Boosting speech intelligibility using spectral reweighting under a constant energy constraint

Yan Tang\(^1\)  
\(^1\)Language and Speech Lab, Universidad del País Vasco, Spain  
Martin Cooke\(^2,1\)  
\(^2\)Ikerbasque (Basque Foundation for Science)

Introduction

Speech intelligibility in noise can be increased without changing overall RMS energy by reallocating energy to frequency regions which result in masking release. Different combinations of maskers and speech signals might be expected to require different spectral weights for optimal masking release. In a previous study [1], a genetic-algorithm based spectral weight optimisation procedure with a simple objective intelligibility measure led to gains of up to 15 percentage points. The current study examines the performance of weights discovered via a pattern search optimisation procedure using a more sophisticated intelligibility metric.

Spectral weighting and objective intelligibility

Spectral weighting

\[ \log|S(f)| = \log|\hat{S}(f)| + \log|\hat{W}(f)| \]

where \( S(f) \) is the original and modified speech spectrum at frequency channel \( f \), \( W \) is the weighting applied to frequency channel \( f \).

Energy and duration constraints

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

Objective intelligibility prediction

The extended glimpse proportion (xGP) [2] augments the raw GP [3] with terms representing audibility, duration, detectability and glimpse redundancy:

\[ xGP = \frac{1}{K_{SNR}} \sum_{k=1}^{K_{SNR}} \sum_{f=1}^{F} \left( S_{k,f}' \right)^{2} \left( (N_{k,f} + \alpha) \wedge (Y_{k,f} > \max(HL, \bar{Y}_{f})) \right) \]

with a compressive nonlinearity

\[ v(x) = \frac{\log(1 + x/\delta)}{\log(1 + 1/\delta)} \]

where \( K_{SNR}, K \) is the number of time frames in unmodified and modified speech, \( F \) is the number of frequency channels (54), \( H_{\text{HL}} \) is a Heaviside unit step function, \( \alpha \) is the local SNR threshold (3 dB), \( S_{k,f}' \) is the modified speech spectro-temporal excitation pattern at \( k \) and \( f \), \( N_{k,f} \) is the noise spectro-temporal excitation pattern at \( k \) and \( f \), \( Y_{k,f} \) is the cross-time average of \( S_{k,f}' \), \( HL \) is the hearing level (25 dB) and \( \delta \) is an offset to avoid log zero (0.01).

Optimised and approximated spectral weightings

**Maskers**

- speech-shaped noise (SSN)
- competing speaker (CS)
- speech-modulated noise (SMN)
- white noise (WN)
- 500 Hz high-pass filtered SSN (HP500)
- 4 kHz low-pass filtered SSN (LP4K)

**Noise- and level-dependent optimal spectral weighting (expt 1)**

- Decrease in intelligibility for the white noise masker is observed. Failure of objective intelligibility model?

**Noise- and level-independent approximated weightings (expt 2)**

- Intelligibility gains for 20-channel boosting similar to those in expt. 1
- Decreasing the number of boosted channels leads to intelligibility reduction
- Very low frequency boost – observed in optimal patterns – proved unhelpful
- Statistically: 20-chan > 10-chan = 5-chan > 5-chan-low (mainly)

Discussion

- Noise- and level-dependent optimal spectral weighting can lead to very substantial intelligibility gains.
- Noise- and level-independent spectral weighting is nearly as effective as those customised for specific maskers. These findings point to a practical mechanism for intelligibility enhancement in some common noise conditions.
- Further work will investigate boosting strategies for maskers with a uniform or high-pass characteristic, and evaluate alternative objective intelligibility measures more sensitive to frequency regions important for speech comprehension.

Evaluation

Keyword identification in phonemically-balanced Spanish sentences (Sharvard Corpus; [5])

**Expt 1** Effect of noise and level-dependent optimal weightings for 6 maskers

**Expt 2** Effect of noise and level-independent approximated weighting: tested with CS, SSN and SMN only

- 22 native Spanish listeners for each experiment
- Two SNRs, different for each masker, predicted to produce 25 and 50% keywords correct using xGP metric