Optimised spectral weightings for noise-dependent speech intelligibility enhancement

Yan Tang[1], *Martin Cooke*[2,1]

[1]Language and Speech Laboratory, Universidad del País Vasco, Vitoria, Spain
[2]Ikerbasque (Basque Science Foundation), Bilbao, Spain

y.tang@laslab.org, m.cooke@ikerbasque.org

Abstract

Natural or synthetic speech is increasingly used in less-than-ideal listening conditions. Maximising the likelihood of correct message reception in such situations often leads to a strategy of loud and repetitive renditions of output speech. An alternative approach is to modify the speech signal in ways which increase intelligibility in noise without increasing signal level or duration. The current study focused on the design of stationary spectral modifications whose effect is to reallocate speech energy across frequency bands. Frequency band weights were selected using a genetic algorithm-based optimisation procedure, with glimpse proportion as the objective intelligibility metric, for a range of noise types and levels. As expected, a clear dependence of noise type and global signal-to-noise ratio on energy reallocation was found. One unanticipated outcome was the consistent discovery of sparse, highly-selective spectral energy weightings, particularly in high noise conditions. In a subjective test using stationary noise and competing speech maskers, listeners were able to identify significantly more words in sentences as a result of spectral weighting, with increases of up to 15 percentage points. These findings suggest that context-dependent speech output can be used to maintain intelligibility at lower sound output levels.

Index Terms: speech intelligibility, noise, optimisation, genetic algorithm, glimpse proportion

1. Introduction

Speech is frequently perceived in less-than-ideal listening conditions. While talkers have available a range of potential strategies to promote successful communication by adapting to the listener's context, speech output technology has in the past been largely insensitive to the changing needs of the listener. Recently, however, context-sensitive speech output algorithms have been proposed and evaluated [1, 2, 3, 4, 5] for both recorded and synthetic speech.

Two problems need to be solved to deliver benefits in context-sensitive speech output technology. First, speech modifications must be sought which lead to intelligibility increases. Second, the context must be known, predictable or estimated with sufficient accuracy to enable speech modification algorithms to make optimal adjustments. In [6] we evaluated a range of modification techniques which reallocated speech energy across time and frequency while preserving overall signal-to-noise ratio (SNR), and showed that modifications can produce a substantial listener benefit. However, some of the most successful modifications required detailed local noise estimates over time. In practice, it may be difficult to deliver and apply robust noise estimates at short timescales.

An intermediate approach to the use of context is to estimate noise descriptors, and to use these to select a modification which has been optimised offline. Descriptors could include low-level attributes such as level, stationarity, or temporal modulation characteristics but might also be high-level identifiers such as 'competing speech' or 'vehicle noise'. Descriptors would not require instant-by-instant noise estimates; the descriptor update rate depends on the listening situation, but might be of the order of seconds or minutes in environments such as transport interchanges or applications such as talking navigation aids.

The current study focuses on the problem of offline optimisation of speech modifications designed to promote intelligibility in the context of different noise types at a range of SNRs. In this initial study, modifications are restricted to stationary spectral reweightings under globally-constant energy and duration constraints. Spectral weights for stationary and non-stationary maskers are optimised using an objective intelligibility model followed by a subjective evaluation conducted to evaluate the effectiveness of the approach for listeners.

2. Optimisation of spectral weights

2.1. Constrained spectral weighting

Spectral energy in different frequency regions can be boosted or attenuated by applying a frequency-dependent weighting W to the speech spectrum S to produce a modified spectrum S'

$$\log|S'(f)| = \log|S(f)| + \log|W(f)| \tag{1}$$

under the constraint of constant input-output signal energy and duration:

$$\sum_{t=1}^{T'} s'(t)^2 = \sum_{t=1}^{T} s(t)^2 \tag{2}$$

$$T = T' \tag{3}$$

2.2. Objective intelligibility prediction

It is not feasible to use subjective intelligibility assessment as part of a practical optimisation procedure, which typically involves evaluating a very large number of candidates. However, objective intelligibility measures (OIMs) which make reasonable predictions for speech mixed with stationary and non-stationary maskers are available (e.g. [7, 8, 9]). These metrics can be used for the rapid prediction of the effectiveness of a population of modification candidates, and are suitable as part of a closed-loop, iterative optimisation process which seeks to

improve objective intelligibility, prior to a one-off listening test of the best candidates.

