A correlation metric in the envelope power spectrum domain for speech intelligibility prediction

Helia Relaño Iborra\textsuperscript{1)}, Tobias May, Johannes Zaar, Christoph Scheidiger and Torsten Dau

Hearing Systems group, Department of Electrical Engineering, Technical University of Denmark, DK-2800, Kgs. Lyngby, Denmark.

Introduction

A powerful tool to investigate speech perception is the use of speech intelligibility prediction models. Recently, a model was presented, termed correlation-based speech-based envelope power spectrum model (sEPSMcorr) \cite{1}, based on the auditory processing of the multi-resolution speech-based Envelope Power Spectrum Model (mr-sEPSM) \cite{2}, with the correlation back-end of the Short-Time Objective Intelligibility measure (STOI) \cite{3}. The sEPSMcorr can accurately predict NH data for a broad range of listening conditions, e.g., additive noise, phase jitter and ideal binary mask processing.

The sEPSMcorr model

sEPSMcorr includes audibility thresholds, such that sensitivity losses can be incorporated based on the more realistic auditory pre-processing. The CASP model \cite{4} is considered, as its front end can be tuned on the audiogram, but other types of impairment (e.g., loss of compression, reduced frequency selectivity) cannot be simulated using this framework. However, speech perception can vary greatly among listeners even when hearing sensitivity is similar. Therefore, the predictive power of the sEPSMcorr back end was further investigated in combination with a more realistic auditory pre-processing front end adopted from the computational auditory signal processing and perception model (CASP) \cite{4}.

Evaluation

- Speech mixed with stationary or non-stationary interferers:
  - Speech shaped noise (SSN), which was also used to fit the model
  - Amplitude modulated SSN (SAM) with $f_{cycles} = 8$ Hz and modulation depth of 1.
- Speech in the presence of reverberation: $T_{60} = 0, 0.4, 0.7, 1.3$ and 2.3 s
- Ideal Binary Mask processing (IBM) with four interferers.
- Speech subjected to Phase jitter distortion: $r(t) = \text{Re}(s(t)\exp^{j(\alpha t)} + \sigma \cos(\Theta(t)))$. $\Theta(t) = [0, 2\pi], \alpha = 0.8, 1.25$.

Towards realistic cochlear processing

sEPSMcorr includes audibility thresholds, such that sensitivity losses can be incorporated based on the audiogram. However, other types of impairment (e.g., loss of compression, reduced frequency selectivity) cannot be simulated using this model.

Therefore, the predictive power of the sEPSMcorr back end is further investigated in combination with a more realistic auditory pre-processing. The CASP model \cite{4} is considered, as its front end can be tuned to individual HI. CASP has been shown to successfully predict behavioral NH data obtained in conditions of, e.g., spectral masking, amplitude-modulation detection, and forward masking \cite{4} as well as individual HI results from simultaneous and forward masking and notch-noise experiments \cite{5}.

Preliminary results

- Additive noise
- SAM + sEPSMcorr
- Jittered speech

Outlook

- Investigate the model’s ability to account for individual hearing impairments, using the parameters available in the CASP framework.
- Consider additional processing stages that could account for inner hair cell loss and auditory nerve de-afferentation (Sumner et al. 2001, López-Poveda and Barrios, 2013), as they are likely to be determinant in speech-in-noise related tasks.
- Determine the conditions on which the HI model will be tested with special focus on supra-threshold distortions that might be challenging for HI subjects.

References

\begin{enumerate}
\end{enumerate}

\textsuperscript{1)} hella@elektro.dtu.dk