Learning static spectral weightings for speech intelligibility enhancement in noise

Tang, Y and Cooke, M

http://dx.doi.org/10.1016/j.csl.2017.10.003

Learning static spectral weightings for speech intelligibility enhancement in noise

Yan Tang[a,*], Martin Cooke[b,c]

[a] *Acoustics Research Centre, University of Salford, UK*
[b] *Ikerbasque (Basque Science Foundation), Bilbao, Spain*
[c] *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 (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

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

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

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

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

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

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

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

2. High energy glimpse proportion

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

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

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

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

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

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

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

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

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

3. Optimised spectral weightings

3.1. Maximising HEGP via Pattern Search optimisation

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

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

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

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

3.2. Speech material

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

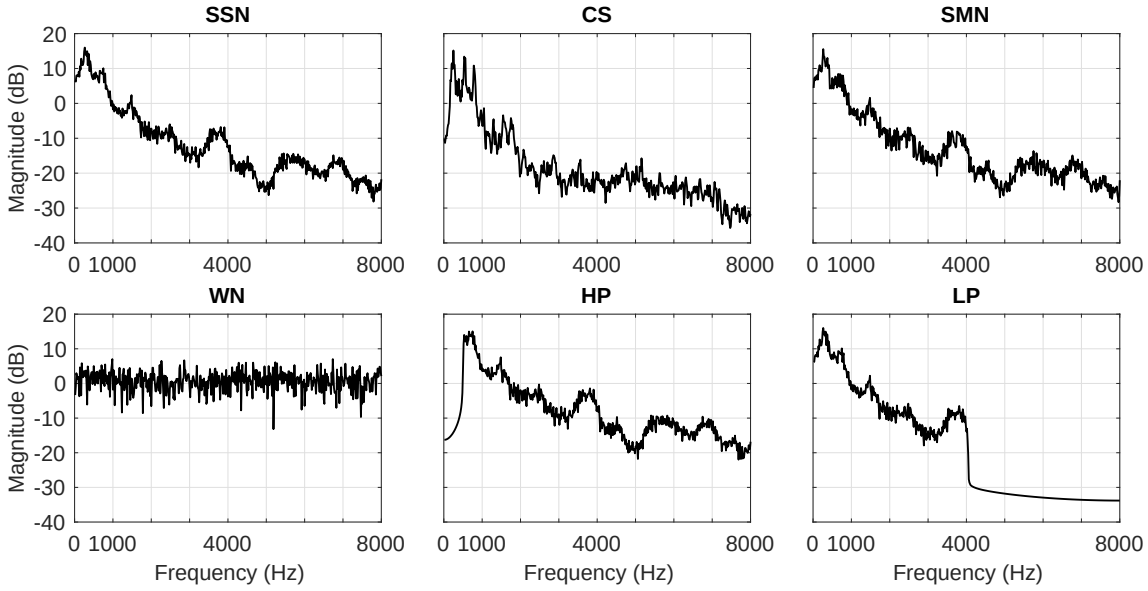
3.3. Maskers and SNRs

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

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

3.4. Spectral weighting candidates

In order to examine the consistency of the spectral weight patterns learnt in individual optimisation runs, the output of two trials of the optimisation 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.*

