

# Learning static spectral weightings for speech intelligibility enhancement in noise

Yan Tang<sup>a,\*</sup>, Martin Cooke<sup>b,c</sup>

<sup>a</sup>*Acoustics Research Centre, University of Salford, UK*

<sup>b</sup>*Ikerbasque (Basque Science Foundation), Bilbao, Spain*

<sup>c</sup>*Language and Speech Laboratory, Universidad del País Vasco, Vitoria, Spain*

---

## Abstract

Near-end speech enhancement works by modifying speech prior to presentation in a noisy environment, typically operating under a constraint of limited or no increase in speech level. One issue is the extent to which near-end enhancement techniques require detailed estimates of the masking environment to function effectively. The current study investigated speech modification strategies based on reallocating energy statically across the spectrum using masker-specific spectral weightings. Weighting patterns were learned offline by maximising a glimpse-based objective intelligibility metric. Keyword scores in sentences in the presence of stationary and fluctuating maskers increased, in some cases by very substantial amounts, following the application of masker- and SNR-specific spectral weighting. A second experiment using generic masker-independent spectral weightings that boosted all frequencies above 1 kHz also led to significant gains in most conditions. These findings indicate that energy-neutral spectral weighting is a highly-effective near-end speech enhancement approach that places minimal demands on detailed masker estimation.

*Keywords:* speech, intelligibility, glimpsing, noise, pattern search, spectral weighting

---

---

\*Corresponding author at: Acoustics Research Centre, University of Salford, UK  
*Email address:* [y.tang@salford.ac.uk](mailto:y.tang@salford.ac.uk) (Yan Tang)

## 1. Introduction

Listening to speech in noisy or reverberant environments is both error-prone and effortful. Consequently, reducing the impact of noise via speech enhancement has been the goal of a significant research effort (e.g. Hu and Loizou, 2004; Paliwal and Alsteris, 2005; Martin, 2005; Chen et al., 2006; Srinivasan et al., 2007; Kim et al., 2009; Williamson et al., 2015). Techniques such as noise cancellation or suppression are widely used in human-machine interfaces, and in technologies such as mobile communication and noise-cancelling headphones. However, these approaches have limited use in applications such as public-address systems where listeners are not directly adjacent to the end-point of the transmission channel since, even when the speech signal is further enhanced at the listener’s end, the ensuing signal may suffer further contamination in a noisy listening environment.

An alternative approach is to manipulate the speech signal itself, analogous to the way human talkers adjust their speaking style in noisy conditions (e.g. Lombard, 1911; Summers et al., 1988; Junqua et al., 1998; Boril and Pollak, 2005; Cooke and Lu, 2010). Many approaches have been proposed in the last decade to increase speech intelligibility under adverse conditions by altering the clean speech signal. The concept of *near-end* listening enhancement, introduced by Sauert and Vary (2006), describes situations where the speech signal originating at the end of the transmission channel distant from the listener is modified to increase speech intelligibility for the near-end listener who is assumed to be located in a noisy environment. Techniques are generally based on raising the speech spectrum above the average noise spectrum using spectro-temporal manipulation of local signal-to-noise ratio (SNR). Bonardo and Zovato (2007) introduced a dynamic range controller to increase perceived loudness of synthetic speech while maintaining the original intensity range. Time-frequency dependent amplification was employed by Brouckxon et al. (2008) in formant-enhancement, leading to a decreased speech reception threshold in noise.

The aforementioned studies show that increasing SNR via amplification provides a clear benefit for listeners. However, the use of excessive output levels may lead to listener discomfort and stress, and sustained exposure can cause damage to hearing (Knobel and Sanche, 2006) or equipment (Sabin and Schoenike, 1998). Methods proposed in more recent studies (e.g. Yoo et al., 2007; Sauert and Vary, 2009; Tang and Cooke, 2010, 2012; Taal et al., 2014; Schepker et al., 2015) operate under a constant input-output regime

1 for the speech signal, precluding any intelligibility gains due simply to an  
2 increase in overall SNR. Even under these constraints speech modification  
3 can be highly-effective. Extensive across-algorithm comparisons involving 26  
4 speech modification techniques and using the same dataset for evaluation  
5 (Cooke et al., 2013a,b) have shown that state-of-the-art approaches are able  
6 to boost intelligibility by an amount equivalent to increasing the gain of  
7 unmodified speech by more than 5 dB.

8 Objective intelligibility or quality metrics (OIMs) have been used in the  
9 design of near-end speech modification techniques based on optimising model  
10 parameters by maximising the objective metric. Sauert and Vary (2009) op-  
11 timised the Speech Intelligibility Index (ANSI S3.5-1997, 1997), while the  
12 algorithm proposed by Taal et al. (2014) transferred energy to consonant-  
13 vowel transients by optimising a perceptual distortion measure developed by  
14 Taal and Heusdens (2009), leading to significant listener gains. In our pre-  
15 vious work (Tang and Cooke, 2012), the glimpse proportion metric (Cooke,  
16 2006) was used as the OIM in closed-loop optimisation process to derive a  
17 series of masker- and level-dependent spectral weightings. Akin to band-  
18 importance functions (Studebaker et al., 1987; Stubebaker and Sherbecoe,  
19 1991; Bell et al., 1992) which quantify the contribution of each frequency re-  
20 gion to overall intelligibility, spectral weightings inject more energy in certain  
21 frequency bands at the expense of others, although unlike band-importance  
22 functions the weightings depend on the masker. Speech with optimised spec-  
23 tral weights was more intelligible than unmodified speech for both stationary  
24 and fluctuating maskers (Cooke et al., 2013b).

25 The current study extends Tang and Cooke (2012) in three directions.  
26 First, the optimisation process makes use of a new glimpse-based OIM re-  
27 cently shown to outperform the original glimpse proportion measure. The  
28 success of an optimisation strategy is limited by the accuracy of the chosen  
29 OIM. In a recent comparison (Tang and Cooke, 2016) of glimpse-based op-  
30 timisation approaches alongside a state-of-the-art OIM (Christiansen et al.,  
31 2010), a metric based on high-energy glimpses led to the most accurate pre-  
32 dictions of listener intelligibility scores across nearly 400 conditions varying  
33 in speech style, masker type and SNR. The high-energy glimpsing metric,  
34 described in section 2, forms the basis for the optimisation approach of the  
35 current study.

36 Second, the effect on intelligibility of both masker-dependent and masker-  
37 independent spectral weightings is evaluated, by questioning the assumption  
38 behind many of the aforementioned modification approaches (e.g. Sauert and

1 Vary, 2009; Tang and Cooke, 2010; Taal et al., 2014) that the background  
2 noise signal is known or capable of being accurately-estimated. In prac-  
3 tice, noise estimation can be problematic, particularly at short time delays.  
4 Consequently, algorithms have been proposed that operate independently  
5 of knowledge of the masker. Such algorithms typically boost those speech  
6 regions or properties believed to convey salient speech information. For ex-  
7 ample, Zorila et al. (2012) demonstrated that subjective intelligibility can  
8 benefit from enhancing formant information and emphasising voicing seg-  
9 ments while preserving high frequency components, with a further intelli-  
10 gibility boost produced by dynamic range compression. In another study,  
11 Jokinen et al. (2016) showed that modifying the phase spectrum of wide-  
12 band telephony speech by enhancing high-amplitude peaks caused by the  
13 glottal excitation in the time domain can also increase speech intelligibility  
14 in noise. Consequently, one of the objectives of the current study was to  
15 determine the effectiveness of spectral weightings learnt offline (Expt. 1) or  
16 based on a generic masker-independent boosting pattern (Expt. 2).

17 Finally, in this study we employ a numerical optimisation approach that  
18 is capable of operating with the high-dimensionality parameter vectors that  
19 result from an auditory-based spectral representation. Although it is possible  
20 to optimise spectral weightings using a low dimensionality representation  
21 such as octave-bands (e.g. Viktorovitch, 2005), it is desirable to make use  
22 of a more realistic finer-scale spectral representation that is known to reflect  
23 auditory frequency resolution. Our earlier approach (Tang and Cooke, 2012)  
24 used genetic algorithms (Holland, 1975; Mitchell, 1996) for this purpose.  
25 In the current study we use a different numerical optimisation technique,  
26 pattern search (Hooke and Jeeves, 1961; Davidon, 1991), which was designed  
27 to be deployed in complex, high-dimensional and potentially-discontinuous  
28 search spaces.

29 Section 2 motivates the high-energy glimpse pro portion metric used at  
30 the core of the optimisation process to predict intelligibility. Spectral weight-  
31 ings which result from pattern search optimisation in the presence of different  
32 maskers are derived in section 3. The following section presents the outcome  
33 of an experiment in which listeners identified keywords in sentences modified  
1 by optimised spectral weights in matched masker/level conditions. Based  
2 on common features of the spectral weightings discovered via optimisation,  
3 section 5 describes the results of a second intelligibility experiment using a  
4 number of generic, masker-independent spectral weightings.

