

Technical University of Denmark



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

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

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

Citation (APA):
Iborra, H. R., May, T., Zaar, J., Scheidiger, C., & Dau, T. (2016). A correlation metric in the envelope power spectrum domain for speech intelligibility prediction. Poster session presented at ARCHES/ICANHEAR 2016, Zurich, Switzerland.

DTU Library
Technical Information Center of Denmark

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

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 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 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].

The sEPSM^{corr} model

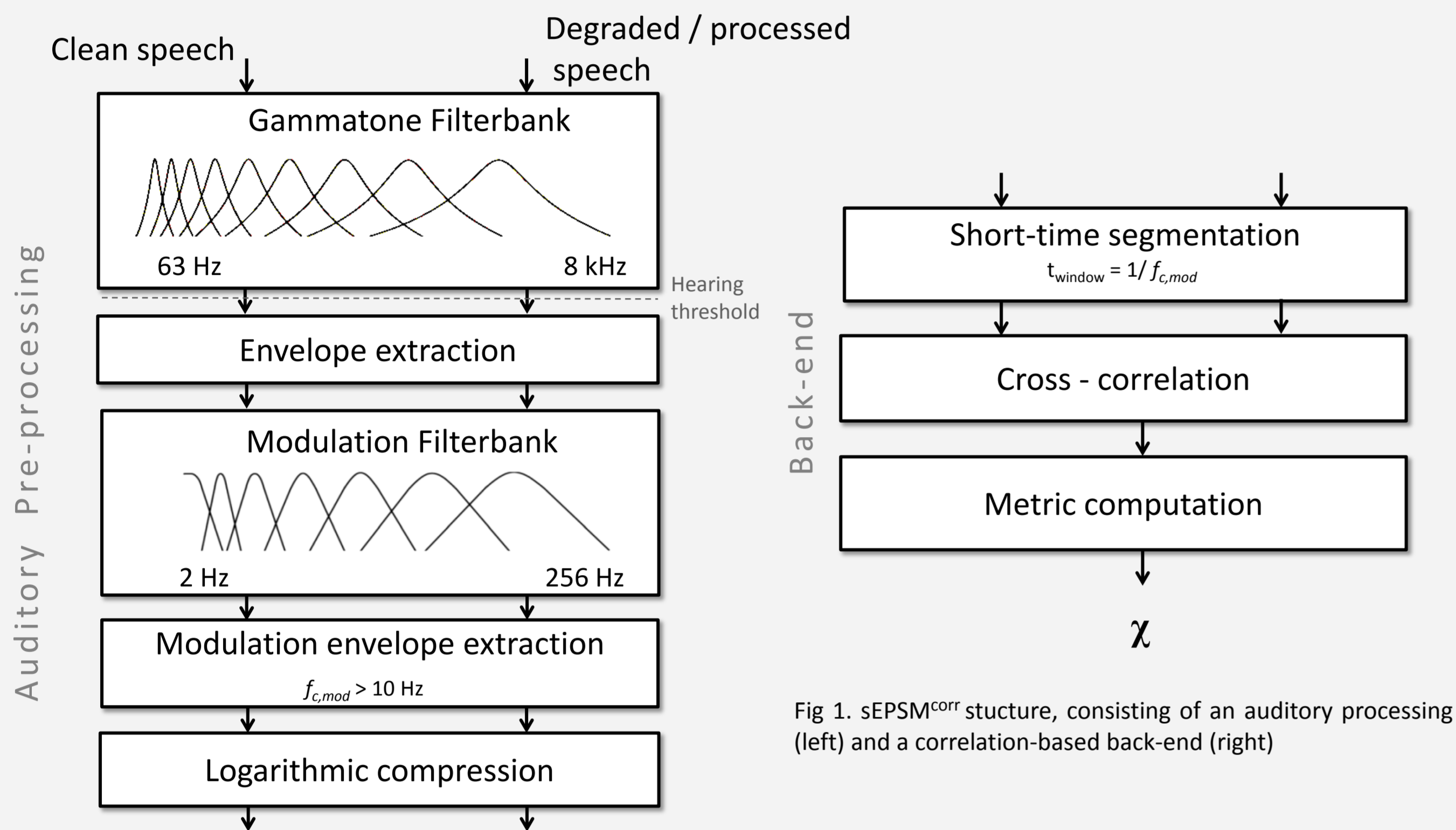


Fig 1. sEPSM^{corr} structure, consisting of an auditory processing (left) and a correlation-based back-end (right)