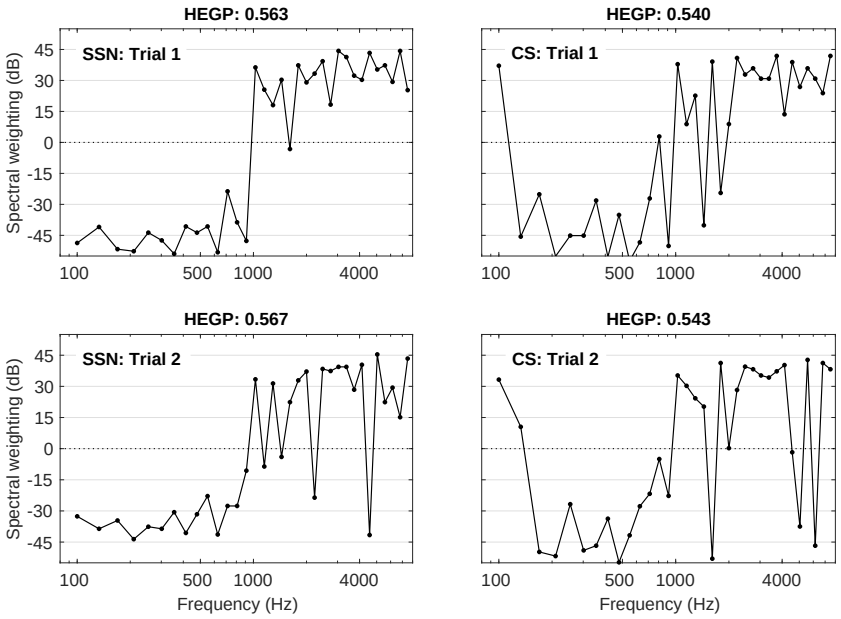


Figure 2: *Best spectral weightings from two separate PS optimisation runs for* SSN *(left) and* CS *(right) maskers in the high SNR condition. The HEGP score for each weighting is also displayed.*

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

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

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

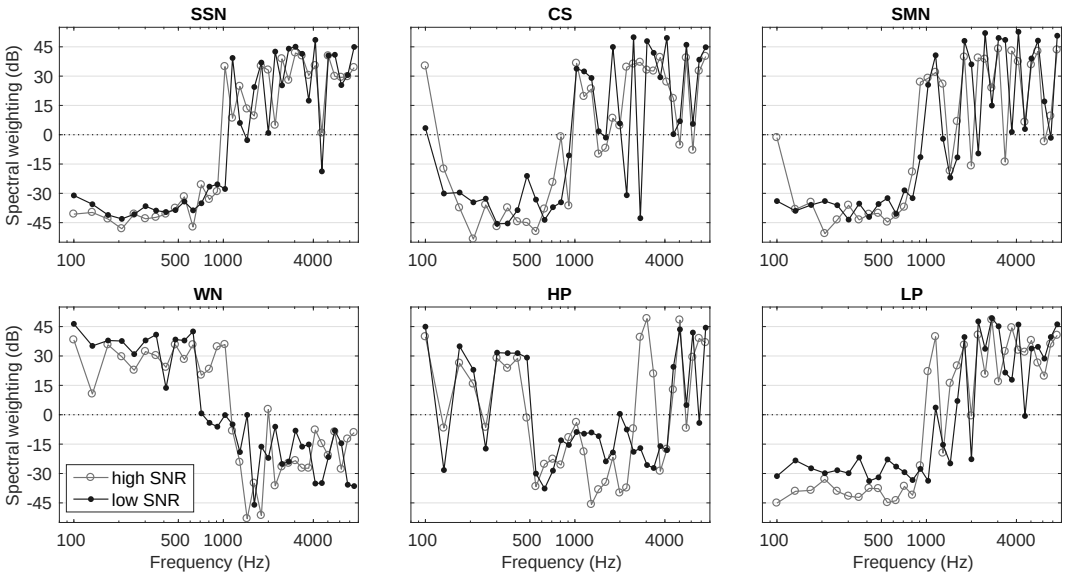


Figure 3: *Optimised spectral weightings discovered by Pattern Search.*

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

4.1. Speech and masker material

A set of 240 sentences drawn from the male talker of the Sharvard corpus (sentences 401-640) was used in Expt. 1. These sentences are different from the 100-member development set employed during pattern search optimisation. Maskers and SNRs were the same as those used in the learning phase.

4.2. Listeners

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

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

4.3. Procedure

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

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

4.4. Post-processing

Subjective intelligibility was computed as the correct keyword recognition rate for each condition. Five keywords per sentence were used for scoring. Due to inconsistent use of diacritics by listeners, all vowel accents were removed prior to scoring so that answers with or without accents were considered to be equivalent e.g. both 'río' and 'rio' were considered to be correct responses for the word 'río'. For statistical purposes, percentages were converted to rationalised arcsine units (RAU; Studebaker, 1985). However, for ease of interpretation, results are shown in percentages in the figures.