5 **2. High energy glimpse proportion**

6 Glimpse proportion (GP) quantifies the proportion of time-frequency re-  
 7 gions of an auditory-inspired representation of the speech signal that exceed  
 8 equivalent regions of the masker by a specific amount. GP is intended to re-  
 9 flect the local audibility of speech in noise, and is correlated with subjective  
 10 intelligibility data (e.g. Barker and Cooke, 2007). Tang and Cooke (2012)  
 11 demonstrated that modifying speech to maximise GP can lead to intelli-  
 12 gibility gains. GP is defined in terms of spectro-temporal excitation patterns  
 13 (Moore, 2003)  $S_f(t)$  and  $N_f(t)$  for speech and noise at time  $t$  in frequency  
 14 region  $f$  as follows:

$$GP = \frac{1}{TF} \sum_{f=1}^F \sum_{t=1}^T \mathcal{H}[S_f(t) > N_f(t) + \alpha] \quad (1)$$

15 where  $F$  denotes the number of frequency bands,  $T$  the number of time  
 16 frames, and  $\mathcal{H}[\cdot]$  is the unit step function which counts the number of spectro-  
 17 temporal regions meeting the local masked audibility criterion  $\alpha$ . GP is a  
 18 normalised measure in the range 0-1.

19 The high-energy glimpse proportion metric (HEGP; Tang and Cooke,  
 20 2016) was inspired by an approach taken in the Coherence Speech Intelligi-  
 21 bility Index (Kates and Arehart, 2005) of separately weighting frames based  
 22 on a tripartite categorisation of the RMS energy of the speech signal in each  
 23 frame. The OIM introduced by Christiansen et al. (2010) used a similar  
 24 notion and employed only the high-energy frames. Rather than classifying  
 25 at the frame level, Tang and Cooke (2016) categorised glimpses based on  
 26 their energy relative to the mean in each frequency region. High-energy  
 27 glimpses are defined as those time-frequency regions deemed to be glimpses  
 28 by eq. 1 with the additional requirement that the local excitation pattern  
 29 for the speech-plus-noise mixture,  $Y_f(t)$ , is greater than the average level in  
 30 frequency region  $f$ . In HEGP the glimpsing criterion in eq. 1 is replaced by:

$$[S_f(t) > (N_f(t) + \alpha)] \wedge (Y_f(t) > \bar{Y}_f) \quad (2)$$

31 where  $\bar{Y}_f$  represents the mean of  $Y_f$  across time.

32 Tang et al. (2016) reported further significant improvements in the pre-  
 33 dictive power of the HEGP metric by removing inaudible (sub-threshold)  
 1 glimpses, and by applying a quasi-logarithmic transformation to the GP  
 2 value, based on the finding that subjective intelligibility scores reach ceiling

3 for relatively low values of GP (Barker and Cooke, 2007). These extensions  
4 increased listener-model correlations from 0.79, 0.71 and 0.53 for the original  
5 GP metric to 0.92, 0.83 and 0.87 across three large-scale datasets.

6 In the version of HEGP used in the current study, excitation patterns  
7 were produced by a bank of  $F = 34$  gammatone filters uniformly distributed  
8 on the equivalent rectangular bandwidth scale, covering the frequency range  
9 100-7500 Hz, sampled in time at a frame rate of 100 Hz. The local SNR  
10 threshold  $\alpha$  was set to 3 dB based on the findings in Cooke (2006).

### 11 **3. Optimised spectral weightings**

#### 12 *3.1. Maximising HEGP via Pattern Search optimisation*

13 In the current study, Pattern Search (PS; Hooke and Jeeves, 1961; Davi-  
14 don, 1991) was used alongside HEGP to estimate spectral weightings. PS is  
15 a member of the direct search family of numerical optimisation methods that  
16 do not require estimates of the gradient or higher derivatives of the objective  
17 function. These methods are suitable for optimisation in a high-dimensional  
18 space (Yu, 1979). In the current context, PS operates by exploring the space  
19 of spectral weight vectors. Each component of the vector is a value in deci-  
20 bels representing a boost or an attenuation in the corresponding frequency  
21 band. At each iteration, the candidate vector is normalised to have zero  
22 mean and the average HEGP metric evaluated across a development set of  
23 sentences in the presence of a given masker at a specified SNR. The final  
24 spectral weighting results when a convergence criterion is reached, or after a  
25 specified maximum number of iterations.

26 It is important to note that the zero mean normalisation step only ap-  
27 proximates the effect of a constant input-output RMS level for the purposes  
28 of HEGP computation in PS optimisation. In practice, the actual RMS level  
29 resulting from the spectral weighting at any step of the optimisation pro-  
30 cess will be different for each sentence in the development set. Crucially, for  
31 the listening experiments reported below, normalisation was performed on a  
32 sentence-by-sentence basis to ensure that the RMS level following boosting  
33 was exactly the same as the level prior to boosting.

34 Spectral weight vectors consisted of 34 components corresponding to the  
35 number of gammatone filters ( $F$ ) used to compute the HEGP metric. Since  
36 PS is a minimisation procedure, the negative of HEGP was used as the  
1 cost function. The development set over which HEGP was evaluated at  
2 each iteration contained 100 sentences (see section 3.2). At each iteration

3 individual spectral weights were constrained to the range  $[-50, 50]$  dB to  
4 prevent excessive boosting or attenuation in specific frequency regions. An  
5 implementation of PS from the MATLAB Global Optimisation Toolbox was  
6 used, with an initial mesh size of 1, and mesh expansion and contraction  
7 factors of 2.0 and 0.5 respectively. An iteration limit of 200 was imposed but  
8 in practice PS converged after only 25-30 iterations.

### 9 *3.2. Speech material*

10 Speech material was drawn from the Sharvard corpus (Aubanel et al.,  
11 2014), a phonemically-balanced Spanish sentence resource inspired by the  
12 original English Harvard sentences (Rothausser et al., 1969). The corpus  
13 contains 700 sentences uttered by both a male and a female native Span-  
14 ish talker. Each sentence contains five keywords for scoring e.g. ‘el grupo  
15 de gente se sumó a la fuerte lucha’. Sentences 1-100, with a sampling fre-  
16 quency of 16 kHz, from the male talker were used to learn optimised spectral  
17 weightings.

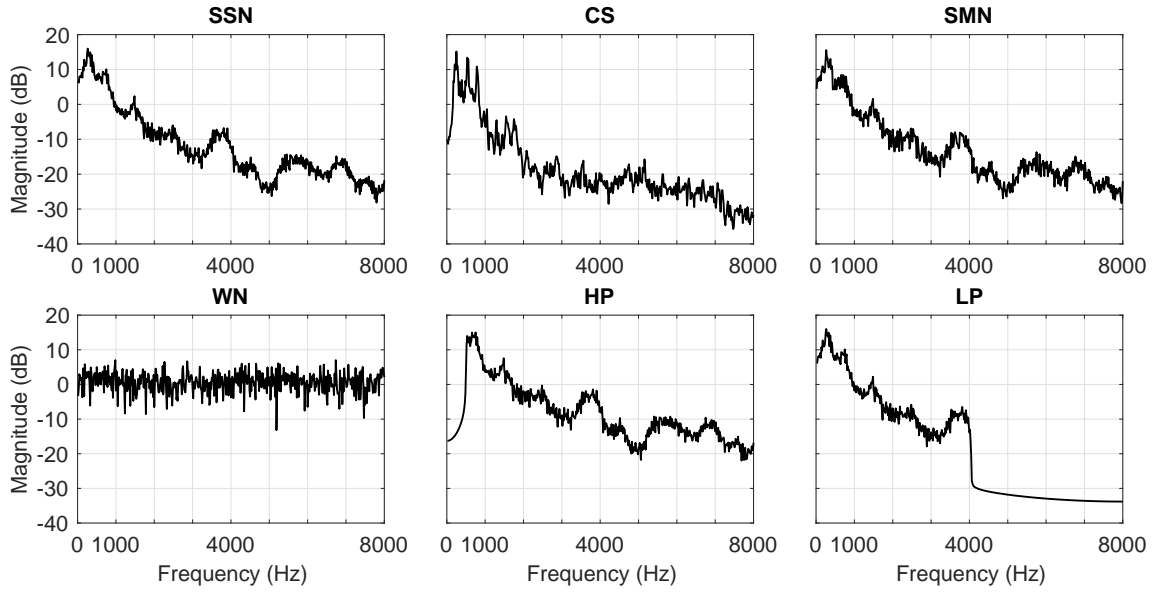
### 18 *3.3. Maskers and SNRs*

19 Six maskers, depicted in Fig. 1, were used in the optimisation procedure  
20 and in subsequent perceptual listening experiments. The SSN masker was  
21 constructed to have a long term spectrum matching that of the male Shar-  
22 vard speaker. The CS masking material came from the female Sharvard  
23 talker. The SMN masker was generated by multiplying the SSN signal by  
24 the envelope of randomly-concatenated CS sentences. Low-pass and high-  
25 pass noise maskers LP and HP were derived by filtering the SSN signal at  
26 cutoff frequencies of 4000 and 500 Hz respectively using IIR Chebyshev fil-  
27 ters with 0 dB passband gain and 80 dB stop-band attenuation. Long term  
28 average spectra of the six maskers are shown in Fig. 1.

29 For each masker, optimisation was performed at two SNR levels, denoted  
30 ‘low SNR’ and ‘high SNR’, whose values are provided in Fig. 1. SNRs were  
31 chosen in pilot tests to result in approximately 25% and 50% keywords correct  
32 scores.