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 $f_{c,mod} = 8$ Hz and modulation depth of 1.
 - The speech like, but non-semantic international speech test signal (ISTS)
- 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:

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

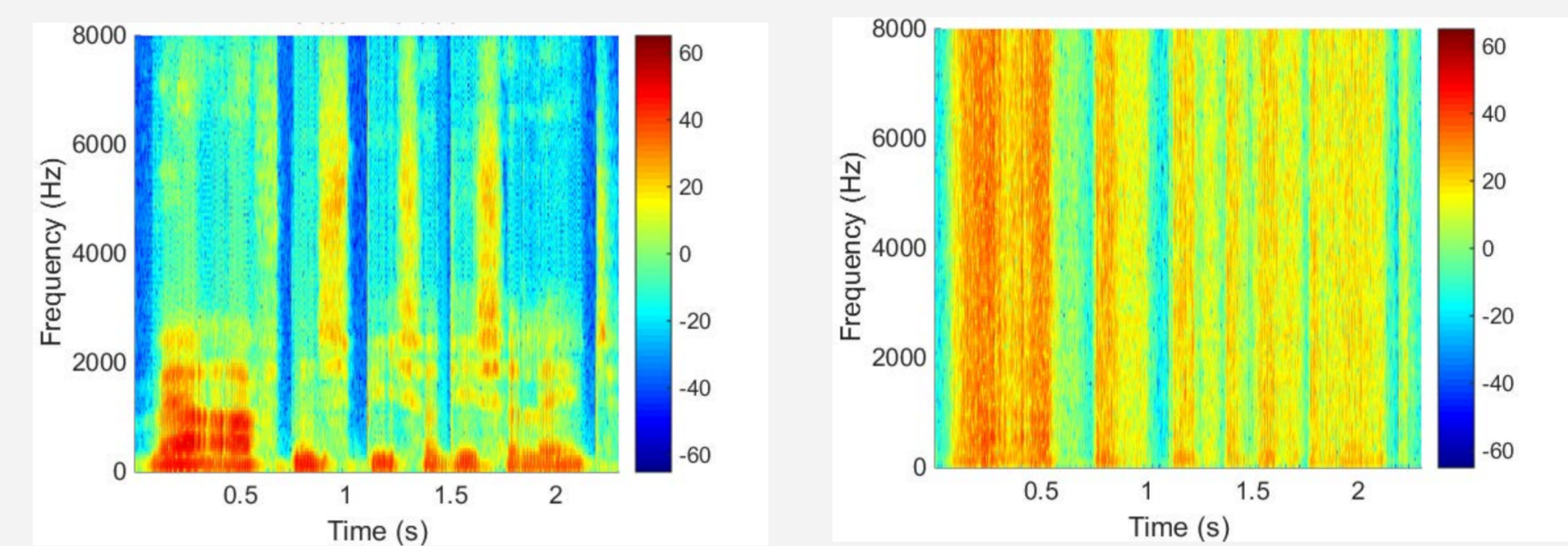


Fig 2. Clean speech (left) and speech with phase jitter distortions of $\alpha = 0.75$ (right)