4.5. Results

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

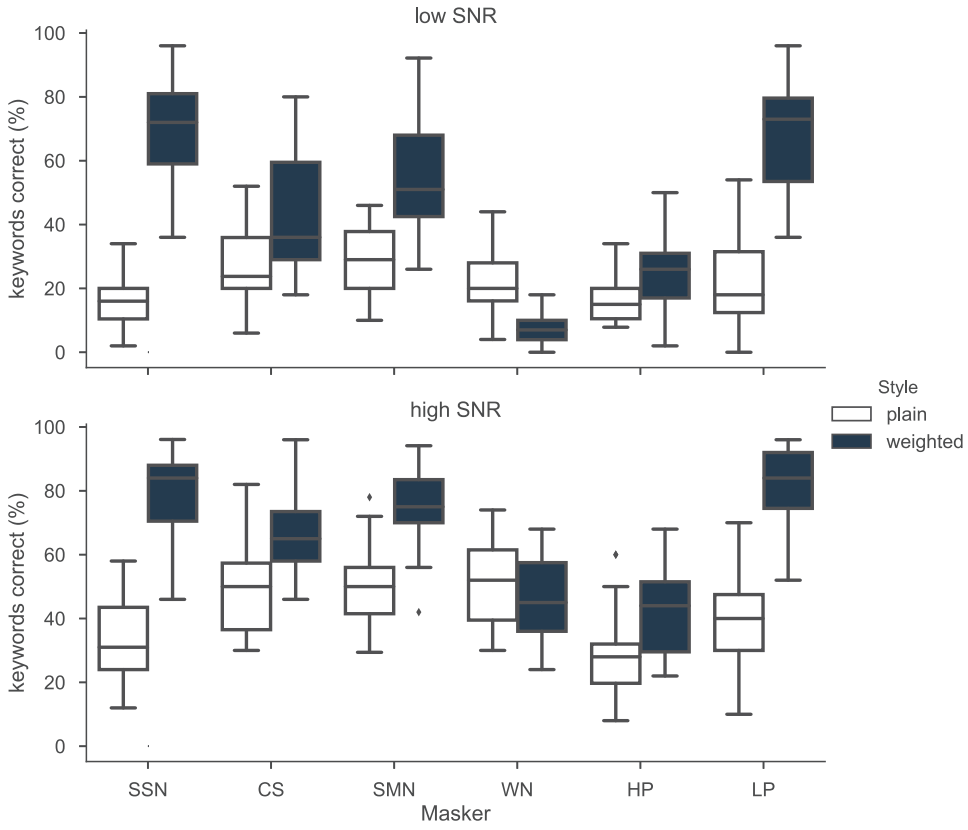


Figure 4: *Intelligibility scores for spectrally-weighted speech and unmodified 'plain' speech in the presence of masking noise at low (top) and high (bottom) SNRs.*

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

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

4.6. Discussion

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

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

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

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

there is no basis for preferring non-neighbouring channels in the HEGP metric.

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

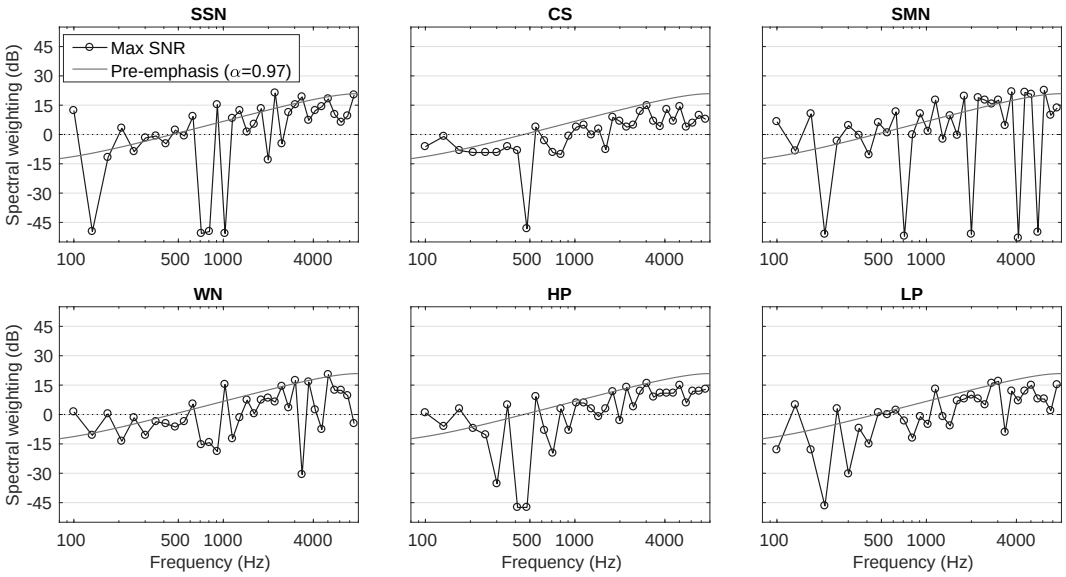


Figure 5: *Spectral weightings learnt by maximising average across-frequency-band SNR. The frequency response of a pre-emphasis filter ($\alpha = 0.97$) is shown with a 15-dB offset.*

