Information-preserving temporal reallocation of speech in the presence of fluctuating maskers

Vincent Aubanel[1,2], *Martin Cooke*[1,2]

[1]Language and Speech Laboratory, University of the Basque Country, Vitoria, Spain
[2]Ikerbasque (Basque Foundation for Science)

v.aubanel@laslab.org

Abstract

How can speech be retimed so as to maximise its intelligibility in the face of competing speech? We present a general strategy which modifies local speech rate to minimise overlap with a known fluctuating masker. Continuous time-scale factors are derived in an optimisation procedure which seeks to minimise overall energetic masking of the speech by the masker while additionally unmasking those speech regions potentially most important for speech recognition. Intelligibility increases are evaluated with both objective and subjective measures and show significant gains over an unmodified baseline, with larger benefits at lower signal-to-noise ratios. The retiming approach does not lead to benefits for speech mixed with stationary maskers, suggesting that the gains observed for the fluctuating masker are not simply due to durational expansion.

Index Terms: speech intelligibility, temporal modification, energetic and informal masking

1. Introduction

Background speech can be disrupting, as anyone who has attempted to maintain a conversation in a busy cafe will know. Speech from competing talkers has both an energetic and informational masking effect, diverting attention and at times preventing successful message decoding. In adverse conditions, speakers usually react by raising their voice and modifying other speech parameters – collectively described as the Lombard effect (e.g., [1, 2]). More recently, it has been shown that speakers also use temporal adaptation to reduce overlap with a fluctuating masker [3, 4], retiming their productions in what appears to be a "wait-and-talk" strategy [5]. Inspired by these behavioural results, the current study investigates their application to temporal modifications of recorded speech whose aim is to increase intelligibility when mixed with non-stationary maskers, in the Hurricane Challenge evaluation framework [6].

One focus of the current study concerns which speech epochs it is most beneficial to 'unmask'. Not all speech segments contribute equally to intelligibility. Vowels have been found to be the primary determinant of sentence intelligibility [7, 8], since replacing them by noise is more disruptive to comprehension than when either consonants or CV/VC transitions are substituted [9]. The greater relative advantage of vowels appears to be due to the fact that in addition to segmental information, they carry sentence-level cues through the amplitude envelope [10] and fundamental frequency contour, facilitating syllabification and providing prosodic cues for sentence recognition [11, 12, 13].

However, a recent study [14] proposed an alternative, information-theoretic, characterisation of the speech stream, arguing that since perceptual systems are most sensitive to change, the rapidly-changing parts of speech should carry the most salient information for perception. The Cochlea-scaled Spectral Entropy (CSE) index [14] better predicted intelligibility than the traditional vowel-consonant distinction. Notwithstanding the importance of vowels for intelligibility – which can be captured through their greater resistance to energetic masking – the information-theoretic approach has a complementary appeal in situations where speech is masked by a fluctuating source, and where only limited fragments are available for processing.

Section 2 presents the general framework of the method, highlighting the contributions of metrics for both energetic masking and salient speech information. Specific implementation details for the Hurricane Challenge scenario are provided in section 3. The objective and subjective performance of the approach is quantified in section 4, while possible applications and extensions are summarised in section 5.

2. Temporal reallocation of speech

2.1. General framework

Temporal reallocation can be achieved in general by strategies such as localised expansion and compression of speech as well as expansion and compression of silent intervals, i.e., pauses, using an objective cost function designed to find temporal changes which improve the predicted intelligibility of the output speech in the presence of a known masker. Optimal time warping of the speech can be carried out using dynamic programming.

Our focus here is on fluctuating maskers. The proposed cost function is based on two linked ideas. The first is to exploit any gaps (or, more generally, regions of reduced energy) in the masker in order to reallocate speech into regions which, overall, minimise the degree of energetic masking. The second is to identify and give priority, in the 'unmasking' process, to speech regions which potentially contain the most salient information for speech perception.

