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 \( S' \):

\[
\log |S'(f)| = \log |S(f)| + \log |W(f)|
\]

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

\[
\sum_{t=1}^{T'} s'(t)^2 = \sum_{t=1}^{T} s(t)^2
\]

\[
T' = T
\]

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 \( \alpha \), expressed in decibels (dB):

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

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 in dB of speech and noise at time \( t \) and frequency \( f \) and the \( \mathcal{L}(\cdot) \) 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.

<table>
<thead>
<tr>
<th>constraint</th>
<th>boosting bounds</th>
<th>[-50 50] dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>population size</td>
<td></td>
<td>300</td>
</tr>
<tr>
<td>initial variable</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>elite count</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>stopping criteria</td>
<td>generations</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td>stall generations</td>
<td>5</td>
</tr>
</tbody>
</table>

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.](image1)

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.
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.

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

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

5. References