### 33 *3.4. Spectral weighting candidates*

34 In order to examine the consistency of the spectral weight patterns learnt  
35 in individual optimisation runs, the output of two trials of the optimisa-  
36 tion process were inspected for each condition. While the resulting spectral



masker		low SNR	high SNR
SSN	speech-shaped noise	-8	-6
CS	competing speech	-21	-17
SMN	speech-modulated noise	-13	-9.5
WN	white noise	-10	-7
LP	low-pass, cutoff 4000 Hz	-23	-18.5
HP	high-pass, cutoff 500 Hz	-10.5	-8.5

Figure 1: *Details of the maskers used in the study.*



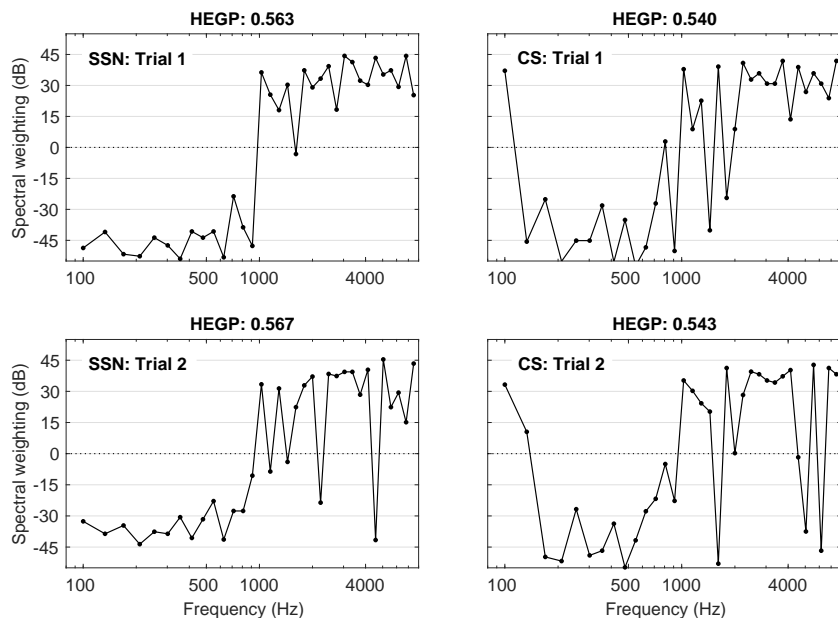


Figure 2: *Best spectral weightings from two separate PS optimisation runs for SSN (left) and CS (right) maskers in the high SNR condition. The HEGP score for each weighting is also displayed.*

37 weights differed in some details, in general the overall patterns were very  
 1 similar, as illustrated in Fig. 2.

2 The spectral weightings for the six maskers at both SNR levels are pre-  
 3 sented in Fig. 3. For each condition, the final weightings were computed as  
 4 the average over the candidates from the two trials. The three maskers whose  
 5 long-term spectrum is that of broadband speech (CS, SSN and SMN) as well  
 6 as the LP masker display a similar spectral weighting pattern with a clear  
 7 boost for frequencies above 1 kHz. The two fluctuating maskers CS and  
 8 SMN additionally exhibit a tendency to boost very low frequencies. The  
 9 WN masker displays a converse pattern, with a clear boosting of mid-to-low  
 10 frequencies. The HP masker leads to spectral boosting applied in the region  
 11 below 500 Hz and above 2000 Hz. To a first-order, profiles are similar at  
 12 both low and high SNRs. However, differences in the low frequency region  
 13 are evident for the CS, SMN and LP maskers, and in the mid-high region  
 14 for the WN and HP maskers. An unexpected feature of the spectral weight-  
 1 ings is the presence of wide-ranging fluctuations in the degree of boosting,  
 2 covering a range of some 60 dB, particularly in the high frequency region.

3 We speculate on possible origins for these features in section 4.6. The next  
 4 section presents the results of a listening experiment which measured the  
 5 effect on intelligibility of spectral weighting.

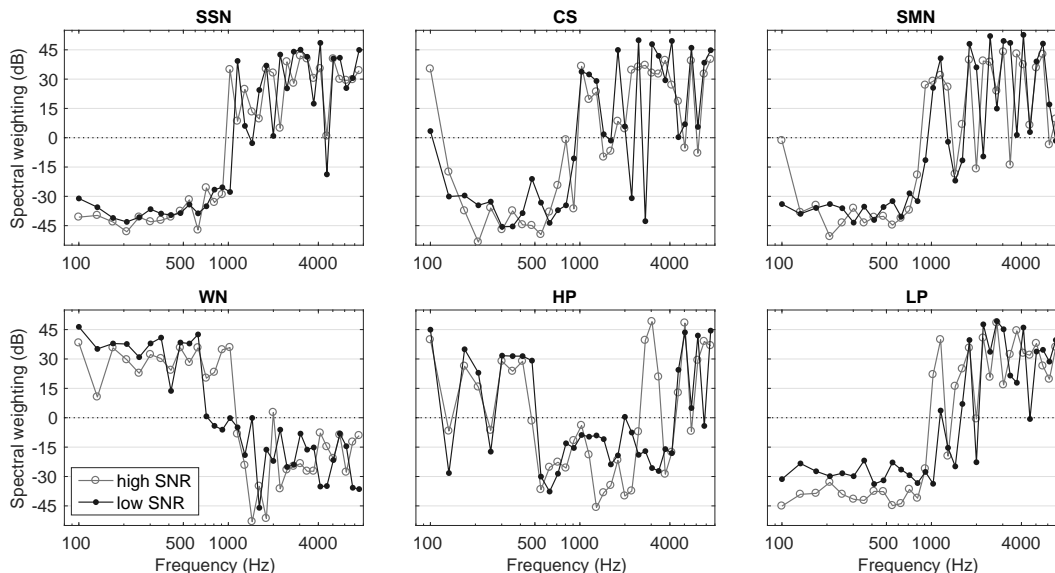


Figure 3: *Optimised spectral weightings discovered by Pattern Search.*

#### 6 **4. Experiment 1: subjective intelligibility of sentences boosted by** 7 **masker- and SNR-dependent optimised spectral weightings**

##### 8 *4.1. Speech and masker material*

9 A set of 240 sentences drawn from the male talker of the Sharvard cor-  
 10 pus (sentences 401-640) was used in Expt. 1. These sentences are different  
 11 from the 100-member development set employed during pattern search op-  
 12 timisation. Maskers and SNRs were the same as those used in the learning  
 13 phase.

##### 14 *4.2. Listeners*

15 Some 22 native Spanish undergraduate and graduate students (age range  
 16 20-38 years, mean 23.8 years, std. dev. 4.8 years) from the University of  
 17 the Basque Country took part in Expt. 1. All participants were given an  
 18 audiometric hearing screening test at octave frequencies between 125 Hz and

19 8000 Hz. Results from two participants who had hearing thresholds above  
1 20 dB HL at two or more frequencies were excluded. All participants were  
2 paid for their participation.

### 3 *4.3. Procedure*

4 The complete set of 240 sentences in both unmodified form ('plain') and  
5 spectrally-weighted ('weighted') using the optimised weightings shown in  
6 Fig. 3 was mixed with each of the 6 maskers at the 2 SNR levels, lead-  
7 ing to 5760 stimuli (240 sentences  $\times$  6 maskers  $\times$  2 SNRs  $\times$  2 modifica-  
8 tions). Each listener was allocated a subset of 240 stimuli using a balanced  
9 design, such that each listener heard the same sentence only once, and each  
10 masker/SNR condition was heard by the same number of listeners. The  
11 240 sentences were blocked into 12 masker/SNR conditions, with 10 plain  
12 and 10 spectrally-weighted stimuli per block. Block presentation order was  
13 randomised for each participant, as was within-block stimulus presentation  
14 order.

15 Stimulus presentation and response collection made use of a custom-built  
16 MATLAB application. Stimuli were normalised to the same RMS level, and  
17 20 ms half-Hamming ramps applied to attenuate onset and offset transients.  
18 Presentation level was fixed at 80 dB(A), calibrated using a Brüel & Kjær  
19 4153 artificial ear with a Brüel & Kjær 2250-L sound pressure level anal-  
20 yser. Listeners heard stimuli via Sennheiser HD380 Pro headphones in a  
21 sound-attenuating booth in the Phonetics Laboratory at the University of  
22 the Basque Country (Vitoria-Gasteiz Campus). A practice session preceded  
23 the formal test in order to accustomise listeners to the testing environment.  
24 Listeners typed their responses into an on-screen text box.

### 25 *4.4. Post-processing*

26 Subjective intelligibility was computed as the correct keyword recognition  
27 rate for each condition. Five keywords per sentence were used for scoring.  
28 Due to inconsistent use of diacritics by listeners, all vowel accents were re-  
29 moved prior to scoring so that answers with or without accents were consid-  
30 ered to be equivalent e.g. both 'ríó' and 'rio' were considered to be correct  
31 responses for the word 'ríó'. For statistical purposes, percentages were con-  
32 verted to rationalised arcsine units (RAU; Studebaker, 1985). However, for  
33 ease of interpretation, results are shown in percentages in the figures.

