**DTU Library**

# Prediction of speech intelligibility based on a correlation metric in the envelope power spectrum domain

**Relano-Iborra, Helia; May, Tobias; Zaar, Johannes; Scheidiger, Christoph; Dau, Torsten**

[Link back to DTU Orbit](#)

# Prediction of speech intelligibility based on a correlation metric in the envelope power spectrum domain

Helia Relaño Iborra[a)], 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 (sEPSM[corr]) [1], based on the auditory processing of the multi-resolution speech-based Envelope Power Spectrum Model (mr-sEPSM) [2], combined with the correlation back-end of the Short-Time Objective Intelligibility measure (STOI) [3]. The sEPSM[corr] can accurately predict NH data for a broad range of listening conditions, e.g., additive noise, phase jitter and ideal binary mask processing.

The sEPSM[corr] model includes audibility thresholds, such that sensitivity loss can be incorporated based on the audiogram, but other types of hearing impairment s (HI) 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 sEPSM[corr] 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) [4]. Here, the speech-based CASP (sCASP) was evaluated in NH conditions and compared to the sEPSM[corr].
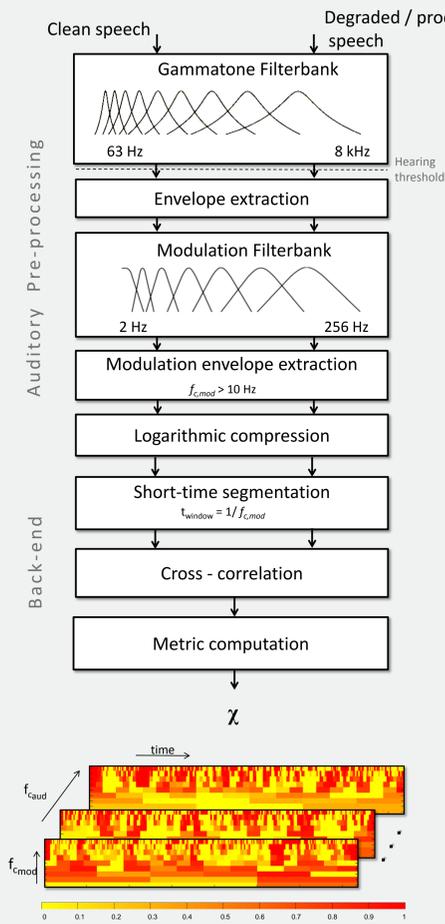
## The sEPSM[corr] model

Clean speech      Degraded / processed speech



Fig 1. sEPSM[corr] stucture, consisting of an auditory processing (left) and a correlation-based back-end (right)