The OIM used in the current study is glimpse proportion (GP). This metric, which essentially quantifies the audibility of speech in the presence of a masker, is part of the glimpsing model of speech perception [9] and has been used to predict the intelligibility of natural [6] and synthetic speech [5], both unmodified and modified. The GP score is the percentage of spectro-temporal regions in modelled auditory excitation patterns whose local SNR exceeds a certain threshold α, expressed in decibels (dB):

$$GP = \frac{100}{TF} \sum_{t=1}^{T} \sum_{f=1}^{F} \mathcal{L}(S_{t,f} - (N_{t,f} + \alpha)) \qquad (4)$$

where T and F are the numbers of time frames and frequency channels, $S_{t,f}$ and $N_{t,f}$ denote the spectro-temporal excitation pattern in dB of speech and noise at time t and frequency f and the $\mathcal{L}(.)$ operator counts the number of 'glimpses' which meet the audibility criterion. Here, $F = 58$ frequency channels distributed uniformly in ERB-rate in the range 50-8000 Hz were used.

2.3. Genetic algorithm optimisation

To more fully explore the set of possible spectral boosting solutions while maintaining coverage of the spectrum at auditory frequency resolution – which demands the use of a large number of frequency locations (F in eqn. 4) – a genetic algorithm (GA) [10] was chosen. GAs are often used to generate solutions in complex, discontinuous or high-dimensional spaces which are not well-suited to other optimisation algorithms [11]. Here, population members are candidate spectral weightings, expressed in log magnitude units. Optimisation was performed using the MATLAB global optimisation toolbox using the parameter values listed in table 1 (default toolbox values were used for parameters not listed). To prevent excessive boosting in localised frequency regions, spectral weights were constrained.

Table 1: GA parameter settings.

constraint	boosting bounds	[-50 50] dB
population	size	300
	initial variable value	0
	elite count	30
stopping criteria	generations	500
	stall generations	5

2.4. Example spectral weighting candidates

A set of 120 sentences from the corpus described in section 3 was used during development, along with 10 maskers each at 5 SNRs (-10, -5, 0, 5, 10 dB). Maskers were: competing talker (CS), speech-modulated noise (SMN), speech-shaped noise (SSN), white noise (WN), and both high-pass noise (HP) and low-pass noise (LP) with cut-off frequencies at 500 Hz, 1000 Hz and 2000 Hz. Spectral weightings were optimised independently for each (masker, SNR) combination. In each generation of the GA, weighting candidates were evaluated in terms of their GP scores over the development corpus. Each optimisation was repeated at least once; to check for any effects of initial conditions, the -10 dB SNR condition for each of the 10 maskers was run 4 times. Figure 1 shows the best candidates for 4 separate runs of the GA in condition (CS, -10 dB). While the details vary from run to run, the overall pattern of weighting is similar, with a boost to the very low frequency region and to frequencies above 1 kHz at the expense of the 200-900 Hz part of the spectrum.