Results

□ Human data ■ mr - sEPSM ◇ STOI ● sEPSM^{corr}

Additive noise

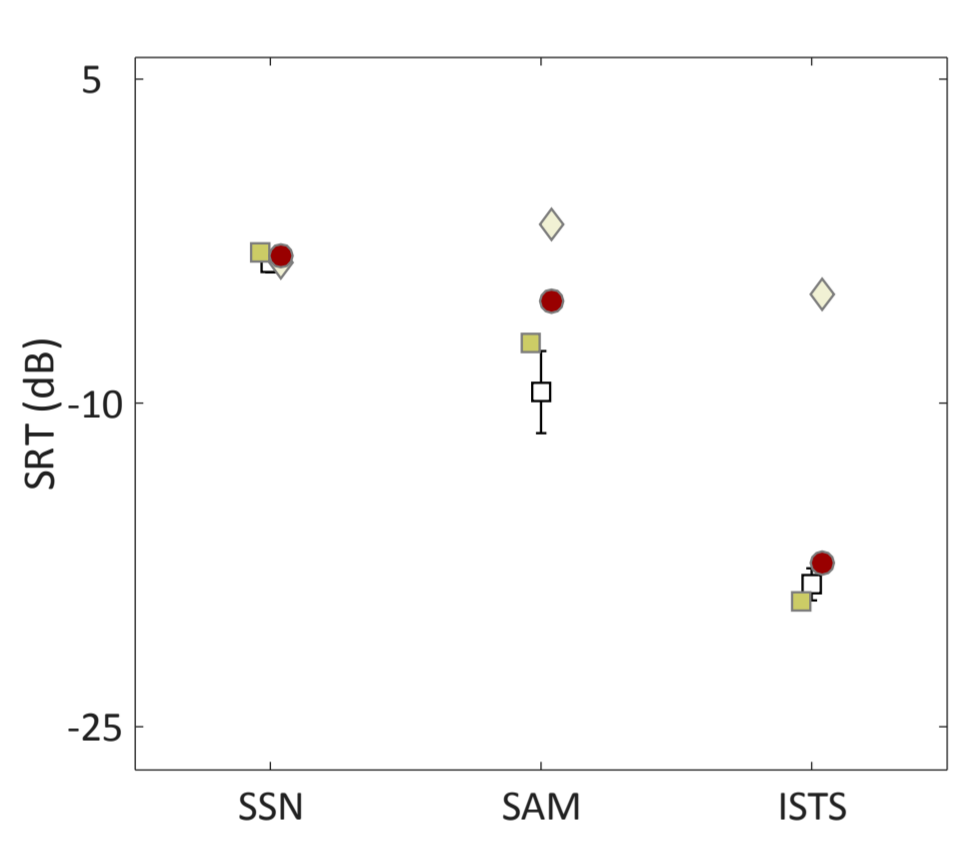


Fig 3. SRT predictions for additive noises: SSN, SNN with and 8-Hz amplitude modulation and the International Speech Test Signal. Human data from [2]. mr-sEPSM $\rho = 0.99$. STOI $\rho = 0.54$. sEPSM^{corr} $\rho = 0.97$.