When comparing the average band SNR of speech modified by maximising

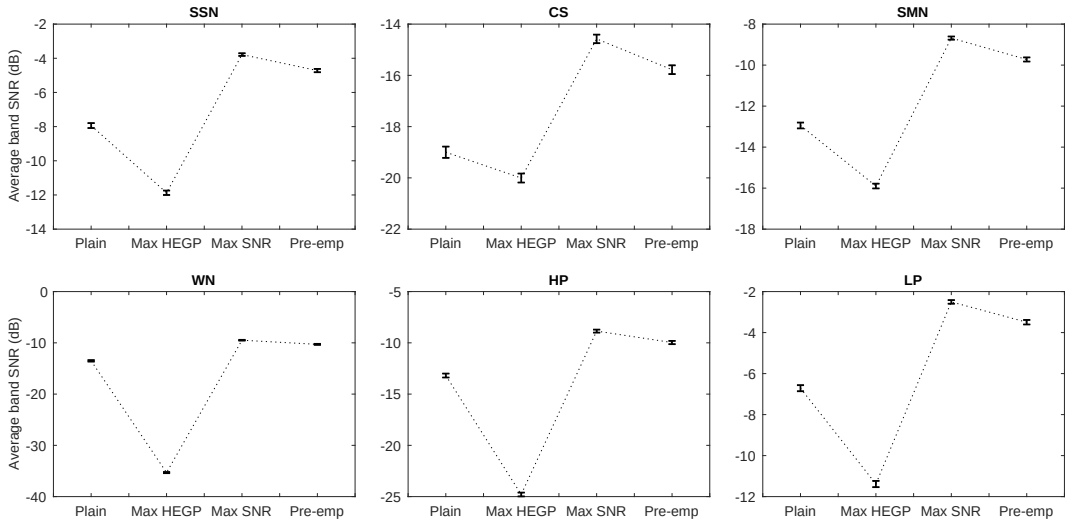


Figure 6: *Average band SNR of unmodified speech and speech modified by the Max HEGP, Max SNR and pre-emphasis weighting. Error bars denote 95% confidence intervals.*

