalent intensity changes (EICs) computed by mapping scores to psychometric curves previously obtained for each masker using Plain speech (see [16] for details). EICs are plotted for SSN against CS to permit a clearer visualisation of which methods are beneficial for one or both types of masker.

Table 1 lists changes relative to Plain in dBs and percentage points. The largest gains for both natural and synthetic entries in each masker condition are highlighted. To permit comparison of entries, Fisher's least significant differences in dBs and percentage points are also tabulated, computed using separate ANOVAs for each SNR level and masker type with a single factor of modification entry.

5. Discussion

Large intelligibility gains equivalent to boosting the level of unmodified speech by up to 5.6 dB were observed, with similar-sized increases over both natural and TTS baselines and for both types of masker. These gains are substantial, reaching up to 37 percentage points of word accuracy. Larger gains were seen at mid and low SNRs, perhaps due to the limited scope for improvement over the baseline in the high SNR conditions, although it is notable that TTS systems operating from a lower baseline also showed smaller gains in the high SNR condition.

A high degree of masker preference can be seen in these results. For natural speech, only 3 methods (SEO, uwSSDRCt, AdaptDRC) produced significant gains for both CS and SSN maskers. Other approaches (optimalSII, phoneLLdscr, phoneLLabso, SINCoFETS) performed well in stationary noise but were more or less harmful for the non-stationary case, where GCRetime scored well.

Not surprisingly, Plain speech was more intelligible than unmodified TTS, although the gap reduced with decreasing SNR from around 4/7 dB to 3/5 dB for SSN/CS respectively. However, one striking outcome of the Challenge is the finding that three modified TTS entries (PSSDRC-syn, TTSLGP-DRC, GlottLombard) reached and even exceeded the intelligibility level of Plain speech in stationary noise, with PSSDRC-syn also showing marginal gains for the CS masker. As noted earlier, the Plain utterances were intrinsically clear, and to boost TTS beyond that level is a significant achievement.

Intriguingly, there was no clear advantage for entries that used prior knowledge of the masker. In fact, two of the best techniques overall for natural speech (SEO, uwSSDRCt) were noise-independent, as was PSSDRC-syn for TTS. Durational changes were used by nearly half of natural speech entries and all TTS systems and appear to have contributed to good performance in several cases, especially for the GCRetime approach which exploits temporal fluctuations in the masker. The performance of SEO is of note given that it exploited neither durational expansion nor knowledge of the masker signal.

While detailed discussion of individual modification algorithms and their components is outside the scope of this summary article, it is clear that most of the natural and TTS entries that incorporated dynamic range compression (AdaptDRC, uwSSDRCt, TTSLGP-DRC, PSSDRC-syn) performed well.

In conclusion, the first large-scale open evaluation of speech modification algorithms designed to enhance intelligibility has demonstrated worthwhile gains over a relatively-clear unmodified speech baseline. It is to be hoped that synergistic combination of techniques or their components is possible, leading to larger gains. Other factors which might be measured in future comparisons include speech quality, perceived loudness and computational complexity.

Acknowledgements. We thank all the entrants for their timely responses at each stage of the Challenge, and Anna Naxiadou and Vasilis Karaiskos for helping to run the listening tests. The research leading to these results was partly funded from the European Community's 7th Framework Programme (FP7/2007-2013) under grant agreement 213850 (SCALE) and by the Future and Emerging Technologies (FET) programme under FET-Open grant number 256230 (LISTA).

6. References

[1] WHO, "Burden of disease from environmental noise: Quantification of healthy life years lost in Europe," World Health Organisation, 2011.

[2] H. Venkatagiri, "Segmental intelligibility of four currently used text-to-speech synthesis methods," *J. Acoust. Soc. Am.*, vol. 113, pp. 2095–2104, 2003.

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

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

[5] 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.

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

[7] 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.

[8] 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, no. 5, pp. 549–558, 2006.

[9] 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.

[10] H. Brouckxon, W. Verhelst, and B. D. Schuymer, "Time and frequency dependent amplification for speech intelligibility enhancement in noisy environments," in *Proc. Interspeech*, 2008, pp. 557–560.

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

[12] R. K. Moore and M. Nicolao, "Reactive speech synthesis: Actively managing phonetic contrast along an H&H continuum," in *ICPhS 2011*, Hong Kong, China, 2011, pp. 1422–1425.

[13] 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, 2011, pp. 333–340.

[14] 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*, 2012, pp. 4061–4064.

[15] 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.

[16] 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 Communication*, vol. 55, pp. 572–585, 2013.

[17] E. H. Rothauser, W. D. Chapman, N. Guttman, H. R. Silbiger, M. H. L. Hecker, G. E. Urbanek, K. S. Nordby, and M. Weinstock, "IEEE Recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[18] A. S. House, C. Williams, M. H. L. Hecker, and K. D. Kryter, "Psychoacoustic speech tests: A modified rhyme test," *J. Acoust. Soc. Am.*, vol. 35, no. 11, pp. 1899–1899, 1963.

[19] W. A. Dreschler, H. Verschuure, C. Ludvigsen, and S. Westermann, "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing aid assessment," *Audiology*, vol. 40, pp. 148–157, 2001.

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

[21] C. Stilp and K. Kluender, "Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility," *Proceedings of the National Academy of Sciences*, vol. 107, no. 27, pp. 12 387–12 392, 2010.

[22] M. E. Hamid, S. Das, K. Hirose, and M. K. I. Molla, "Speech enhancement using EMD-based adaptive soft-thresholding (EMD-ADT)," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 5, no. 2, June 2012.

[23] R. Dokku and R. Martin, "Detection of stop consonants in continuous noisy speech based on an extrapolation technique," in *Proc. EUSIPCO*, 2012, pp. 2338–2342.

[24] C. Taal, J. Jensen, and A. Leijon, "On optimal linear filtering of speech for near-end listening enhancement," *Signal Processing Letters, IEEE*, vol. 20, no. 3, pp. 225 –228, 2013.

[25] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. Workshop Speech Natural Lang.*, 1992, pp. 357–362.

[26] P. N. Petkov, G. E. Henter, and W. B. Kleijn, "Maximizing phoneme recognition accuracy for enhanced speech intelligibility in noise," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 21, no. 5, pp. 1035–1045, 2013.

[27] M. Demol, W. Verhelst, K. Struyve, and P. Verhoeve, "Efficient non-uniform time-scaling of speech with WSOLA," in *Proc. ISCA-ITRW Multiling 2006*, Stellenbosch, South Africa, 2006.

[28] N. Hodoshima, T. Arai, A. Kusumoto, and K. Kinoshita, "Improving syllable identification by a preprocessing method reducing overlap-masking in reverberant environments," *J. Acoust. Soc. Am.*, vol. 119, pp. 4055–4064, 2006.

[29] 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, 2012.

[30] 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, 2012.

[31] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 19, no. 1, pp. 153–165, 2011.