Energetic masking is estimated using the glimpse proportion metric (GP) [15], computed as the percentage of spectro-temporal regions in modelled auditory excitation patterns whose local SNR exceeds a certain threshold α, expressed in decibels (dB):

$$GP = \frac{1}{TF} \sum_{f=1}^{F} \sum_{t=1}^{T} \mathcal{H}[\, S(t,f) - M(t,f) - \alpha \,] \qquad (1)$$

where T and F are the number of time frames and frequency

channels, $S(t, f)$ and $M(t, f)$ denote the spectro-temporal excitation pattern values in dB of speech and masker at time t and at frequency f, and $\mathcal{H}[\cdot]$ is the unit step function. Excitation patterns are derived by filtering with a 32-channel gammatone filterbank [16] using an implementation described in [17]. Centre frequencies cover the $50 - 7500$ Hz range, linearly-spaced on the equivalent rectangle bandwidth scale. The Hilbert envelope of each filter output is computed and smoothed by a leaky integrator with a 8 ms time constant [18]. Finally, the smoothed envelope is log-compressed and downsampled to 100 Hz (i.e., 10 ms frame rate).

The index used to identify supposedly informative regions is Cochlea-scaled spectral entropy [14] (CSE), computed as the running sum of Euclidean distance d between successive 16 ms adjacent frames of auditorily-transformed speech spectra:

$$d^2(t) = \sum_{f=1}^{F} [S(t+1, f) - S(t, f)]^2 \tag{2}$$

$$CSE(t) = \sum_{k=-b/2}^{b/2} d(t+k) \tag{3}$$

where b the number of the adjacent frames over which to sum. Following [14] we use $b = 5$ (i.e., 80 ms), which corresponds roughly to the mean consonant duration.

The task of per-utterance global maximisation of glimpse proportion while unmasking high-CSE regions can now be expressed as an optimisation problem which seeks the best realignment of the speech by minimising the local cost function c, defined in eqn. 4. We distinguish between two versions of the local cost function which differ in whether CSE weighting is applied (eqn. 4b) or not (eqn. 4a). Given speech and masker time frames i and j respectively, the local cost matrices can be expressed as:

$$c(i, j) = GP_f(i, j) \tag{4a}$$

$$c(i, j) = GP_f(i, j)\, W_{\text{CSE}}(i) \tag{4b}$$

where $GP_f(i, j)$ represents the proportion of frequency channels in speech frame i glimpsed in the presence of frame j of the masker, i.e.,

$$GP_f(i, j) = \frac{1}{F} \sum_{f=1}^{F} \mathcal{H}[\, S(i, f) - M(j, f) - \alpha\,] \tag{5}$$

Here, $\alpha = 0$ was used. Finally, W_{CSE} boosts high-CSE regions (defined as those higher than a value β) by a factor w:

$$W_{\text{CSE}} = (w - 1)\, \mathcal{H}[CSE - \beta] + 1 \tag{6}$$

2.2. Alternative temporal strategies

Depending on the application scenario, different types of temporal adjustment strategies might be preferred:

1. The most general form of adjustment involves **arbitrary expansion and compression** of different regions of speech and pauses. This might be used, for instance, to introduce a pause to avoid an energetic stretch of the masker signal.
2. In situations with more relaxed time constraints, an **expansion-only** approach would avoid potentially harmful effects of faster speech rates.
3. To avoid possibly disruptive effects of speech rate changes, an alternative approach is to maintain the original speech rate but to **insert pauses** to avoid the masker.