Reverberant speech

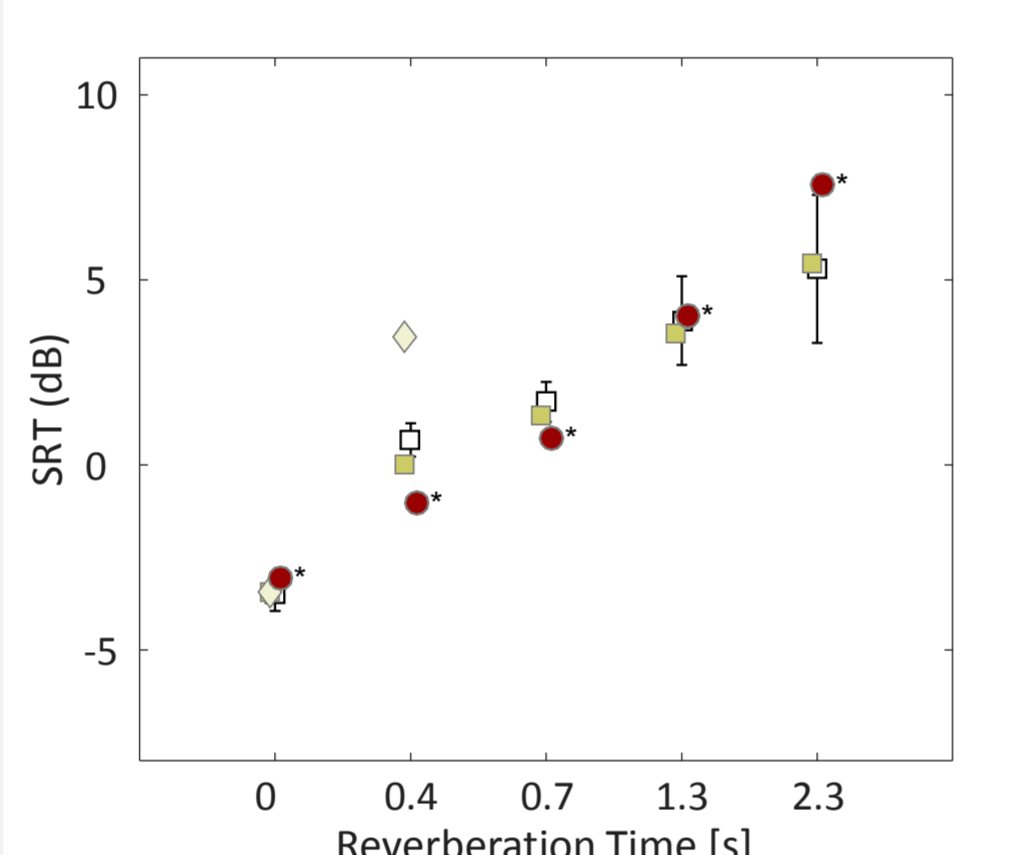


Fig 4. SRT predictions for speech under reverberation. Human data from [2]. An alternative (long-term version) of sEPSM^{corr} is shown. mr-sEPSM $\rho = 0.99$. STOI $\rho = \text{NA}$. sEPSM^{corr,LT} $\rho = 0.94$.

Jittered speech

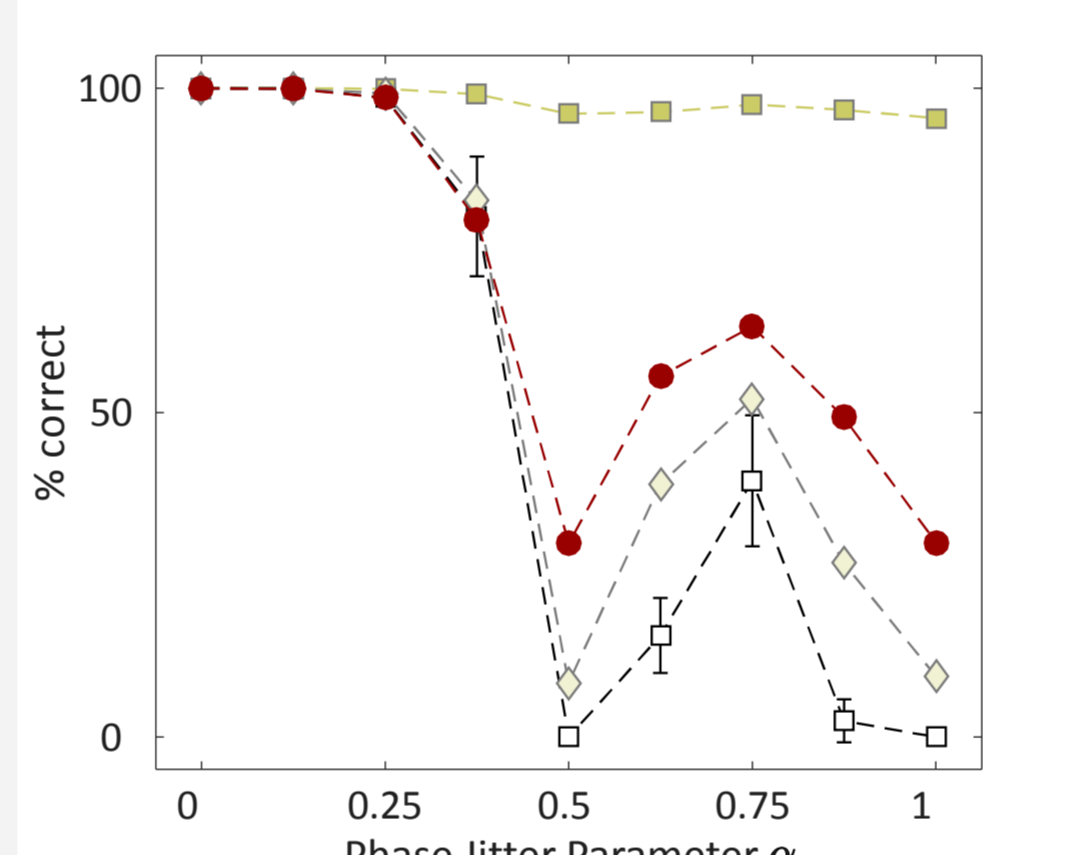


Fig 5. Intelligibility scores for noisy speech with phase jitter. Human data from [5]. mr-sEPSM: MAE = 49.4%. STOI: MAE = 9.0%. sEPSM^{corr}: MAE = 17.0%