34 4.5. Results

35 The boxplots in Fig. 4 indicate means and ranges of the percentage of  
 36 keywords identified correctly for plain and spectrally-weighted speech in the  
 1 two SNR and six masking conditions. Spectral weighting led to increases  
 2 in all conditions apart from those related to the white noise masker, with  
 3 improvements ranging from 8 to 55 percentage points. A similar pattern of  
 4 gains was observed at each SNR. Averaged across the two SNRs, the largest  
 5 gains amounted to 51 and 44 percentage points respectively. These occurred  
 6 for the two stationary maskers with a low-pass characteristic (SSN and LP).  
 7 Gains for the two fluctuating maskers (CS and SMN) were more modest,  
 8 with increases averaged across SNR levels of 17 and 26 percentage points

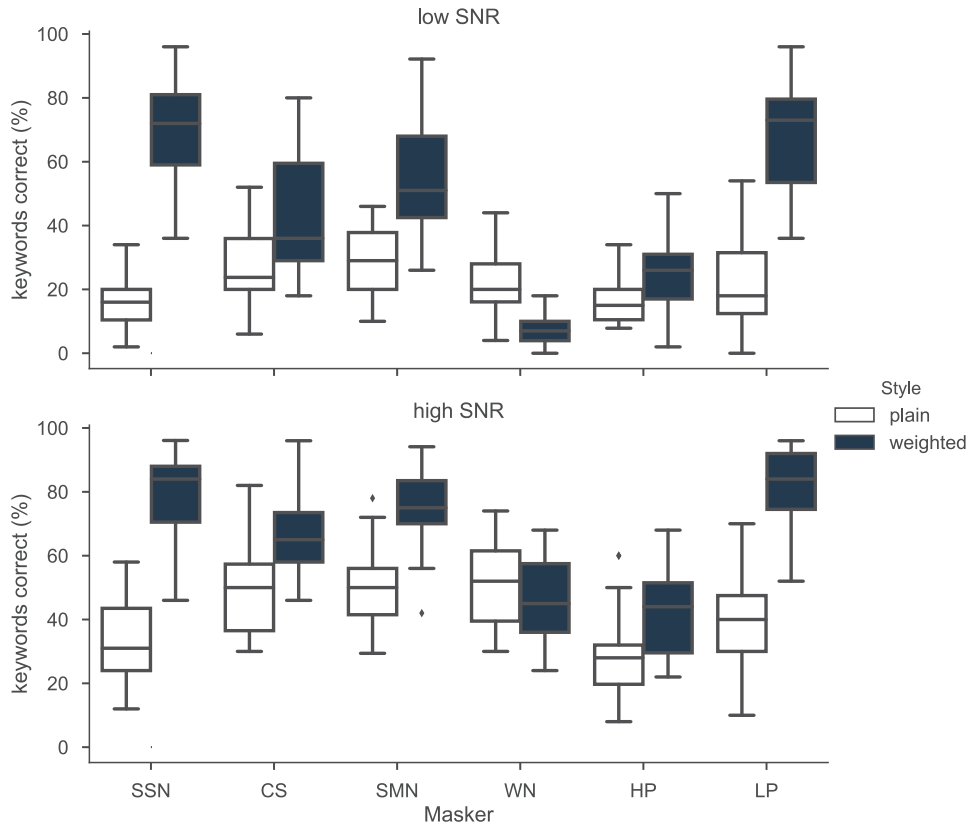


Figure 4: *Intelligibility scores for spectrally-weighted speech and unmodified ‘plain’ speech in the presence of masking noise at low (top) and high (bottom) SNRs.*

9 respectively. The HP condition led to a smaller increase of 10 percentage  
10 points. The exception to the pattern of intelligibility increases was white  
1 noise, with decreases of 15 and 6 percentage points at the low and high SNR  
2 levels (all figures have been rounded to the nearest integer).

3 A three-way repeated measures ANOVA with within-subjects factors of  
4 masker type, SNR level and modification style (plain, weighted) performed on  
5 RAUs confirmed the statistical significance of the reported outcomes. A sig-  
6 nificant main effect of modification style [ $F(1, 21) = 476.98, p < .001, \eta^2 =$   
7  $.41]$  was observed, as well as a two-way interaction between modification style  
8 and masker type [ $F(5, 105) = 117.5, p < .001, \eta^2 = .39]$ , and a three-way  
9 interaction among the three main factors [ $F(5, 105) = 5.99, p < .001, \eta^2 =$   
10  $.021]$  but no significant interaction between the modification style and SNR  
11 level [ $F(1, 21) = .03, p = .87, \eta^2 < .001]$ . Post-hoc analyses based on  
12 a Fisher's least significant difference of 5.1 RAUs confirmed that spectral  
13 weighting led to a significant intelligibility gain for all masker types apart  
14 from WN.

#### 15 *4.6. Discussion*

16 The outcome of experiment 1 provides a clear demonstration that the  
17 simple expedient of reallocating spectral energy by boosting some frequency  
18 bands at the expense of others is capable of increasing speech intelligibility  
19 substantially without increasing SNR. Further, spectral boosting profiles can  
20 be learnt by closed-loop optimisation with an objective intelligibility metric  
21 at its core.

22 While gains from static spectral boosting might be expected for stationary  
23 maskers, benefits were also evident for the fluctuating maskers CS and SMN.  
24 Both maskers led to similar scores in plain speech, but SMN-based weighting  
25 produced larger gains than CS-based weighting. Since both maskers have  
26 similar temporal properties – the temporal modulation pattern of SMN is  
27 derived from the temporal envelope of the CS masker – it is possible that  
28 the difference stems from additional informational masking in the case of the  
29 competing talker (Brungart et al., 2001). An alternative explanation for the  
30 smaller gains observing in the CS condition might be a loss of audibility: in  
31 order to achieve similar intelligibilities across maskers for the plain baseline,  
32 a rather low SNR of -21 dB was required in the CS condition. Since the  
33 presentation level was constant, the level of the speech target was reduced  
34 for the CS condition relative to the SMN masking condition.

35 The spectral boosting pattern learnt for the white noise masking condi-  
36 tions was not only ineffective, but actually harmed intelligibility. An inspec-  
37 tion of Fig. 3 shows that spectral energy was transferred from mid and high  
38 frequencies to the region below 1 kHz. While the HP pattern has a similar  
1 characteristic in the low frequencies, it is apparent that it is only in the WN  
2 case that the high frequencies are entirely attenuated. It may be that the  
3 reduction in salience of cues to the location of the higher formants and the  
4 presence of fricative energy resulted in an intelligibility loss in this condition.  
5 An alternative possibility is that the HEGP predictions are poor in the case  
6 of WN, leading to an inappropriate boosting pattern. There is some evi-  
7 dence to support this hypothesis: the least accurate predictions were seen in  
8 the WN and HP conditions, with a Pearson correlation coefficient of just  
9 0.43 for the comparison of HEGP and subjective scores, while the equivalent  
10 correlation for the remaining four maskers was 0.91. Further work is required  
11 to understand why the HEGP metric makes poorer predictions for white and  
12 high-pass maskers.

13 The largest gains came from the four maskers with a similar spectral  
14 boosting pattern (SSN, CS, SMN, LP). Indeed, these maskers have near-  
15 identical long-term average spectra up to 4 kHz (fig. 1). Given that on  
16 average a speech-shaped masker (by definition) masks speech equally in all  
17 frequency regions (Mayer, 1894; Wegel and Lane, 1924), it is surprising to  
18 observe that a strategy that essentially boosts all information above 1 kHz  
19 is so effective. One explanation might be that due to the cochlear tonotopic  
20 mapping there are proportionally more frequency channels devoted to the  
21 0-1 kHz region than any other 1 kHz wide band, and thus more potential  
22 glimpses available to be reallocated elsewhere. Another possibility is that  
23 the regions above 1 kHz are more important for speech perception, although  
24 the evidence for such a claim is mixed (see General Discussion). A related  
25 possibility is that voicing information available in the lower frequency region  
26 does not necessarily require the transmittance of all resolved harmonics, lead-  
27 ing to some redundancy of information. In a similar vein, one unexpected  
28 characteristic of the learnt spectral weighting patterns is what appears to be  
29 selective boosting of channels, mainly in the mid and high frequencies. A  
30 similar sparse boosting pattern was found in Tang and Cooke (2012) using  
31 a GP-based metric and genetic algorithm optimisation. It is tempting to  
32 conclude that under a constant input-output energy constraint, it is better  
33 to ensure that a range of frequencies is boosted rather than enhancing neigh-  
34 bouring frequency channels that contain redundant information, although

35 there is no basis for preferring non-neighbouring channels in the HEGP met-  
 36 ric.

37 Having observed that the weightings obtained by maximising HEGP in  
 38 the main emphasise mid-to-high frequencies (Fig. 3), one hypothesis is that  
 1 intelligibility benefits from an increase in the average across-frequency-band  
 2 SNR. To investigate this possibility, the optimisation procedure described  
 3 in Section 3.1 was used to determine weightings based on maximising the  
 4 average SNR across frequency bands (denoted ‘Max SNR’) in the low SNR  
 5 condition for each masker. We also looked at the effect of pre-emphasis,  
 6 since this also results in the transfer of energy from low to high frequencies  
 7 under the constant input-output energy constraint adopted in this study.  
 8 Fig. 5 displays the weightings for each masker that are suggested by the  
 9 optimisation. The frequency response of a pre-emphasis filter with  $\alpha = 0.97$   
 10 is also shown. It is clear that weightings based on maximising average band  
 11 SNR, along with those for pre-emphasis, are qualitatively different from those  
 12 based on HEGP (Fig. 3) with the latter showing a far steeper transition  
 13 between low and mid-high frequencies.