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

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

5. Experiment 2: The effect of generic spectral weightings

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

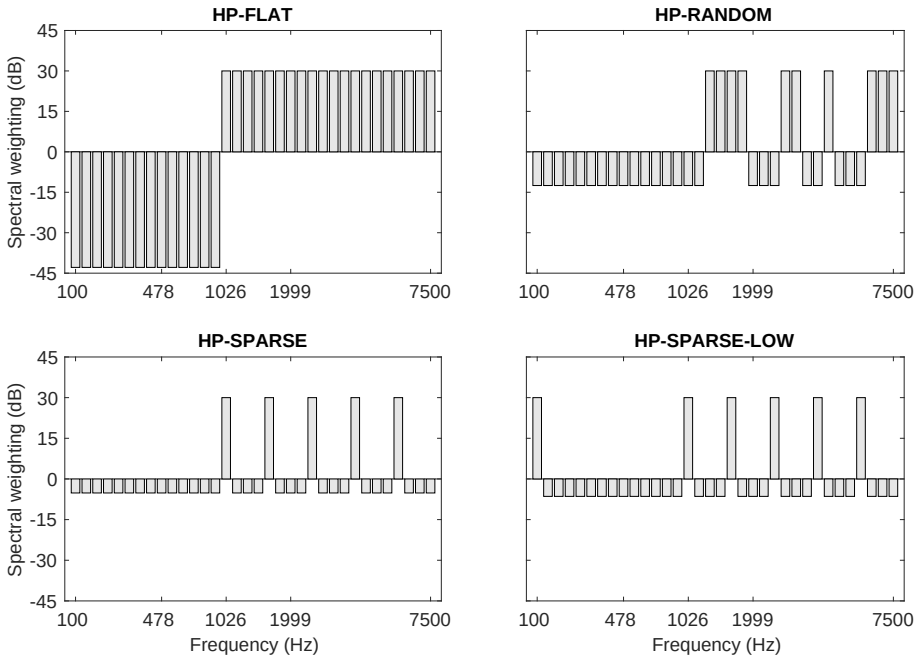


Figure 7: *Boosting patterns investigated in Expt. 2.*

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

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

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

5.1. Listeners and procedure

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

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

5.2. Results

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

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

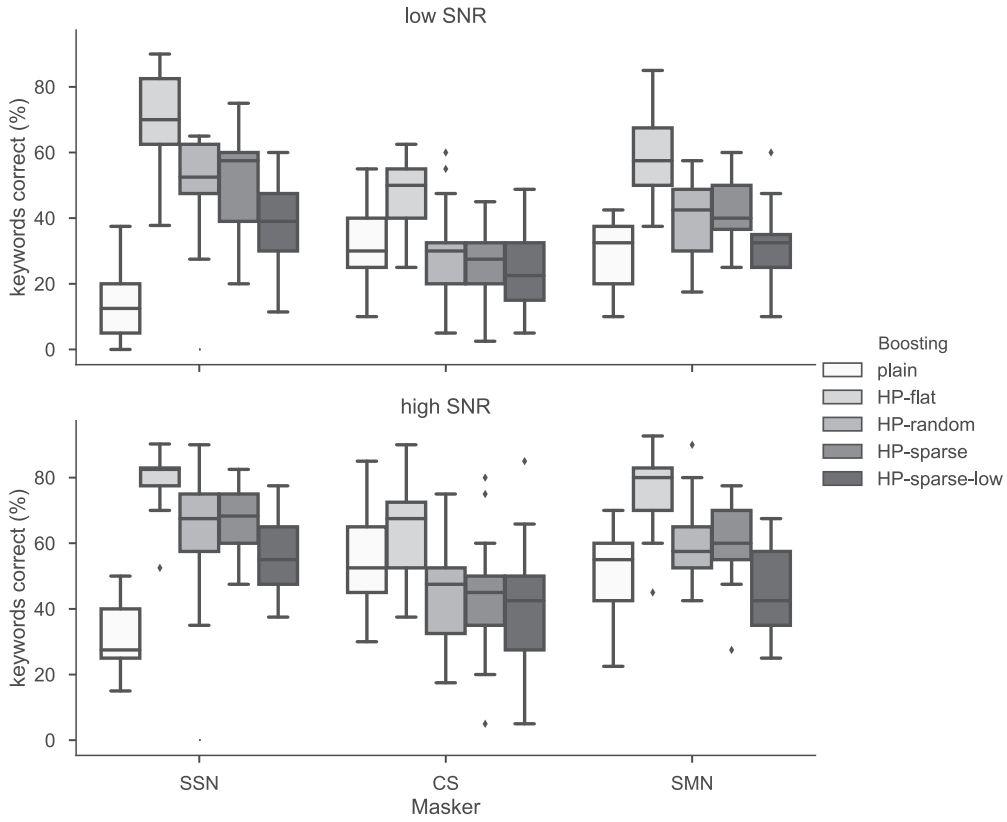


Figure 8: *Keyword scores in plain and modified speech resulting from the four static boosting patterns in the presence of speech-shaped noise (SSN), competing speech (CS) and speech-modulated noise (SMN).*

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

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

5.3. Discussion

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

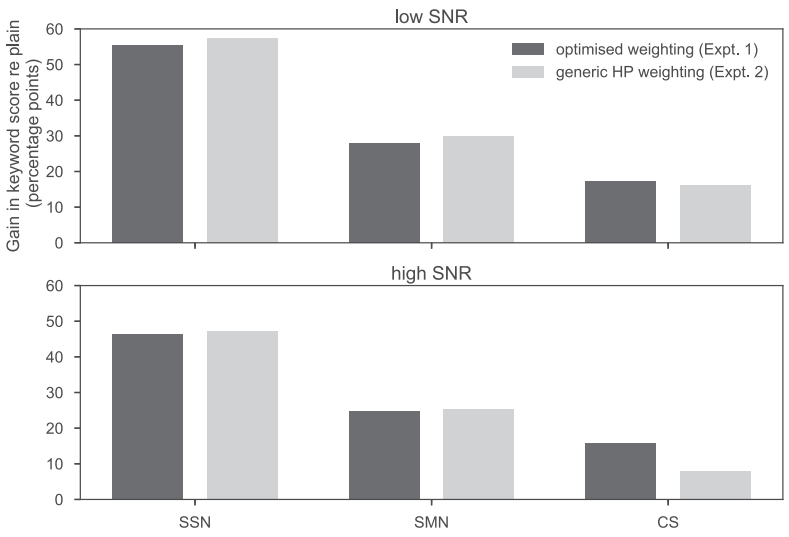


Figure 9: *Comparison of gains resulting from optimised spectral weighting (Expt. 1) and generic* HP *weighting (Expt. 2).*

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

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

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

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

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

6. General discussion

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

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

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

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

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

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

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

Conclusions

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

Acknowledgements

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

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

ANSI S3.5-1997, 1997. Methods for the calculation of the Speech Intelligibility Index.