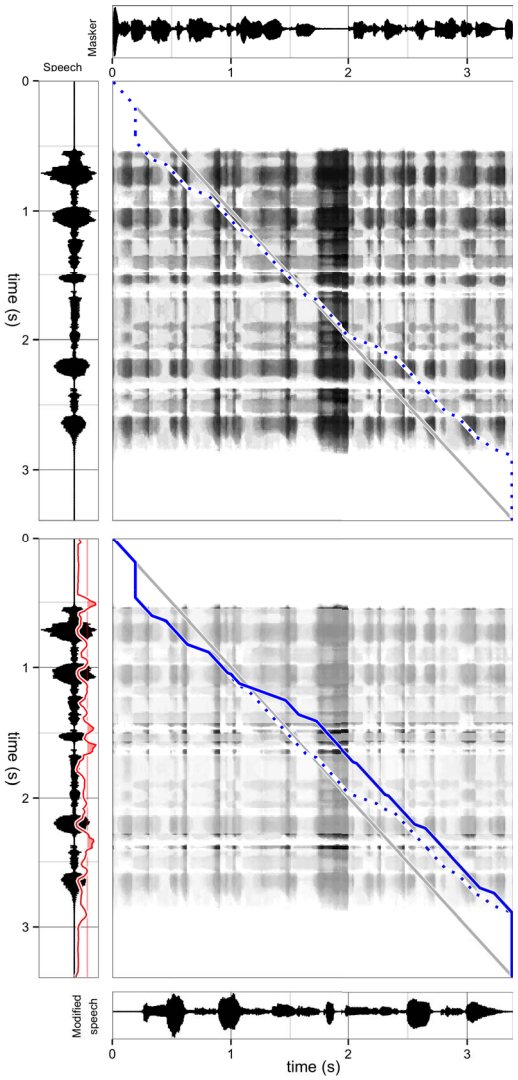

Figure 1: *Retiming of "*A large size in stockings is hard to sell*" in the presence of a competing speech masker. Upper panel: GP cost matrix for the mixture, with the optimal path shown as the dotted line. High GP regions appear darker. Lower panel: CSE-weighted cost matrix. The CSE index is overlaid on the speech signal, with high-CSE regions shaded. The solid line shows the best path, with the best path for GP-only drawn for comparison. Here, the high-CSE segments* [st] *and* [k] *in "*stockings*" are time-shifted to a low-energy part of the masker (1.7–2.0 s). Modified speech is shown at the base of the Figure.*

These techniques can be implemented within the general framework by defining different subsets of permitted path transitions within the dynamic programming optimisation process. For example, expansion-only can be achieved using an asymmetric transition function, and pause insertion by maintaining the original speech rate by allowing only diagonal transitions in addition to local cost-free horizontal transitions at pre-identified speech locations (e.g., word boundaries).

3. Hurricane Challenge: implementation

In the Challenge the active speech content is 1 s shorter than the masker, relaxed time constraints that favour the use of either an expansion-only or pause-insertion approach. The former was chosen since pause insertion has been shown in one study [19] to produce a significant reduction in intelligibility for sentence material, possibly since pause insertion in read speech can disrupt a listener's ability to detect the locations of words presented under masked conditions. To implement the expansion-only approach, the following path constraints were applied:

$$p_1 = (1,1)$$
$$p_L = (T,T)$$
$$p_{l+1} - p_l \in \begin{cases} \{(1,1)\} & \text{if } i_l < init \\ \{(1,0),(1,1)\} & \text{if } i_l \in \text{lead/lag silences} \\ \{(1,1)\} & \text{if } i_l \in \text{high-CSE} \\ \{(1,1),(1,2),(1,3)\} & \text{otherwise} \end{cases}$$

where $P = \{p_1, p_2, ..., p_L\}$ defines an alignment between S and M for pairs of time frames $p_l = (i_l, j_l)$, for $l \in [1 : L-1]$. The $init$ constant (set to 20 frames) prevents the speech from being retimed to coincide with the initial 200 ms of the masker where it might be perceptually more difficult to separate [20]. The $(1,0)$ path option in the lead/lag portions permits compression of the silent intervals in these regions. The restriction to a pure diagonal transition $(1,1)$ in the high-CSE regions ensures that speech rate is left unchanged in these portions of the signal, to avoid disrupting potentially salient information. Finally, the asymmetric path function in the remaining regions of the signal permits speech expansion and prohibits compression. High-CSE regions were detected with $\beta = 0.6$ (eqn. 6), and a boosting factor of $w = 3$ was used to balance the contribution of these regions with that made by the glimpse proportion components. These values were chosen empirically by observing their effect on an objective intelligibility estimate (section 4).

Figure 1 illustrates the speech realignment produced as a result of the above optimisation procedure for the two local cost functions defined in eqn. 4, while Figure 4 shows the effect of the modification in terms of gained and lost glimpses. Only the modification using CSE-weighting was submitted to the Hurricane Challenge, under the name **GCRetime**.

4. Results

The Hurricane Challenge provides speech and noise mixtures with a mean duration across the 180 mixtures of 3.04 s ($SD = 0.22$ s). The mean time scale factor resulting from the above optimisation procedure was 1.37 ($SD = 0.04$). Elongation factors were statistically identical across SNR conditions.