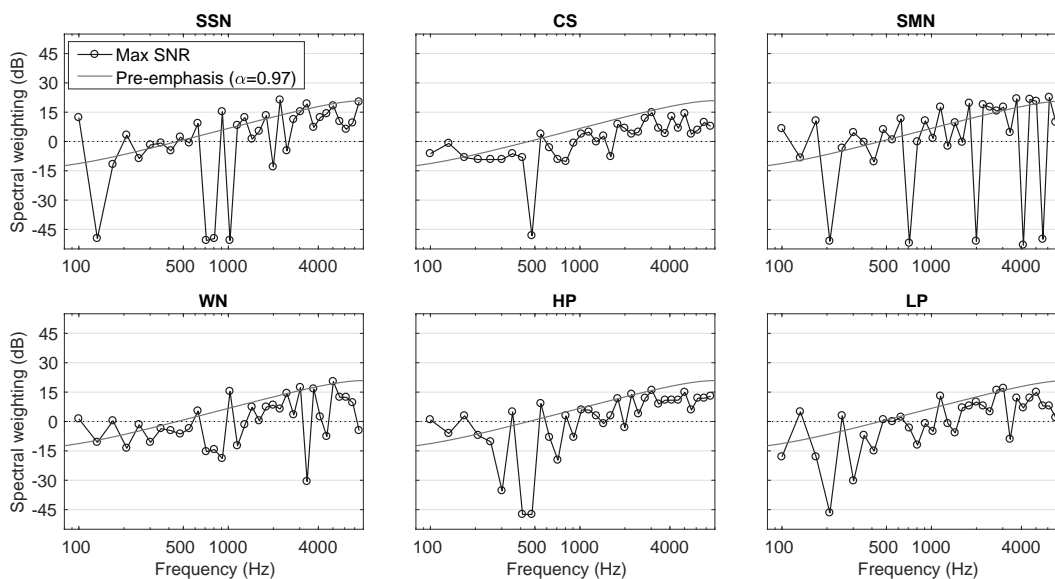


Figure 5: *Spectral weightings learnt by maximising average across-frequency-band SNR. The frequency response of a pre-emphasis filter ( $\alpha = 0.97$ ) is shown with a 15-dB offset.*

14 When comparing the average band SNR of speech modified by maximising

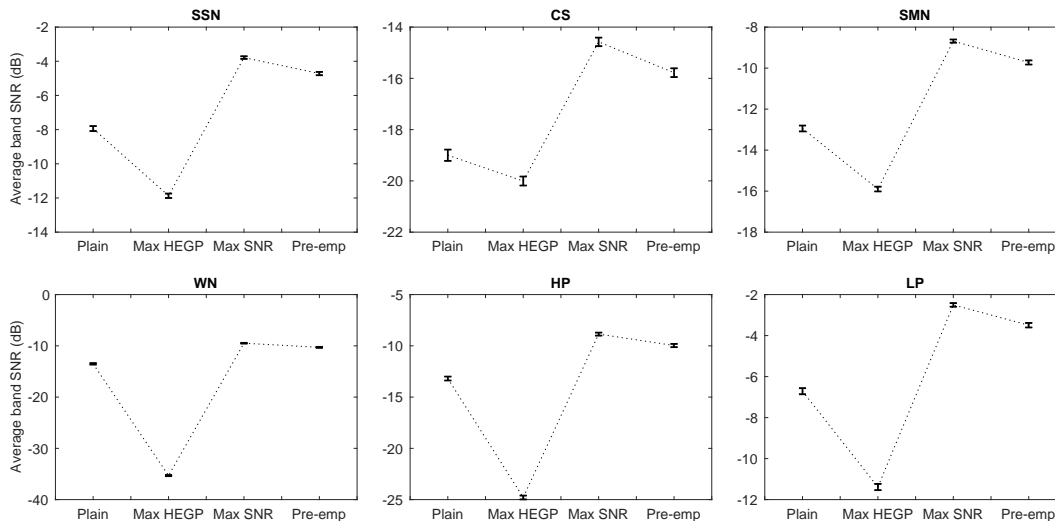


Figure 6: Average band SNR of unmodified speech and speech modified by the Max HEGP, Max SNR and pre-emphasis weighting. Error bars denote 95% confidence intervals.

15 HEGP, maximising average SNR, and by pre-emphasis, to that of unmodified  
 16 speech (Fig. 6), the highest average SNRs are indeed produced by the max  
 1 SNR weightings, with pre-emphasis also showing increases over the Plain  
 2 baseline. However, the approach proposed in this study, max HEGP, results  
 3 in a clear reduction in average SNR, demonstrating that intelligibility gains  
 4 from spectral weightings produced by maximising HEGP are not due to  
 5 increased average SNR. Enhancing speech intelligibility under a constant  
 6 input-output energy constraint does not necessarily require a better average  
 7 band SNR.

8 The observed similarities of the spectral weighting patterns for the three  
 9 wideband speech-based maskers (CS, SSN, SMN) raises the question of  
 10 whether weighting patterns might be generalised across these maskers, and  
 11 the extent to which masker- and SNR-dependent weightings are needed at  
 12 all. A second listening experiment was performed to address this issue.

## 13 5. Experiment 2: The effect of generic spectral weightings

14 The primary goal of Expt. 2 was to evaluate the effectiveness of a generic  
 15 (masker independent) spectral weighting pattern based on a schematic ver-  
 16 sion of the mid-high frequency energy reallocation pattern observed in the



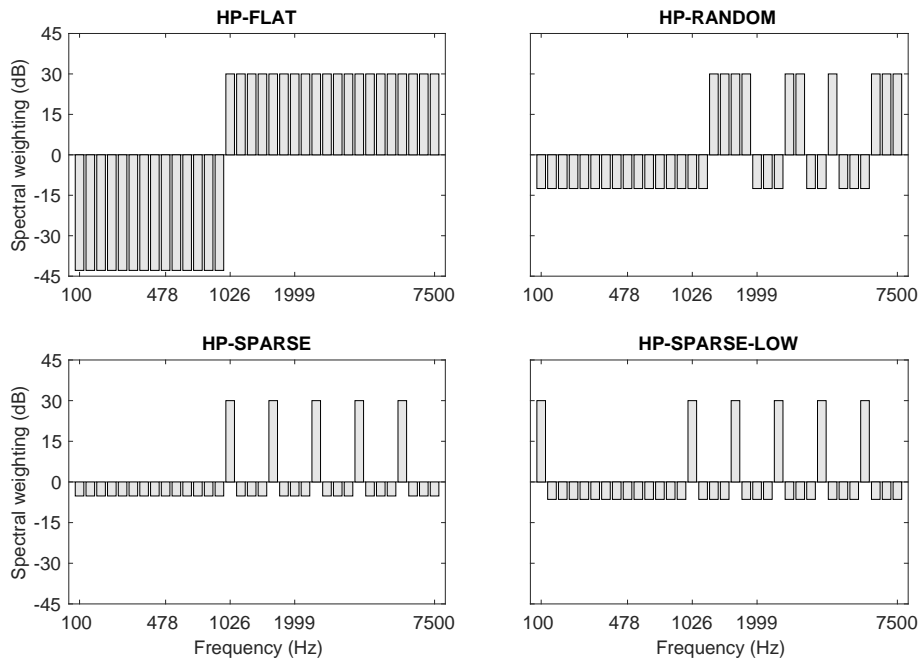


Figure 7: *Boosting patterns investigated in Expt. 2.*

17 optimised weightings derived from the CS, SSN and SMN maskers. At the  
 18 same time, the role of sparse boosting was investigated using three further  
 19 static weightings in which a smaller number of frequency channels received  
 1 a boost, with the side effect of reducing the amount of attenuation in the  
 2 non-boosted channels. The four weightings used in Expt. 2 are depicted in  
 3 Fig. 7. In all cases a constant maximum boost of 30 dB was applied to a range  
 4 of channels, with a commensurate reduction in the level of the non-boosted  
 5 channels to produce a constant speech level following boosting. The 30 dB  
 6 value was chosen as approximately the boosting value observed in this range  
 7 for these maskers in Expt. 1.

- 8 1. HP-FLAT: boost applied to all 20 frequency bands in the region above
- 9 1 kHz, and attenuation of the 14 bands below 1 kHz.
- 10 2. HP-RANDOM: boost 10 frequency bands randomly chosen in the range
- 11 above 1 kHz; attenuation of the remaining 24 bands. Unlike the other
- 12 boosting patterns, a different random set of 10 channels was used for
- 13 each of the 240 sentences.

- 14 3. HP-SPARSE: boost applied sparsely to 5 frequency bands in the range  
15 above 1 kHz; attenuation of the remaining 29 bands. The 5 boosted  
16 locations were evenly-spaced in the range of 1000-7500 Hz.
- 1 4. HP-SPARSE-LOW: as HP-SPARSE, with the addition of the lowest band  
2 centred on 100 Hz. This pattern was motivated by the presence of a  
3 similar low-frequency boost in the CS and SMN optimised spectral  
4 weights observed earlier.

### 5 5.1. Listeners and procedure

6 Participants in Expt. 2 constituted a non-overlapping cohort of 22 native  
7 Spanish listeners (mean 21.1 years; std. dev. 2.3 years) with the same profile  
8 as the listeners of Expt. 1. One participant was excluded from further  
9 analyses following audiometric screening, as was a further participant who  
10 responded to the non-target talker in the CS condition.

11 The stimuli for Expt. 2 were the same 240 sentences used in Expt. 1.  
1 Using a similar balancing procedure as for Expt. 1, stimuli were blocked  
2 by masker (CS, SSN, SMN) and SNR (low, high) into 6 conditions. Eight  
3 sentences from each of the four modifications plus the plain baseline were  
4 presented in a random order in each block.