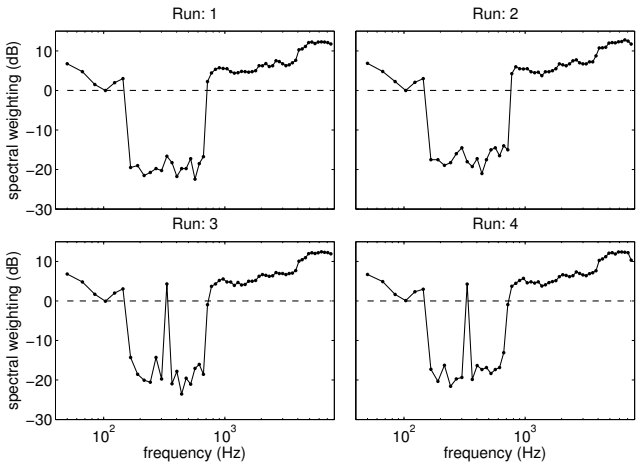

Figure 1: *Best individuals from 4 separate GA runs.*

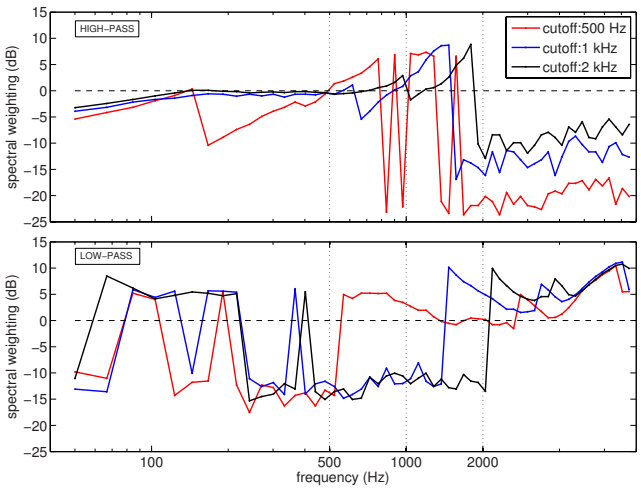

Figure 2: *Best candidate spectral weightings for high- and low-pass noise with cut-off frequencies at 500, 1000 and 2000 Hz (SNR = -10 dB)*

The dependence of the optimum weighting on masker spectrum is illustrated in figure 2 which shows optimal weightings for low and high-pass maskers with different cutoff frequencies. While the boosting pattern shows a clear sensitivity to the location of the transition band, the pattern itself shows some surprising elements: for HP noise, one might expect to see speech to be boosted across the passband, but there is only partial boosting in the case of the 500 and 1000 Hz cutoffs, and none at all for the 2 kHz case. Most of the boosting occurs in the region on the low-frequency side of the transition band. In the LP case, low frequencies are boosted, but not entirely at the expense of the high frequencies.

Further examples of the best candidates from a selection of maskers and SNRs are shown in figure 3. Two intriguing findings – seen in all noise types tested – are visible in this figure.