Figure 2 reports changes in glimpse proportion as a result of the proposed modification for both GP-only (eqn. 4a) and CSE-weighted GP (eqn. 4b). Modified speech was normalised to meet the original SNRs used for the Hurricane Challenge ([21], eqn. 1). Note that since the proposed method was not designed for the stationary masker of the Hurricane Challenge (speech-shaped noise, SSN), *speech modified in the presence of the competing speech masker (CS) was also used in the SSN condition*. This can be seen as a reference condition which measures the effect of temporal modifications on stationary maskers.

For CS, glimpse proportion increased significantly from plain to modified speech for both GP and CSE-GP in the snrMid and snrLo conditions of the Challenge [all $p<.01$]. Similarly, both modifications resulted in a significant decrease in the SSN condition [all $p<.05$]. An ANOVA on a linear model with MODIFICATION METHOD (i.e., GP vs. CSE-GP), SNR and MASKER as predictors confirmed the significant main effects of SNR and masker [both $p<.001$], but no effect of modification method [$p=.35$]. The observed decrease in SSN is due to the expansion of low-GP regions which are those most affected by the optimisation procedure, whose goal is to shift lower-energy speech parts to regions not dominated by the masker. As a consequence of expansion, for the stationary masker the proportion of low-energy regions *increases* with the resulting decrease in overall GP.

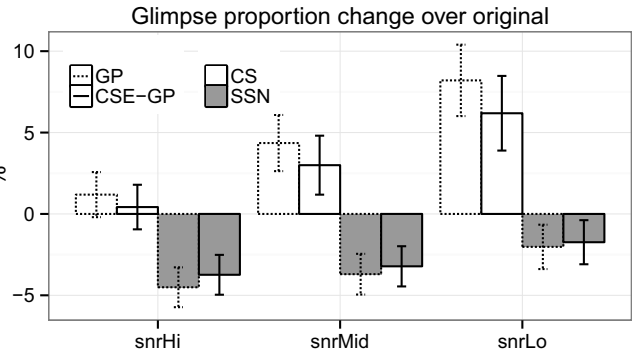

Figure 2: *GP changes in modified speech relative to non-modified speech, for the three SNR conditions. Error bars, here and elsewhere, show 95% confidence intervals ($N = 180$).*

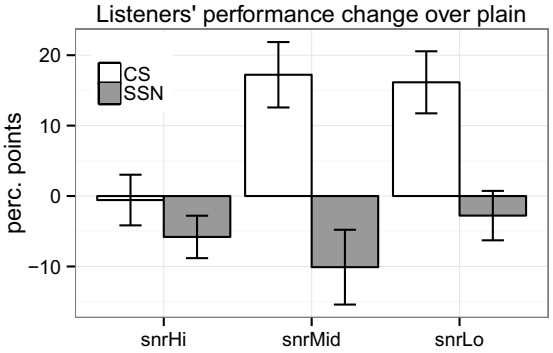

Figure 3: *Listeners' keyword scores for the modified speech relative to plain speech ($N = 175$)*

Figure 3 shows listeners' performance [6] in the CSE-GP modified speech condition expressed as changes in keyword scores relative to plain in percentage points. Highly-significant gains of up to 18 percentage points over the unmodified baseline are obtained in the snrMid and snrLo conditions for the CS masker [both $p<.001$], while smaller decreases in scores are observed for SSN in the snrHi and snrMid conditions [both $p<.001$]. These gains are equivalent to boosting the level of the unmodified speech by up to 4.4 dB [6].

5. Discussion

Temporal warping of speech to shift potentially-informative parts to regions where they suffer less energetic masking produced very substantial intelligibility gains in the Hurricane Challenge. Since the temporally-modified speech did not yield