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

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

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

Bonardo, D., Zovato, E., 2007. Speech synthesis enhancement in noisy environments. In: Proc. Interspeech. pp. 2853–2856.

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

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

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

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

Christiansen, C., Pedersen, M. S., Dau, T., 2010. Prediction of speech intelligibility based on an auditory preprocessing model. Speech Communication 52 (7-8), 678–692.

Cooke, M., 2006. A glimpsing model of speech perception in noise. J. Acoust. Soc. Am. 119 (3), 1562–1573.

Cooke, M., Lu, Y., 2010. Spectral and temporal changes to speech produced in the presence of energetic and informational maskers. J. Acoust. Soc. Am. 128 (4), 2059–2069.

Cooke, M., Mayo, C., Valentini-Botinhao, C., 2013a. Intelligibility-enhancing speech modifications: the Hurricane Challenge. In: Proc. Interspeech. pp. 3552–3556.

Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., Tang, Y., 2013b. Evaluating the intelligibility benefit of speech modifications in known noise conditions. Speech Communication 55, 572–585.

Davidon, W., 1991. Variable metric method for minimization. SIAM Journal on Optimization 1 (1), 1–17.

DePaolis, R. A., Janota, C. P., Frank, T., 1996. Frequency importance functions for words, sentences, and continuous discourse. J. Speech, Lang. Hear. Res. 39, 714–723.

French, N. R., Steinberg, J. C., 1947. Factors governing the intelligibility of speech sounds. J. Acoust. Soc. Am. 19 (1), 90–119.

Gabrielsson, A., Schenkman, B. N., Hagerman, B., 1988. The effects of different frequency responses on sound quality judgments and speech intelligibility. J. Speech Lang. Hear. Res. 31 (2), 166–177.

Hall, J. L., Flanagan, J. L., 2010. Intelligibility and listener preference of telephone speech in the presence of babble noise. J. Acoust. Soc. Am. 127 (1), 280–285.

Healy, E. W., Yoho, S. E., Apoux, F., 2013. Band importance for sentences and words reexamined. J. Acoust. Soc. Am. 133 (1), 463–473.

Holland, J. H., 1975. Adaptation in Natural and Artificial Systems. Vol. Ann Arbor. University of Michigan Press.

Hooke, R., Jeeves, T., 1961. "Direct search" solution of numerical and statistical problems. Journal of the Association for Computing Machinery 8 (2), 212–229.

Hu, Y., Loizou, P. C., 2004. Speech enhancement based on wavelet thresholding the multitaper spectrum. IEEE Trans. Speech Audio Processing, 59–67.

ITU-T P.862, 2001. P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.

Jokinen, E., Pulakka, H., Alku, P., 2016. Phase modification for increasing the intelligibility of telephone speech in near-end noise conditions – evaluation of two methods. Speech Communication 83, 64–80.

Jokinen, E., Yrttiaho, S., Pulakka, H., Vainio, M., Alku, P., 2012. Signal-to-noise ratio adaptive post-filtering method for intelligibility enhancement of telephone speech. J. Acoust. Soc. Am. 132 (6), 3990–4001.

Junqua, J. C., Fincke, S., Field, K., 1998. Influence of the speaking style and the noise spectral tilt on the Lombard reflex and automatic speech recognition. In: International Conference Spoken Language Proceedings. pp. 467–470.

Kates, J., Arehart, K., 2005. Coherence and the speech intelligibility index. J. Acoust. Soc. Am. 117 (4), 2224–2237.

Kim, G., Lu, Y., Hu, Y., Loizou, P. C., 2009. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. J. Acoust. Soc. Am. 126 (3), 1486–1494.

Knobel, K. A., Sanche, T. G., 2006. Loudness discomfort level in normal hearing individuals. Pro Fono. 18 (1), 31–40.