### 5 5.2. Results

6 Keyword scores for the conditions of Expt. 2 are plotted in Fig. 8.  
7 Clear gains, ranging from 8 to 57 percentage points, were produced in the  
8 HP-FLAT boosting condition, while the other boosting patterns were less suc-  
9 cessful in general, and in some cases led to falls in intelligibility, particularly  
10 in the CS masker.

11 A three-way within-subjects ANOVA with factors of masker, SNR level  
12 and boosting type (plain i.e. none, HP-FLAT, HP-RANDOM, HP-SPARSE,  
13 HP-SPARSE-LOW) was performed on RAU-transformed scores. Apart from  
14 expected main effects of SNR level and masker, there was a strong main  
15 effect of the type of boost applied [ $F(4, 80) = 98.9, p < .001, \eta^2 = .42$ ],  
16 as well as two-way interactions between boost and masker [ $F(8, 160) =$   
1  $33.6, p < .001, \eta^2 = .24$ ] and between boost and SNR [ $F(4, 80) = 4.19, p <$   
2  $.01, \eta^2 = .01$ ]. Based on a Fisher's least significant difference of 7.2 RAUs,  
3 HP-FLAT boosting led to gains in all masking and SNR conditions. HP-  
4 FLAT boosting also outperformed every other boosting type in all condi-  
5 tions. HP-RANDOM and HP-SPARSE produced statistically-equivalent scores

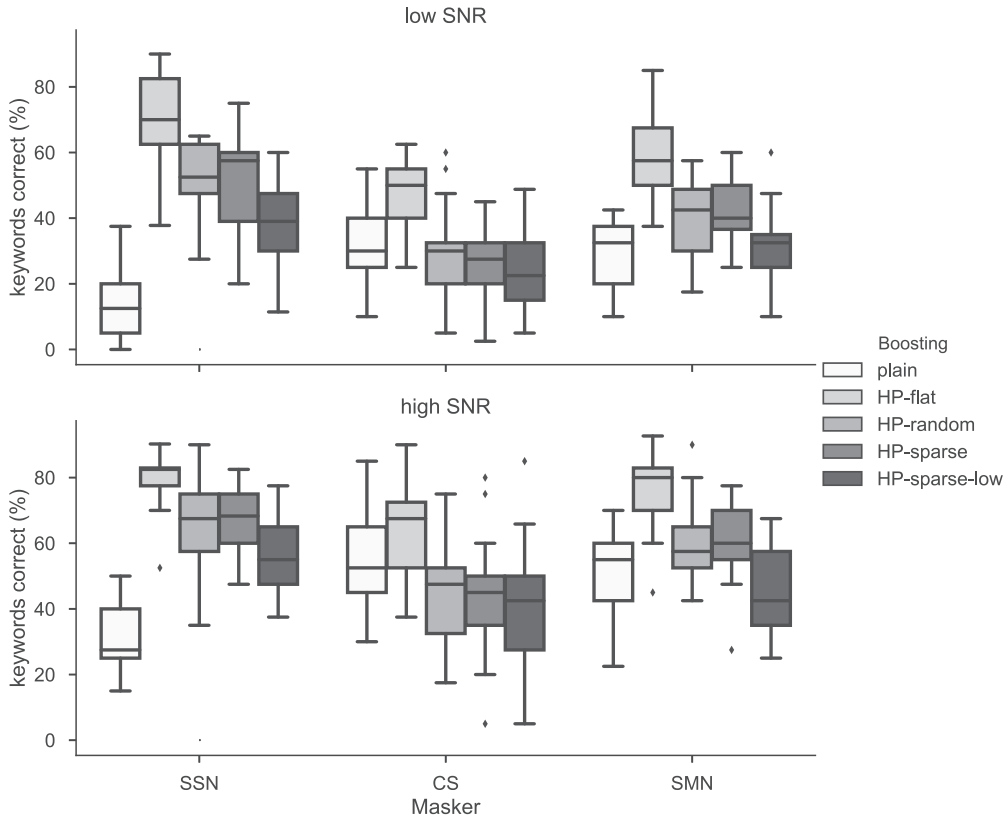


Figure 8: *Keyword scores in plain and modified speech resulting from the four static boosting patterns in the presence of speech-shaped noise (SSN), competing speech (CS) and speech-modulated noise (SMN).*

6 in each condition, while HP-SPARSE-LOW led to smaller gains than either in  
 7 the SSN and SMN masking conditions.

8 Figure 9 compares the mean gains produced by optimised spectral weight-  
 9 ing in Expt. 1 and the generic HP-FLAT weighting in Expt. 2. Apart from  
 10 the high SNR CS masking condition, the gains are of a similar magnitude in  
 11 each condition.

### 12 5.3. Discussion

13 Experiment 2 demonstrates that a very simple generic boosting pattern  
 14 that reallocates energy from the region below 1 kHz to the region above 1 kHz  
 15 is capable of generating similar intelligibility gains in most conditions as pro-

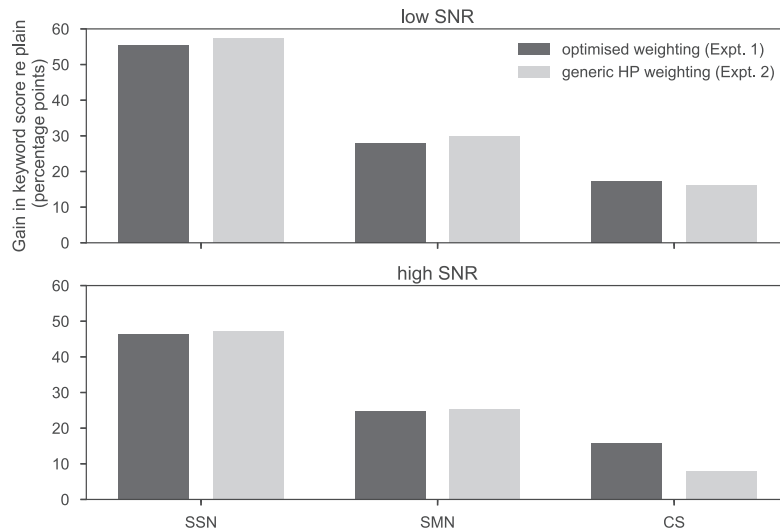


Figure 9: Comparison of gains resulting from optimised spectral weighting (Expt. 1) and generic HP weighting (Expt. 2).

16 duced by the masker- and SNR-specific boosting patterns derived in the first  
 17 experiment. Sparse boosting of a limited number of frequency channels, ei-  
 18 ther fixed (HP-SPARSE) or randomly-chosen (HP-RANDOM) while beneficial,  
 19 did not produce the same degree of improvement. Under the constant input-  
 20 output SNR constraint, boosting the larger number of frequency bands in  
 21 HP-FLAT is compensated for by a greater attenuation of the remaining fre-  
 22 quency bands. It appears that this tradeoff favours uniform high-frequency  
 23 boosting. One speculation is that salient speech information below 1 kHz  
 24 consists in the main of F1 frequency and evidence of voicing in resolved har-  
 25 monic components, and that both of these are available in a limited region  
 26 around the F1 formant frequency whose amplitude resists the increase in  
 27 attenuation implied by the HP-FLAT boosting spectrum.

28 It is not clear why boosting only 5 fixed channels was equivalent to boost-  
 29 ing 10 at random. Given the preceding discussion, it seems unlikely that the  
 30 slightly smaller degree of low frequency attenuation in the 5 channel case  
 31 was responsible for the preservation of intelligibility gain. Instead, it may be  
 32 that listeners find the changing trial-by-trial choice of the 10 boosted chan-

33 nels to be disruptive. Alternatively, the broader boosting regions may have  
34 created ‘false formants’, something that is less likely to occur for the narrower  
35 channels of the sparse boosting condition.

1 One notable finding concerns the effect of boosting in the presence of  
2 competing speech material. In this condition, only HP-FLAT boosting was  
3 helpful. One possibility is that, while HP boosting leads to a smooth spectral  
4 change, preserving local amplitude relationships in all frequency regions apart  
5 from at the 1 kHz boundary, the other boosting patterns tend to disrupt  
6 these relationships, perhaps leading an increase in cognitive load, a form  
7 of informational masking that is particularly deleterious in the presence of  
8 competing speech (Koelewijn et al., 2012a,b).

9 The rather large drops in intelligibility for HP-SPARSE-LOW boosting in  
10 the SSN and SMN conditions is hard to explain, given that the sole difference  
11 with respect to the HP-SPARSE condition is a 30 dB boost to the 100 Hz  
12 frequency band, and a consequent modest attenuation elsewhere. It may  
13 be that such an enhancement creates an artificial grouping cue relating the  
14 very low and the mid-high frequency regions, leading to an artificial source.  
15 The smaller intelligibility drop relative to HP-SPARSE in the CS condition  
16 might be the result of cognitive masking of this spurious grouping cue in  
17 the presence of competing speech. Further studies are needed to explore the  
18 origins of the deleterious effect of low frequency boosting.

## 19 **6. General discussion**

20 Both experiments indicate that speech intelligibility in noise can be in-  
21 creased by large amounts without raising overall RMS level by modifying  
22 the spectral energy distribution of the speech content, using automatically-  
23 derived spectral weighting patterns. Experiment 2 further suggests that a  
24 generic boosting pattern can be as effective as a masker- and SNR-specific  
25 pattern, at least for maskers with a long-term spectrum similar to that of  
26 speech.