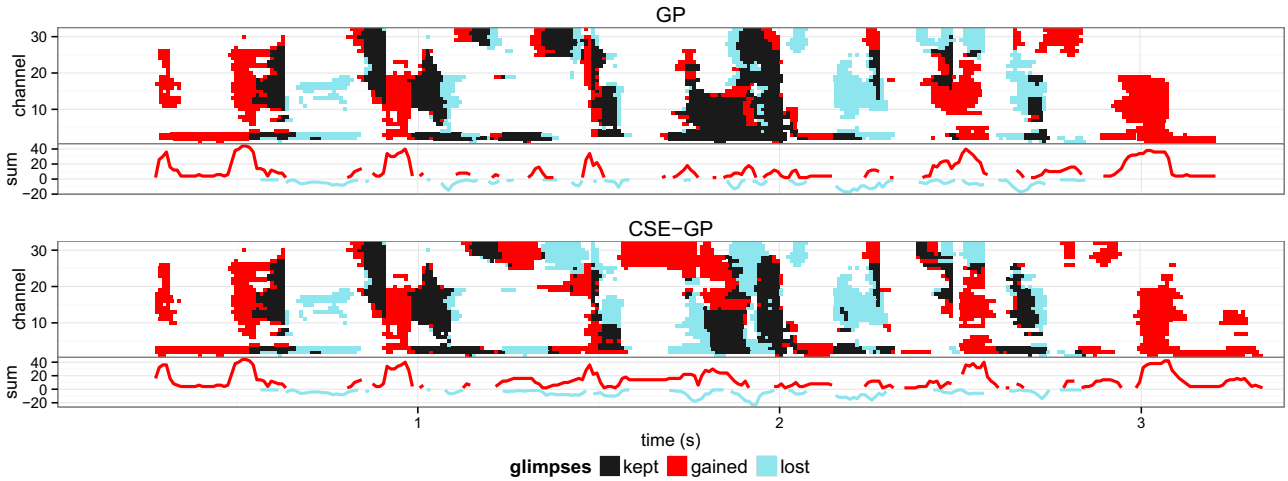

Figure 4: *Glimpse analysis of the modified speech mixture in comparison to the original mixture for GP-only (upper) and CSE-GP (lower). Black regions are glimpses which are present in the original mixture and are retained by the modification. Red regions [color online] indicate the glimpses created as a result of retiming, while blue regions are glimpses which were in the original mixture but became masked as a result of retiming. The lower part of each panel quantifies glimpse gains and losses.*

intelligibility benefits when mixed with the stationary masker, it appears that the gains observed with the fluctuating masker do not come from mere durational increases. In fact, it rather seems that temporal distortions that are independent of the masker can hinder sentence decoding to a small extent.

At present, the method assumes a known but unmodifiable masker signal. However, in some application scenarios which lead to temporally-overlapped speech signals whose source is remote from the listener (e.g., multiparty teleconferencing, air traffic control), it may be the case that both 'target' and masker speech are available for temporal modification. Here, the optimisation task could be to maximise the intelligibility of both signals. Non-interactive situations in which the delivered speech does not call for an answer or feedback allows for even more flexibility: consider the common situation of a tv or radio broadcast where the non-native speech of an interviewee is simultaneously presented with its spoken translation. However, one downside of improved temporal alignment of speech and masker is the potential for increased informational masking ([22]; see also [23] for a review of the effect of listening under adverse conditions). Further studies of the cognitive load associated with these different kinds of temporal modifications are needed.

Even without the ability to modify the masker, the current framework allows for a further method which might improve intelligibility, namely by using the target speech to mask the masker itself. Under the GP-only cost function, regions of the target speech with 'excess' energy could be retimed to disrupt energetic parts of the masker. In the case where the masker is competing speech, this approach could be further refined using metrics such as CSE to identify – and mask – the most informative regions of the *masker*.

An obvious limitation of the current approach is the need for a known masker signal. However, it may be possible to relax this constraint in several ways. First, although the current algorithm is based on moderate resolution spectro-temporal representations of speech and masker, it may be possible to use solely time-domain measures or estimates of energetic masking. Second, the temporal resolution itself (currently 10 ms) could be made coarser, enabling the approach to be applied in situations where the masker's temporal modulations are distorted (e.g., by reverberation).

More generally, the time realignment framework could be modified to use statistical models of the noise in place of the masker signal. This latter approach could be achieved by adaptations to the HMM model combination technique [24, 25]. Indeed, for applications using text-to-speech, HMM model combination could be integrated with HMM synthesis [26]. In application domains such as public transport interchanges, noise models for known time-varying sources such as trains or alarm sounds could be employed. Here, the goal would be to estimate the phase of the known noise source in order to permit effective retiming of generated speech.

Finally, we note that temporal realignment is largely orthogonal to spectral modification. The latter has received most attention in intelligibility enhancement and has proved capable of sizeable gains [6]. Combining the GCRetime method with, for example, boosting of mid-frequency energy, may well yield further benefits for listeners.