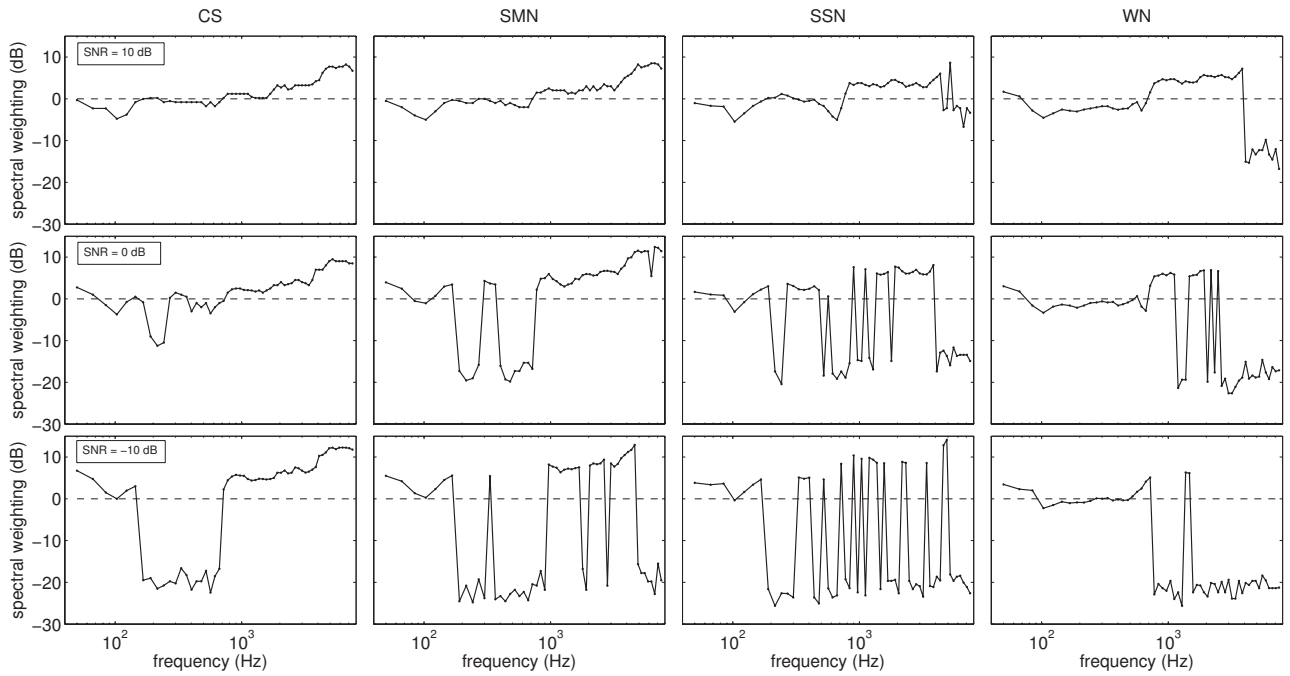

Figure 3: *GA-based optimised weightings for competing speaker (CS), speech-modulated noise (SMN), speech-shaped noise (SSN) and white noise (WN) for SNRs of 10, 0 and -10 dB.*

First, as noise conditions become more adverse, the variance of the boosting increases to the extent of becoming almost a binary choice of whether to boost or attenuate any given channel. Second, with increasing noise, boosting becomes distributed and sparse in frequency, with one or a small number of channels in any region being selected to receive the energy redistributed from neighbouring channels.

2.5. Objective intelligibility results

Table 2 shows objective intelligibility predictions based on glimpse proportion before and after boosting with the optimal spectral weighting resulting from the particular (masker, SNR) combination. In all cases, the objective intelligibility improved as a result of spectral weighting. The key test, of course, is whether these translate into gains in listening tests. The next section details the results of such an evaluation.

3. Subjective evaluation

Evaluation was based on a set of 180 sentences from the Harvard Corpus [12] read by a British English speaker. These were mixed with two maskers (CS, SSN) at three SNRs chosen in pilots to lead to word scores of 25, 50 and 75% (for CS: -7, -14, -21 dB; for SSN: +1, -4, -9 dB). The optimal spectral weighting computed additionally for every experimental SNR, was applied in each of the 6 experimental conditions. A total of 154 listeners identified the sentences for unmodified and spectrally-weighted speech. Condition orders were balanced across listeners, and the design precluded hearing the same sentence more than once. 15 listeners were removed from the study following audiometric screening: results here are based on the remaining 139 listeners.

In all conditions, frequency-weighted spectral modification led to intelligibility improvements from 4.1 to 14.7 percentage points (figure 4). Across SNRs, the gain was 10.0 percentage

Table 2: Intelligibility prediction using the glimpse proportion metric for original and modified speech.

	CS	SSN	WN	SMN	HPass500	HPass1K	HPass2K	LPass500	LPass1K	LPass2K
unmodified										
10 dB	68.0	42.2	36.6	54.7	43.6	56.7	68.8	81.4	65.8	53.3
5 dB	59.3	32.2	28.9	45.5	38.1	53.0	66.9	75.5	59.1	45.7
0 dB	50.1	22.3	22.4	36.1	34.0	50.4	64.3	68.5	51.0	38.6
-5 dB	40.7	13.5	16.2	27.4	31.1	48.8	61.4	61.7	42.3	31.1
-10 dB	31.8	7.4	10.1	20.1	28.8	47.4	56.0	55.7	34.2	25.7
frequency-weighted										
10 dB	72.7	47.1	40.1	60.0	47.7	63.0	70.2	81.4	66.5	54.8
5 dB	64.9	35.5	31.0	51.7	41.2	56.8	68.8	75.9	59.5	48.2
0 dB	56.0	25.0	23.0	41.7	35.7	52.7	66.4	69.0	51.2	40.4
-5 dB	47.6	15.4	17.4	31.3	31.8	50.0	63.2	62.3	44.5	36.6
-10 dB	37.4	8.8	11.8	22.2	32.2	49.4	58.2	58.2	39.2	31.5