Koelewijn, T., Zekveld, A. A., Festen, J. M., Kramer, S. E., 2012a. Pupil dilation uncovers extra listening effort in the presence of an interfering speaker. Ear Hear. 33, 291–300.

Koelewijn, T., Zekveld, A. A., Festen, J. M., Rönnberg, J., Kramer, S. E., 2012b. Processing load induced by informational masking is related to linguistic abilities. Int. J. Otolaryngol. 865731.

Lombard, E., 1911. Le signe de l'elevation de la voix. Annals maladiers oreille, Larynx, Nez, Pharynx 37, 101–119.

Martin, R., 2005. Speech Enhancement Based on Minimum Mean-Square Error Estimation and Supergaussian Priors. IEEE Trans. Speech Audio Processing 13 (5), 845–856.

Mayer, A. M., 1894. Research in acoustics. Lond. Edinb. Dubl. Phil. Mag. Ser. 5, 259–288.

Mitchell, M., 1996. An Introduction to Genetic Algorithms. MIT Press.

Moore, B. C. J., 2003. Temporal integration and context effects in hearing. Journal of Phonetics 31 (3–4), 563–574.

Paliwal, K. K., Alsteris, L. D., 2005. On the usefulness of STFT phase spectrum in human listening tests. Speech Communication 45 (2), 153–170.

Rombouts, G., van Waterschoot, T., Struyve, K., Moonen, M., 2006. Acoustic feedback cancellation for long acoustic paths using a nonstationary source model. IEEE Transactions on Signal Processing 54 (9), 3426–3434.

Rothauser, E. H., Chapman, W. D., Guttman, N., Silbiger, H. R., Hecker, M. H. L., Urbanek, G. E., Nordby, K. S., Weinstock, M., 1969. IEEE Recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoust 17, 225–246.

Sabin, W. E., Schoenike, E. O., 1998. HF Radio Systems & Circuits, rev. 2nd ed Edition. Noble.

Sauert, B., Vary, P., 2006. Near end listening enhancement: speech intelligibility improvement in noise environments. In: Proc. ICASSP. pp. 493–496.

Sauert, B., Vary, P., 2009. Near End Listening Enhancement Optimized with Respect to Speech Intelligibility Index. In: Proc. EUSIPCO. Glasgow, Scotland, pp. 1844–1848.

Schepker, H., Rennies, J., Doclo, S., 2015. Speech-in-noise enhancement using amplification and dynamic range compression controlled by the speech intelligibility index. J. Acoust. Soc. Am. 138 (5), 2692–2706.

Srinivasan, S., Samuelsson, J., Kleijn, W., 2007. Codebook-Based Bayesian speech enhancement for nonstationary environments. IEEE Trans. Audio, Speech, and Language Processing, 441–452.

Stubebaker, G. A., Sherbecoe, R. L., 1991. Frequency-importance and transfer functions for recorded CID W-22 word lists. J. Speech Hear. Res. 34, 427–438.

Studebaker, G. A., 1985. A 'rationalized' arcsine transform. J. Speech Hear. Res. 28, 455–462.

Studebaker, G. A., Pavlovic, C. V., Sherbecoe, R. L., 1987. A frequency importance function for continuous discourse. J. Acoust. Soc. Am. 81 (4), 1130–1138.

Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., Stokes, M. A., 1988. Effects of noise on speech production: acoustic and perceptual analyses. J. Acoust. Soc. Am. 84 (3), 917–928.

Taal, C., Hendriks, R. C., Heusdens, R., 2014. Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure. Computer Speech and Language 28 (4), 858–872.

Taal, C., Heusdens, R., 2009. A low-complexity spectro-temporal based perceptual model. In: Proc. ICASSP. pp. 153–156.

Tang, Y., Cooke, M., 2010. Energy reallocation strategies for speech enhancement in known noise conditions. In: Proc. Interspeech. Makuhari, Japan, pp. 1636–1639.

Tang, Y., Cooke, M., 2012. Optimised spectral weightings for noise-dependent speech intelligibility enhancement. In: Proc. Interspeech. Portland, US, pp. 955–958.

Tang, Y., Cooke, M., Valentini-Botinhao, C., 2016. Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech. Computer Speech and Language 35, 73–92.

Tang, Y., Cooke, M. P., 2016. Glimpse-Based metrics for predicting speech intelligibility in additive noise conditions. In: Proc. Interspeech. San Francisco, 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.