Binary mask processing

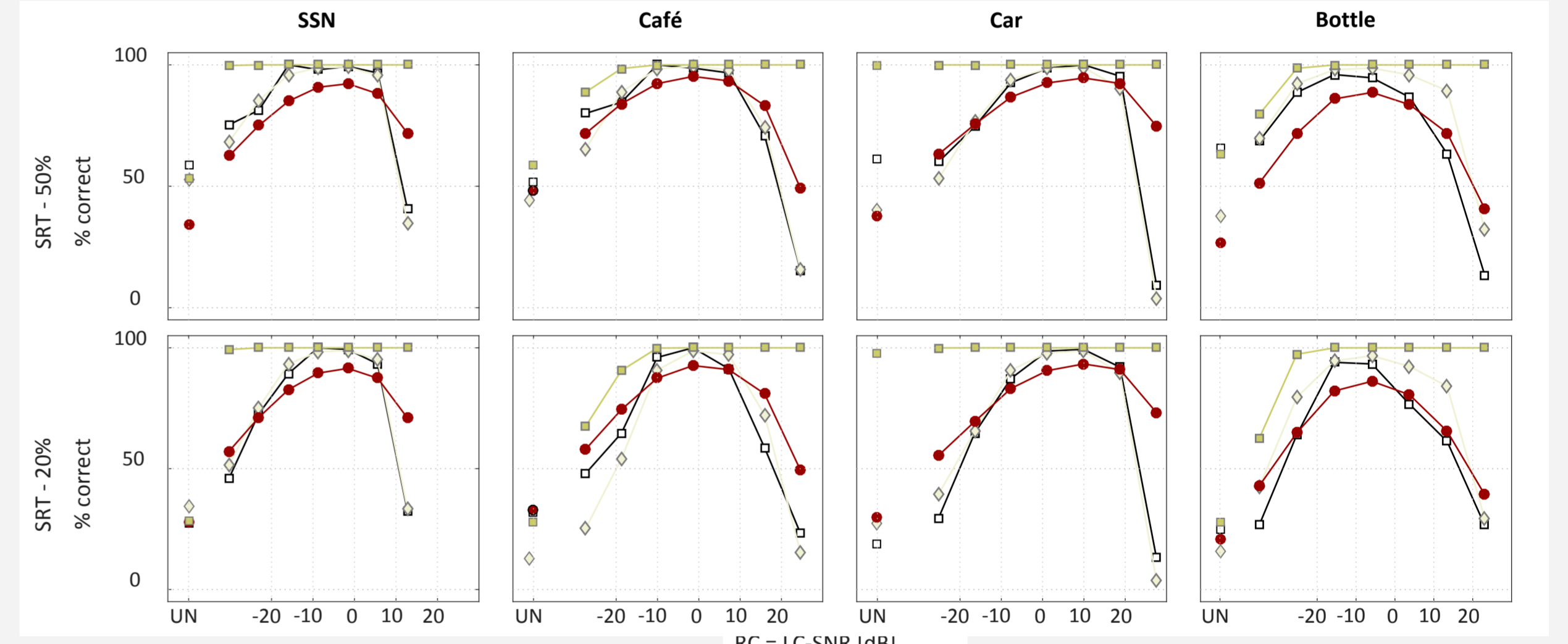


Fig 6. Intelligibility scores for IBM processed speech with four different interferers (columns) and two SNRs (rows). Human data from [2]. mr-sEPSM: $\rho = 0.39$ STOI: $\rho = 0.94$ sEPSM^{corr}: $\rho = 0.79$

Towards realistic cochlear processing

sEPSM^{corr} 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 sEPSM^{corr} back end is further investigated in combination with a more realistic auditory pre-processing. The CASP model [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 [4] as well as individual HI results from simultaneous and forward masking and notched-noise experiments [6].