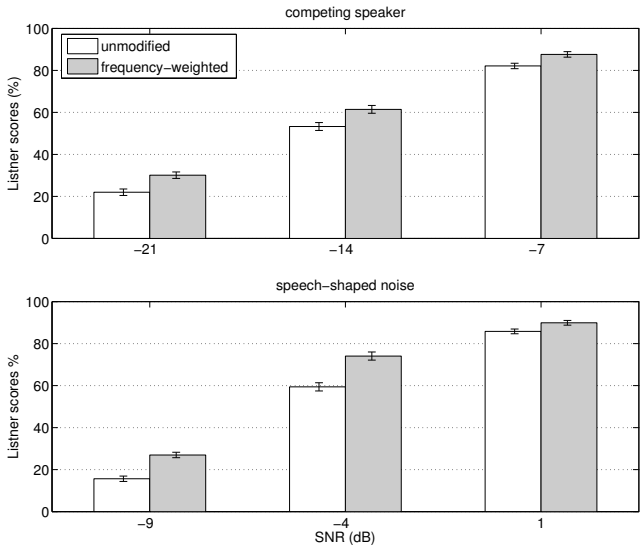

Figure 4: *Listeners' word identification scores for original and modified speech in presence of competing speaker (upper) and speech-shaped (lower) noise. Error bars indicate ± 1 standard error.*

points for SSN and slightly less (7.3) for the competing speech masker.

A three-way repeated measure ANOVA with within-subjects factors of noise type, SNR level and modification confirmed visual impressions of a significant effect of spectral weighting [$F(1, 138) = 79.13, p < .001$].

4. Discussion

Speech modified by spectral weightings learnt offline in response to specific noise types and SNRs under constant energy and duration constraints produced significant intelligibility gains for both stationary and highly-nonstationary maskers at a range of SNRs which probe the full range of word scores. Gains predicted by an objective intelligibility model used as an optimisation criterion were realised in an extensive subjective listening test.

The patterns of spectral weights discovered during optimisation were surprising. Rather than undoing the effect of the masker with weights which inverting its spectrum, spectral profiles were characterised by sparse spectral boosting which increased with noise level. We speculate that a positive effect of sparse boosting is to focus energy in a number of bands distributed across frequency where some evidence of speech will be preserved in spite of an adverse SNR. Since complete spectra are not required to recognise speech, this strategy makes effective use of the limited energy available under the constraints of this task. Sparse boosting would not be observed if a coarser spectral representation (e.g. octave band energies) were used.

Given the promising results from a rather simple technique, it is worth pursuing other types of modification (e.g. changes to the temporal modulation spectrum or harmonic structure) as well as different approaches to optimisation. Further, other objective intelligibility models will lead to different boosting patterns. Glimpsing puts a premium on audibility, but other OIMs such as STOI [8] or the Dau measure [7] which favour lack of distortion could produce different results.

Changes in spectral profile may lead to modified speech which does not meet articulatory constraints. This is not necessarily a problem in applications of speech output technology used in everyday conditions if intelligibility goals are met. Speech quality should also be evaluated, although spectral modifications tend to produce less distortion than temporal changes. Another direction for future study is the use of loudness models rather than energy preservation constraints.

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. We thank Cassie Mayo and Vasilis Karaiskos of the University of Edinburgh for their help in running data collection and listening tests.

5. References

[1] B. Sauert and P. Vary, "Near end listening enhancement: Speech intelligibility improvement in noisy environments," in *Proc. ICASSP*, Toulouse, France, 2006, pp. 1636–1639.

[2] S. D. Yoo, J. R. Boston, A. El-Jaroudi, C. Li, J. D. Durrant, K. Kovacyk, and S. S., "Speech signal modification to increase intelligibility in noisy environments," *J. Acoust. Soc. Am.*, vol. 122, no. 2, pp. 1138–1149, 2007.

[3] B. Sauert and P. Vary, "Near end listening enhancement optimized with respect to speech intelligibility index and audio power limitations," in *Proc. EUSIPCO-2010*, Aalborg, Denmark, 2010, pp. 1919–1923.

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

[5] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?" in *Proc. Interspeech*, Florence, Italy, 2011.

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

[7] C. Christiansen, M. S. Pedersen, and T. Dau, "Prediction of speech intelligibility based on an auditory preprocessing model," *Speech Comm.*, vol. 52, no. 7-8, pp. 678–692, 2010.

[8] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, Dallas, USA, 2010, pp. 4214–4217.

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

[10] J. H. Holland, *Adaptation in Natural and Artificial Systems*. University of Michigan Press, 1975, vol. Ann Arbor, no. 53.

[11] M. Mitchell, *An Introduction to Genetic Algorithms*. MIT Press, 1996.

[12] "IEEE Recommended Practice for Speech Quality Measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225 – 246, 1969.