Gammatone Filterbank
63 Hz    8 kHz
Hearing threshold

Envelope extraction

Modulation Filterbank
2 Hz    256 Hz

Modulation envelope extraction
$f_{c,mod} > 10$ Hz

Logarithmic compression

Short-time segmentation
$t_{window} = 1/f_{c,mod}$

Cross - correlation

Metric computation

$\chi$

Fig 2. sEPSM[corr] metric before computation

## sCASP model

Clean speech      Degraded / processed speech

Outer and middle ear TF

DRNL filterbank

Envelope extraction

Expansion

Adaptation

Modulation Filterbank

Modulation envelope extraction
$f_{c,mod} > 10$ Hz

Grouping by modulation channel
$f_{caud}$
Samples    $f_{cmod}$

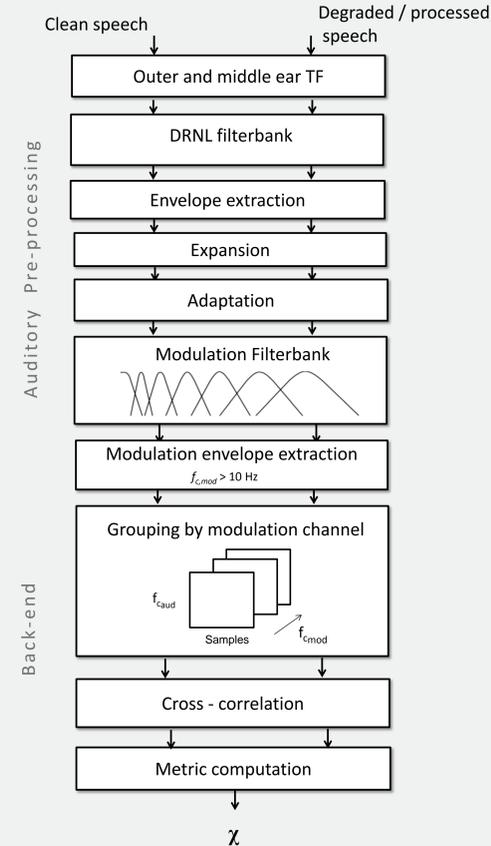Cross - correlation

Metric computation

$\chi$

Fig 3. Structure of the sCASP model

## Towards prediction of HI data

The CASP model offers more flexibility to model hearing impairments, beyond the audiogram, due to the Dual Resonance Non-Linear filterbank (DRNL), [5]. The model has been shown to account for psychoacoustic data from individual HI subjects.

Linear Gain — Gammatone filterbank — Lowpass filter

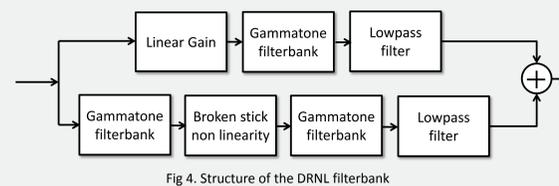Gammatone filterbank — Broken stick non linearity — Gammatone filterbank — Lowpass filter

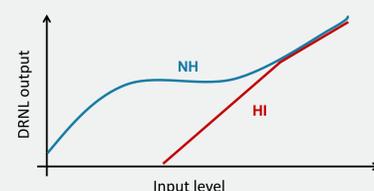Fig 4. Structure of the DRNL filterbank

Fig 5. Example diagram of NH and HI basilar membrane I/O functions

## Test conditions

The models were evaluated in conditions with:
☐ 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 $f_{c,mod}$ = 8 Hz and modulation depth of 1. and the speech like, but non-semantic international speech test signal (ISTS).
☐ Noisy speech in the presence of reverberation : $T_{60}$ = 0, 0.4, 0.7, 1.3 and 2.3 s
☐ Noisy speech subjected to different types of non-linear processing.
  • Ideal Binary Mask processing (IBM) with four interferers.

$$IBM(t,f) = \begin{cases} 1 & \text{if } SNR(t,f) > LC \\ 0 & \text{otherwise} \end{cases}$$

  • Phase Jitter distortion
  $r(t) = \text{Re}\{s(t)e^{j\Theta(t)}\} = s(t)\cos(\Theta(t))$    $\Theta(t) = [0, 2\alpha\pi], \alpha = 0:0.125:1$

Fig 6. Clean speech   (top) and speech with phase jitter distortions  of α =0.75 (bottom)

## Fitting of the models

The models are fitted per speech material to the condition of clean speech with SSN by fitting a sigmoid function between the model outputs and the human scores.

Fig 7. Diagram of model fit

## Results

☐ Human data   ▪ mr-sEPSM   ◇ STOI   ● sEPSM[corr]   ● sCASP
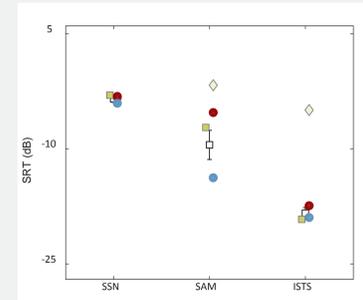
### Additive noise

Fig 8. SRT predictions for additive noises: SSN, SNN with an 8-Hz amplitude modulation  and the International Speech Test Signal. Human data from [2]. mr-sEPSM ρ = 0.99. STOI ρ = 0.54. sEPSM[corr] ρ = 0.96. sCASP ρ = 0.96.
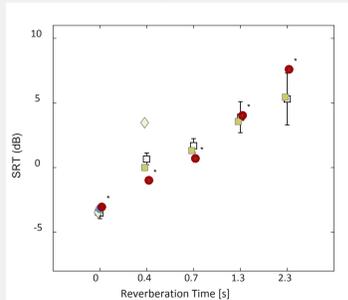
### Reverberant speech