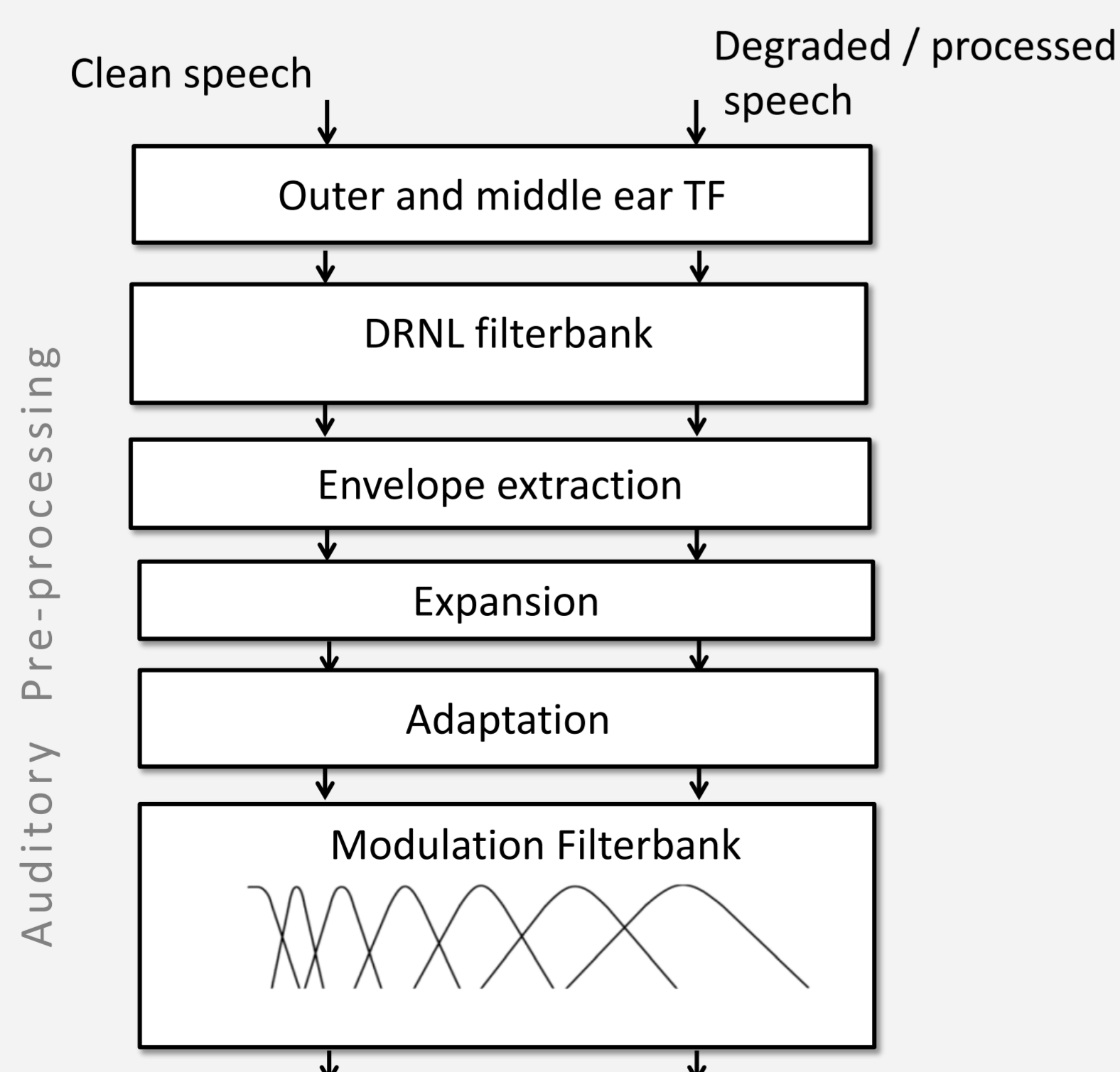


Fig 7. CASP auditory pre-processing front end.