27 Methods based on post-filtering (e.g. Hall and Flanagan, 2010; Jokinen  
28 et al., 2012) might achieve similar effects in terms of reallocating energy  
29 across frequencies. While the frequency response of the filter in Jokinen  
30 et al. (2012) is close to the static weighting found in the current study (i.e.  
31 it tends to be flat after approximately 1.5 kHz), the filter proposed in Hall  
32 and Flanagan (2010) has an incremental response from approximately 0.4 to  
33 3.5 kHz, which then holds constant thereafter 3.5 kHz. By comparing the

34 performance of the two filters, Jokinen et al. (2012) demonstrated that the  
35 filter with nearly equal frequency response for mid-high frequencies is more  
36 efficient than that with an incremental response in increasing narrowband  
37 (up to 4 kHz) intelligibility for listeners, especially in more severe conditions.  
1 Having observed that the speech processed by both post-filtering methods is  
2 more intelligible than the unprocessed speech, this could further suggest that  
3 additional intelligibility gain may be obtained by boosting the frequencies in  
4 the region 1.5-3.5 kHz equally, along with high frequencies. Post-filtering  
5 studies confirm the effectiveness of a strategy of injecting energy from else-  
6 where to mid-high frequencies to enhance intelligibility under the constant  
7 input-output SNR constraint.

8 As a practical approach to near-end listening enhancement, spectral weight-  
9 ing is appealing because it is fast to implement and does not call for a de-  
10 tailed, time-varying estimate of the ongoing masker spectrum. Instead, the  
11 method requires an estimate of masker type and SNR (e.g. Rombouts et al.,  
12 2006) in order to select the appropriate frequency response. Whether such  
13 a masker-dependent approach can be implemented in real-time will depend  
14 on how quickly the masker type and SNR are changing. In circumstances  
15 where both the masker type and SNR tend to be constant, it is possible to  
16 estimate these properties prior to applying the speech modification. Further  
17 work will determine how fine-grained a classification of both factors is nec-  
18 essary, but the outcome of Expt. 2 suggests that even a coarse estimate will  
19 lead to some benefits. The static weighting approach requires only a constant  
20 high-pass filter to be applied to the input signal, and can be implemented  
21 with a minimal delay.

22 The basis for the intelligibility improvements from spectral weighting  
23 remain unclear. For example, there is no simple relationship between the  
24 weightings uncovered in Expt. 1 and previously-reported frequency impor-  
25 tance functions for speech, which themselves present a mixed picture as to  
26 where the salient spectral bands lie (Studebaker et al., 1987). While some  
27 studies have reported peaks of importance in the 2 kHz region (e.g. French  
28 and Steinberg, 1947; DePaolis et al., 1996), others have suggested a near-  
29 equal weighting of importance above and below 1 kHz (Studebaker et al.,  
30 1987; Healy et al., 2013). The effect of incorporating frequency importance  
31 functions into the spectral weighting procedure is worthy of further investi-  
32 gation.

33 The quality of the modified speech may be an issue when being presented  
34 in mild noise or noise-free conditions, since any artefacts introduced by the

35 modification to the speech signal may become perceptually noticeable by the  
36 listener, leading to a potential reduction in speech quality. Previous stud-  
1 ies (e.g. Tang and Cooke, 2010) investigating the quality of algorithmically-  
2 modified speech using the perceptual evaluation of speech quality (PESQ,  
3 ITU-T P.862, 2001) suggest that spectral modification alone has a relatively  
4 small negative impact on PESQ, compared to modifications performed in  
5 the time and the time-frequency domains. PESQ scores were also calculated  
6 for all modifications in the current study. For the noise-dependent spectral  
7 weightings, the PESQ values fall in the range between 3.8 and 4.2 across the  
8 six maskers. For the static weightings, HP-FLAT and HP-SPARSE-LOW lead  
9 to the best (4.0) and the worst (3.7) PESQ scores, respectively. These results  
10 are consistent with our previous findings in Tang and Cooke (2010). How-  
11 ever, there is some evidence revealing that severely attenuating frequencies  
12 where pitch and harmonic information exist may lead to poor perceptual  
13 speech quality in quiet (Gabrielsson et al., 1988). This might explain the  
14 finding of Jokinen et al. (2012) that post-filtering with a frequency response  
15 which has smooth transition between low and mid frequencies leads to better  
16 naturalness of the processed speech than when the filter has a steep cut-off.  
17 Thus, when deploying speech modification techniques in practice, it may be  
18 worthwhile performing SNR estimation online (e.g. Jokinen et al., 2012), in  
19 order to determine the threshold for modification (de)activation in respect  
20 to speech quality.

## 21 **Conclusions**

22 Modifying clean speech prior to presentation by the simple expedient of  
23 altering its spectral balance without changing its RMS level can be a highly-  
24 effective way to increase intelligibility in the presence of masking noise. The  
25 current study demonstrates that masker-dependent spectral weightings can  
26 be learnt by maximising the value of an objective intelligibility metric, ob-  
27 viating the need for detailed time-varying masker estimates during speech  
28 presentation. Further, generic spectral weighting patterns that boost en-  
29 ergy above 1 kHz are beneficial for maskers with a speech-shaped long-term  
1 spectrum.

## 2 **Acknowledgements**

3 This study was supported by the LISTA Project (<http://listening-talker.org>),  
4 funded by the Future and Emerging Technologies programme within the 7th

5 Framework Programme for Research of the European Commission, FET-  
6 Open grant number 256230. The authors thank Máté Attila Toth for re-  
7 cruiting participants for the listening experiments.

8 ANSI S3.5-1997, 1997. Methods for the calculation of the Speech Intelligibil-  
9 ity Index.

10 Aubanel, V., Garcia Lecumberri, M. L., Cooke, M., 2014. The Sharvard  
11 corpus: A phonemically-balanced Spanish sentence resource for audiology.  
12 *Int. J. Audiology* 53, 633–638.

13 Barker, J., Cooke, M., 2007. Modelling speaker intelligibility in noise. *Speech*  
14 *Communication* 49, 402–417.

15 Bell, S. T., Dirks, D. D., Trine, T. D., 1992. Frequency-importance functions  
16 for words in high- and low-context sentences. *J. Speech Hear. Res.* 35,  
17 950–959.

18 Bonardo, D., Zovato, E., 2007. Speech synthesis enhancement in noisy envi-  
19 ronments. In: *Proc. Interspeech*. pp. 2853–2856.

20 Boril, H., Pollak, P., 2005. Analysis of Lombard effect appearance in several  
21 Czech databases. In: *Electronic speech signal processing conference; 16th,*  
22 *Electronic speech signal processing*.

23 Brouckxon, H., Verhelst, W., Schuymer, B. D., 2008. Time and frequency  
24 dependent amplification for speech intelligibility enhancement in noisy en-  
25 vironments. In: *Proc. Interspeech*. pp. 557–560.

26 Brungart, D. S., Simpson, B. D., Ericson, M. A., Scott, K. R., 2001. In-  
27 formational and energetic masking effects in the perception of multiple  
28 simultaneous talkers. *J. Acoust. Soc. Am.* 110 (5), 2527–2538.

29 Chen, J., Benesty, J., Huang, Y., Doclo, S., 2006. New insights into the  
30 noise reduction Wiener filter. *IEEE Trans. Audio, Speech, and Language*  
1 *Processing* 14 (1), 1218–1234.



- 2 Christiansen, C., Pedersen, M. S., Dau, T., 2010. Prediction of speech intelli-  
3 gibility based on an auditory preprocessing model. *Speech Communication*  
4 52 (7-8), 678–692.
- 5 Cooke, M., 2006. A glimpsing model of speech perception in noise. *J. Acoust.*  
6 *Soc. Am.* 119 (3), 1562–1573.
- 7 Cooke, M., Lu, Y., 2010. Spectral and temporal changes to speech produced  
8 in the presence of energetic and informational maskers. *J. Acoust. Soc.*  
9 *Am.* 128 (4), 2059–2069.
- 10 Cooke, M., Mayo, C., Valentini-Botinhao, C., 2013a. Intelligibility-enhancing  
11 speech modifications: the Hurricane Challenge. In: *Proc. Interspeech*. pp.  
12 3552–3556.
- 13 Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., Tang,  
14 Y., 2013b. Evaluating the intelligibility benefit of speech modifications in  
15 known noise conditions. *Speech Communication* 55, 572–585.
- 16 Davidon, W., 1991. Variable metric method for minimization. *SIAM Journal*  
17 *on Optimization* 1 (1), 1–17.
- 18 DePaolis, R. A., Janota, C. P., Frank, T., 1996. Frequency importance func-  
19 tions for words, sentences, and continuous discourse. *J. Speech, Lang. Hear.*  
20 *Res.* 39, 714–723.
- 21 French, N. R., Steinberg, J. C., 1947. Factors governing the intelligibility of  
22 speech sounds. *J. Acoust. Soc. Am.* 19 (1), 90–119.
- 23 Gabrielsson, A., Schenkman, B. N., Hagerman, B., 1988. The effects of dif-  
24 ferent frequency responses on sound quality judgments and speech intelli-  
25 gibility. *J. Speech Lang. Hear. Res.* 31 (2), 166–177.
- 26 Hall, J. L., Flanagan, J. L., 2010. Intelligibility and listener preference of  
27 telephone speech in the presence of babble noise. *J. Acoust. Soc. Am.*  
28 127 (1), 280–285.
- 29 Healy, E. W., Yoho, S. E., Apoux, F., 2013. Band importance for sentences  
30 and words reexamined. *J. Acoust. Soc. Am.* 133 (1), 463–473.
- 31 Holland, J. H., 1975. *Adaptation in Natural and Artificial Systems*. Vol. Ann  
1 Arbor. University of Michigan Press.