Fig 9. SRT predictions for reverberant noisy speech reverberation. Human data from [2]. An alternative (long-term version) of sEPSM[corr] is shown. mr-sEPSM ρ = 0.99. STOI ρ = NA. sEPSM[corr, LT] ρ = 0.94. sCASP ρ =NA .
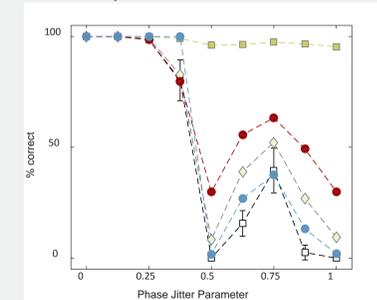
### Jittered speech

Fig 10. Intelligibility scores for noisy speech with phase jitter. Human data from [6]. mr-sEPSM: MAE  49.4%. STOI: MAE = 9.0%.  sEPSM[corr]: MAE = 17.0%. sCASP: MAE = 5.4%.

### Binary mask processing

SSN    Café    Car    Bottle

SRT - 50%  % correct

SRT - 20%  % correct

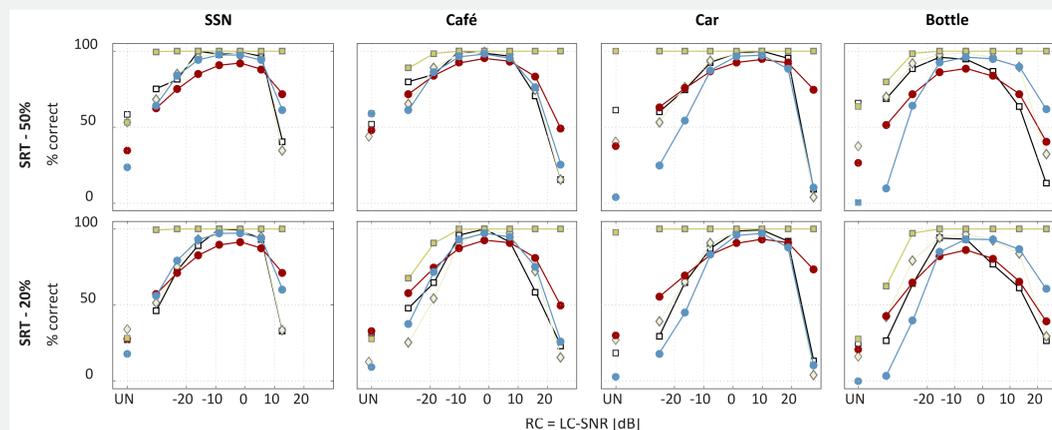UN  -20 -10  0  10  20

RC = LC-SNR [dB]

Fig 11. Intelligibility scores for IBM processed speech with four different interferers (columns) and two SNRs (rows). Human data from [3]. mr-sEPSM: ρ =0.39 STOI: ρ = 0.94 sEPSM[corr]: ρ = 0.79 sCASP: ρ = 0.86.

### Summary of results

The sCASP model  provides similar (and in some conditions better) results than the sEPSM[corr].

The model can now serve as foundation for the development of a HI model, since the DRNL-based framework allows for fitting to individual impairments.

## 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 deafferentation (Sumner et al., 2002 [8]; López-Poveda and Barrios, 2013 [9]), 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.

[1] Relaño-Iborra et al. J. Acoust. Soc. Am. 2016. 140(4):2670-2679
[5] Lopez-Poveda and Meddis. J. Acoust. Soc. Am 110.6 (2001): 3107–3118
[2] Jørgensen et al. J. Acoust. Soc. Am. 2013. 134(1):436–446.
[6] Chabot-Leclerc, et al.  J. Acoust. Soc. Am. 2014. 135(6):3502–3512.
[3] Taal et al. IEEE Trans. Audio Speech Lang. Process. 2011. 19(7):2125–2136.
[7] Jepsen and Dau. J. Acoust. Soc. Am. 2011. 129(1):262-281.
[4] Jepsen, et al. J. Acoust. Soc. Am. 2008 124(1):422-438.
[8] Sumner, et al . J. Acoust. Soc. Am. 2002. 111.5. 2178–2188.
[9] Lopéz-Poveda and Barrios. 2013. Front. Neurosci. 7(7), art.124.

[a)] heliaib@elektro.dtu.dk

Get the code!