6. Conclusions

A time-warping framework that aims to ameliorate the effect of a fluctuating masker on salient information in a target speech signal is shown to be effective on the Hurricane Challenge, producing substantial gains in keyword scores over unmodified speech in mid and low-SNR conditions, increases equivalent to increasing the SNR of the unmodified speech by up to 4.4 dB.

Acknowledgements. The research leading to these results was partly funded from the European Community's 7th Framework Programme (FP7/2007-2013) under grant agreement 256230 (LISTA).

7. References

[1] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *J. Acoust. Soc. Am.*, vol. 84, no. 3, pp. 917–928, 1988.

[2] J. C. 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.

[3] M. Cooke and Y. Lu, "Spectral and temporal changes to speech produced in the presence of energetic and informational maskers," *J. Acoust. Soc. Am.*, vol. 128, no. 4, pp. 2059–2069, 2010.

[4] V. Aubanel, M. Cooke, E. Foster, M. L. García Lecumberri, and C. Mayo, "Effects of the availability of visual information and presence of competing conversations on speech production," in *Proc. Interspeech*, Portland, US, 2012.

[5] V. Aubanel and M. Cooke, "Strategies adopted by talkers faced with fluctuating and competing speech maskers," *J. Acoust. Soc. Am.*, under review.

[6] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane Challenge," submitted to *Interspeech*, Lyon, France, 2013.

[7] R. A. Cole, Y. Yan, B. Mak, M. Fanty, and T. Bailey, "The contribution of consonants versus vowels to word recognition in fluent speech," *Proc. ICASSP*, pp. 853–856, 1996.

[8] D. Kewley-Port, T. Z. Burkle, and J. H. Lee, "Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 122, no. 4, pp. 2365–2375, 2007.

[9] D. Fogerty and D. Kewley-Port, "Perceptual contributions of the consonant-vowel boundary to sentence intelligibility," *J. Acoust. Soc. Am.*, vol. 126, no. 2, pp. 847–857, 2009.

[10] D. Fogerty and L. E. Humes, "The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences," *J. Acoust. Soc. Am.*, vol. 131, no. 2, pp. 1490–1501, 2012.

[11] A. Wingfield, L. Lombardi, and S. Sokol, "Prosodic features and the intelligibility of accelerated speech: Syntactic versus periodic segmentation," *J. Speech Hear. Res.*, vol. 27, pp. 128–134, 1984.

[12] A. Waibel, "Prosodic knowledge sources for word hypothesization in a continuous speech recognition system," in *Proc. ICASSP*, 1987, pp. 856–859.

[13] S. Rosen, "Temporal information in speech: Acoustic, auditory, and linguistic aspects," *Phil. Trans. R. Soc. Lond. B*, vol. 336, no. 1278, pp. 367–373, 1992.

[14] C. Stilp and K. Kluender, "Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility," *PNAS*, vol. 107, no. 27, pp. 12 387–12 392, 2010.

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

[16] R. D. Patterson, J. Holdsworth, I. Nimmo-Smith, and P. Rice, "SVOS Final Report: The Auditory Filterbank," Tech. Rep. 2341, 1988.

[17] M. Cooke, Modelling Auditory Processing and Organisation. Cambridge: Cambridge University Press, 1993.

[18] B. C. J. Moore, B. R. Glasberg, C. J. Plack, and A. K. Biswas, "The shape of the ear's temporal window," *J. Acoust. Soc. Am.*, vol. 83, pp. 1102–1116, 1988.

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

[20] T. Cervera and W. A. Ainsworth, "Effects of preceding noise on the perception of voiced plosives," *Acta Acustica*, vol. 91, pp. 132–144, 2005.

[21] 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, 2013.

[22] D. S. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," *J. Acoust. Soc. Am.*, vol. 109, no. 3, pp. 1101–1109, 2001.

[23] S. Mattys, M. H. Davis, A. R. Bradlow, and S. K. Scott, "Speech recognition in adverse conditions: A review," *Lang. Cognitive Proc.*, vol. 27, no. 7-8, pp. 953–978, Sep. 2012.

[24] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. ICASSP*, 1990, pp. 845–848.

[25] M. Gales and S. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. on Speech and Audio Processing*, vol. 4, pp. 352–359, 1996.

[26] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Comm.*, vol. 51, no. 11, pp. 1039–1064, 2009.