Optimal frequency filtering for speech intelligibility boosting under a constant energy constraint

Yan Tang\(^1\), Martin Cooke\(^2,1\), Petko N. Petkov\(^3\)

\(^1\)Language and Speech Laboratory, Universidad del País Vasco, Vitoria, Spain
\(^2\)Ikerbasque (Basque Science Foundation), Bilbao, Spain
\(^3\)Sound and Image Processing laboratory, School of Electrical Engineering, KTH-Royal Institute of Technology, Stockholm, Sweden

y.tang@laslab.org, m.cooke@ikerbasque.org, petkov@kth.se

Natural or synthetic speech is increasingly used in less-than-ideal listening conditions. While talkers have a range of potential strategies available 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, context-sensitive speech output algorithms have been proposed and evaluated [1, 2, 3, 4, 5] for both recorded and synthetic speech. These approaches typically modify the speech signal to maintain speech intelligibility without increasing signal level or duration. Two problems need to be solved to deliver benefits in context-sensitive speech output technology. First, speech modifications must be found which lead to intelligibility increases. Second, the acoustic context (e.g. background noise type, level) 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), demonstrating substantial listener benefits. However, the most successful modifications required detailed local noise estimates over time which may be difficult to deliver in practice. 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.

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. Frequency band weights were selected using a genetic algorithm-based optimisation procedure [7], with glimpse proportion [8] as an objective intelligibility metric, for a range of noise types (competing talker, speech-shaped, speech-modulated, high-pass, low-pass and white noise) and noise levels producing global signal-to-noise ratios in the range +10 to -10 dB. Speech and noise signals were filtered into 58 frequency bands whose bandwidths were based on auditory filters in the range 50-8000 Hz, to later give a high quality reconstructed signal. One unanticipated outcome was the consistent discovery of sparse, highly-selective spectral energy weightings, particularly in high noise conditions (e.g. figure 1). With speech modified by applying optimal spectral weightings in a subjective test, listeners were able to identify significantly more words in a modification in the presence of stationary noise and competing speech maskers, with increases of up to 15 percentage points. These findings suggest that context-dependent speech

\[ \text{SNR} = -10 \text{ dB} \]

output can be used to maintain intelligibility at lower sound output levels.

Acknowledgement: This study was supported by the LISTA Project, funded by the EU Future and Emerging Technologies programme (grant number 256230).