- 2 Hooke, R., Jeeves, T., 1961. “Direct search” solution of numerical and statis-  
3 tical problems. *Journal of the Association for Computing Machinery* 8 (2),  
4 212–229.
- 5 Hu, Y., Loizou, P. C., 2004. Speech enhancement based on wavelet thresh-  
6 olding the multitaper spectrum. *IEEE Trans. Speech Audio Processing*,  
7 59–67.
- 8 ITU-T P.862, 2001. P.862: Perceptual evaluation of speech quality (PESQ):  
9 An objective method for end-to-end speech quality assessment of narrow-  
10 band telephone networks and speech codecs.
- 11 Jokinen, E., Pulakka, H., Alku, P., 2016. Phase modification for increas-  
12 ing the intelligibility of telephone speech in near-end noise conditions –  
13 evaluation of two methods. *Speech Communication* 83, 64–80.
- 14 Jokinen, E., Yrttiaho, S., Pulakka, H., Vainio, M., Alku, P., 2012. Signal-to-  
15 noise ratio adaptive post-filtering method for intelligibility enhancement  
16 of telephone speech. *J. Acoust. Soc. Am.* 132 (6), 3990–4001.
- 17 Junqua, J. C., Fincke, S., Field, K., 1998. Influence of the speaking style  
18 and the noise spectral tilt on the Lombard reflex and automatic speech  
19 recognition. In: *International Conference Spoken Language Proceedings*.  
20 pp. 467–470.
- 21 Kates, J., Arehart, K., 2005. Coherence and the speech intelligibility index.  
22 *J. Acoust. Soc. Am.* 117 (4), 2224–2237.
- 23 Kim, G., Lu, Y., Hu, Y., Loizou, P. C., 2009. An algorithm that improves  
24 speech intelligibility in noise for normal-hearing listeners. *J. Acoust. Soc.*  
25 *Am.* 126 (3), 1486–1494.
- 26 Knobel, K. A., Sanche, T. G., 2006. Loudness discomfort level in normal  
27 hearing individuals. *Pro Fono.* 18 (1), 31–40.
- 28 Koelewijn, T., Zekveld, A. A., Festen, J. M., Kramer, S. E., 2012a. Pupil  
29 dilation uncovers extra listening effort in the presence of an interfering  
1 speaker. *Ear Hear.* 33, 291–300.
- 2 Koelewijn, T., Zekveld, A. A., Festen, J. M., Rönnerberg, J., Kramer, S. E.,  
3 2012b. Processing load induced by informational masking is related to  
4 linguistic abilities. *Int. J. Otolaryngol.* 865731.

- 5 Lombard, E., 1911. Le signe de l'elevation de la voix. *Annals maladiers oreille,*  
6 *Larynx, Nez, Pharynx* 37, 101–119.
- 7 Martin, R., 2005. Speech Enhancement Based on Minimum Mean-Square  
8 Error Estimation and Supergaussian Priors. *IEEE Trans. Speech Audio*  
9 *Processing* 13 (5), 845–856.
- 10 Mayer, A. M., 1894. Research in acoustics. *Lond. Edinb. Dubl. Phil. Mag.*  
11 *Ser. 5*, 259–288.
- 12 Mitchell, M., 1996. *An Introduction to Genetic Algorithms*. MIT Press.
- 13 Moore, B. C. J., 2003. Temporal integration and context effects in hearing.  
14 *Journal of Phonetics* 31 (3–4), 563–574.
- 15 Paliwal, K. K., Alsteris, L. D., 2005. On the usefulness of STFT phase spec-  
16 trum in human listening tests. *Speech Communication* 45 (2), 153–170.
- 17 Rombouts, G., van Waterschoot, T., Struyve, K., Moonen, M., 2006. Acous-  
18 tic feedback cancellation for long acoustic paths using a nonstationary  
19 source model. *IEEE Transactions on Signal Processing* 54 (9), 3426–3434.
- 20 Rothauser, E. H., Chapman, W. D., Guttman, N., Silbiger, H. R., Hecker,  
21 M. H. L., Urbanek, G. E., Nordby, K. S., Weinstock, M., 1969. IEEE Rec-  
22 ommended practice for speech quality measurements. *IEEE Trans. Audio*  
23 *Electroacoust* 17, 225–246.
- 24 Sabin, W. E., Schoenike, E. O., 1998. *HF Radio Systems & Circuits*, rev.  
25 2nd ed Edition. Noble.
- 26 Sauert, B., Vary, P., 2006. Near end listening enhancement: speech intelli-  
27 gibility improvement in noise environments. In: *Proc. ICASSP*. pp. 493–496.
- 28 Sauert, B., Vary, P., 2009. Near End Listening Enhancement Optimized  
29 with Respect to Speech Intelligibility Index. In: *Proc. EUSIPCO*. Glas-  
30 gow, Scotland, pp. 1844–1848.
- 1 Schepker, H., Rennies, J., Doclo, S., 2015. Speech-in-noise enhancement using  
2 amplification and dynamic range compression controlled by the speech  
3 intelligibility index. *J. Acoust. Soc. Am.* 138 (5), 2692–2706.

- 4 Srinivasan, S., Samuelsson, J., Kleijn, W., 2007. Codebook-Based Bayesian  
5 speech enhancement for nonstationary environments. *IEEE Trans. Audio,  
6 Speech, and Language Processing*, 441–452.
- 7 Stubebaker, G. A., Sherbecoe, R. L., 1991. Frequency-importance and trans-  
8 fer functions for recorded CID W-22 word lists. *J. Speech Hear. Res.* 34,  
9 427–438.
- 10 Studebaker, G. A., 1985. A ‘rationalized’ arcsine transform. *J. Speech Hear.  
11 Res.* 28, 455–462.
- 12 Studebaker, G. A., Pavlovic, C. V., Sherbecoe, R. L., 1987. A frequency  
13 importance function for continuous discourse. *J. Acoust. Soc. Am.* 81 (4),  
14 1130–1138.
- 15 Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., Stokes,  
16 M. A., 1988. Effects of noise on speech production: acoustic and perceptual  
17 analyses. *J. Acoust. Soc. Am.* 84 (3), 917–928.
- 18 Taal, C., Hendriks, R. C., Heusdens, R., 2014. Speech energy redistribution  
19 for intelligibility improvement in noise based on a perceptual distortion  
20 measure. *Computer Speech and Language* 28 (4), 858–872.
- 21 Taal, C., Heusdens, R., 2009. A low-complexity spectro-temporal based per-  
22 ceptual model. In: *Proc. ICASSP*. pp. 153–156.
- 23 Tang, Y., Cooke, M., 2010. Energy reallocation strategies for speech enhance-  
24 ment in known noise conditions. In: *Proc. Interspeech. Makuhari, Japan*,  
25 pp. 1636–1639.
- 26 Tang, Y., Cooke, M., 2012. Optimised spectral weightings for noise-  
731 dependent speech intelligibility enhancement. In: *Proc. Interspeech. Port-  
732 land, US*, pp. 955–958.
- 733 Tang, Y., Cooke, M., Valentini-Botinhao, C., 2016. Evaluating the predic-  
734 tions of objective intelligibility metrics for modified and synthetic speech.  
735 *Computer Speech and Language* 35, 73–92.
- 736 Tang, Y., Cooke, M. P., 2016. Glimpse-Based metrics for predicting speech  
737 intelligibility in additive noise conditions. In: *Proc. Interspeech. San Fran-  
738 cisco, US*, pp. 2488–2492.

- 739 Viktorovitch, M., 2005. Implementation of a new metric for assessing and  
740 optimizing the speech intelligibility inside cars. Tech. rep., SAE Technical  
741 Paper.
- 742 Wegel, R. L., Lane, C. E., 1924. The auditory masking of one sound by  
743 another and its probable relation to the dynamics of the inner ear. *Phys.*  
744 *Rev.* 23, 266–285.
- 745 Williamson, D. S., Wang, Y., Wang, D., 2015. Estimating nonnegative matrix  
746 model activations with deep neural networks to increase perceptual speech  
747 quality. *J. Acoust. Soc. Am.* 138 (3), 1399–1407.
- 748 Yoo, S. D., Boston, J. R., El-Jaroudi, A., Li, C., Durrant, J. D., Kovacyk, K.,  
749 S., S., 2007. Speech signal modification to increase intelligibility in noisy  
750 environments. *J. Acoust. Soc. Am.* 122 (2), 1138–1149.
- 751 Yu, W. C., 1979. The convergent property of the simplex evolutionary tech-  
752 nique. *Scientia Sinica [Zhongguo Kexue]*, 69–77.
- 753 Zorila, T. C., Kandia, V., Stylianou, Y., 2012. Speech-in-noise intelligibility  
754 improvement based on spectral shaping and dynamic range compression.  
755 In: *Proc. Interspeech*. pp. 635–638.