Preliminary results

● CASP + sEPSM^{corr}

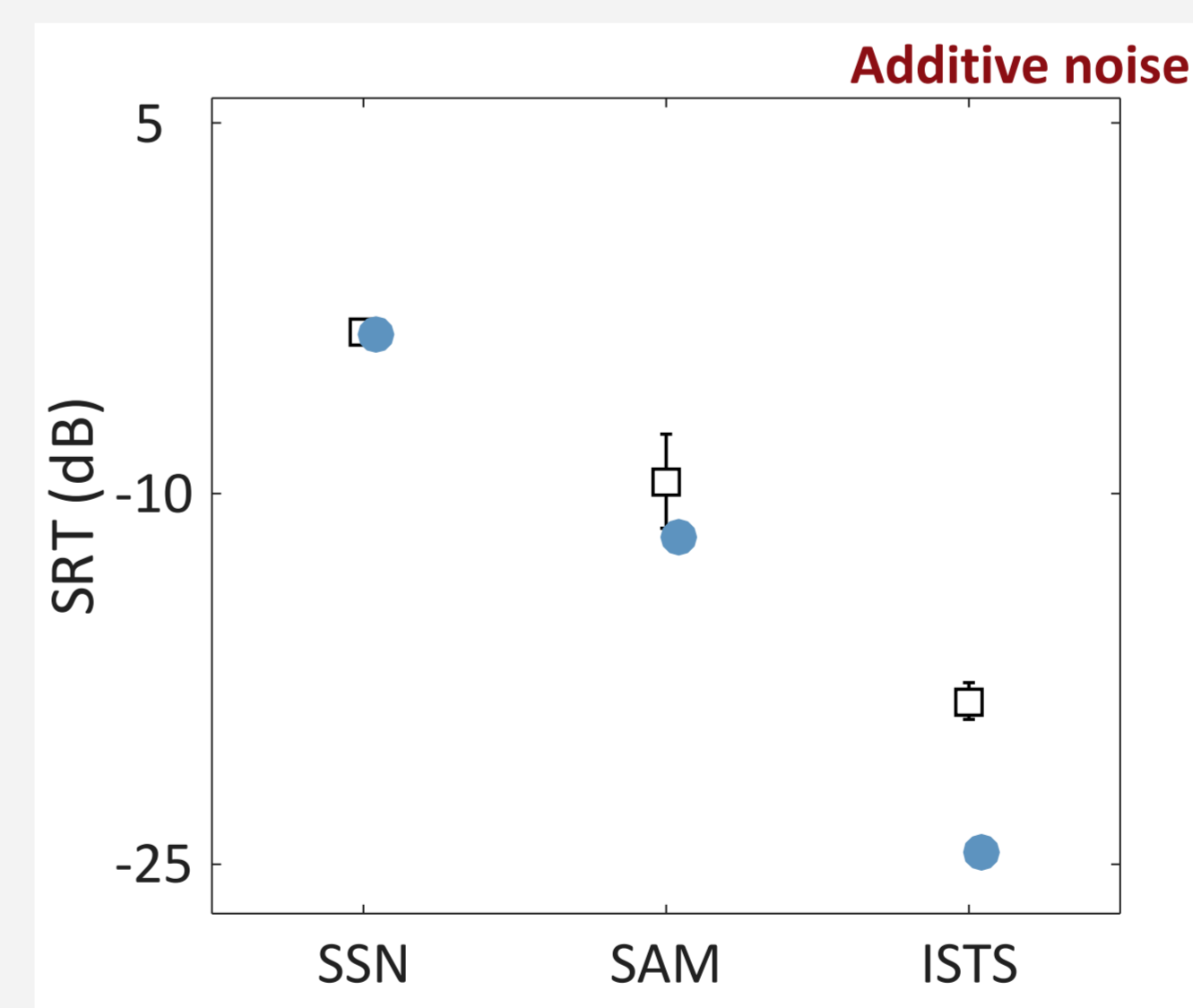


Fig 8. SRT predictions for additive noises: SSN, SNN with and 8-Hz amplitude modulation and the International Speech Test Signal. Human data from [2]. $\rho = 0.99$.

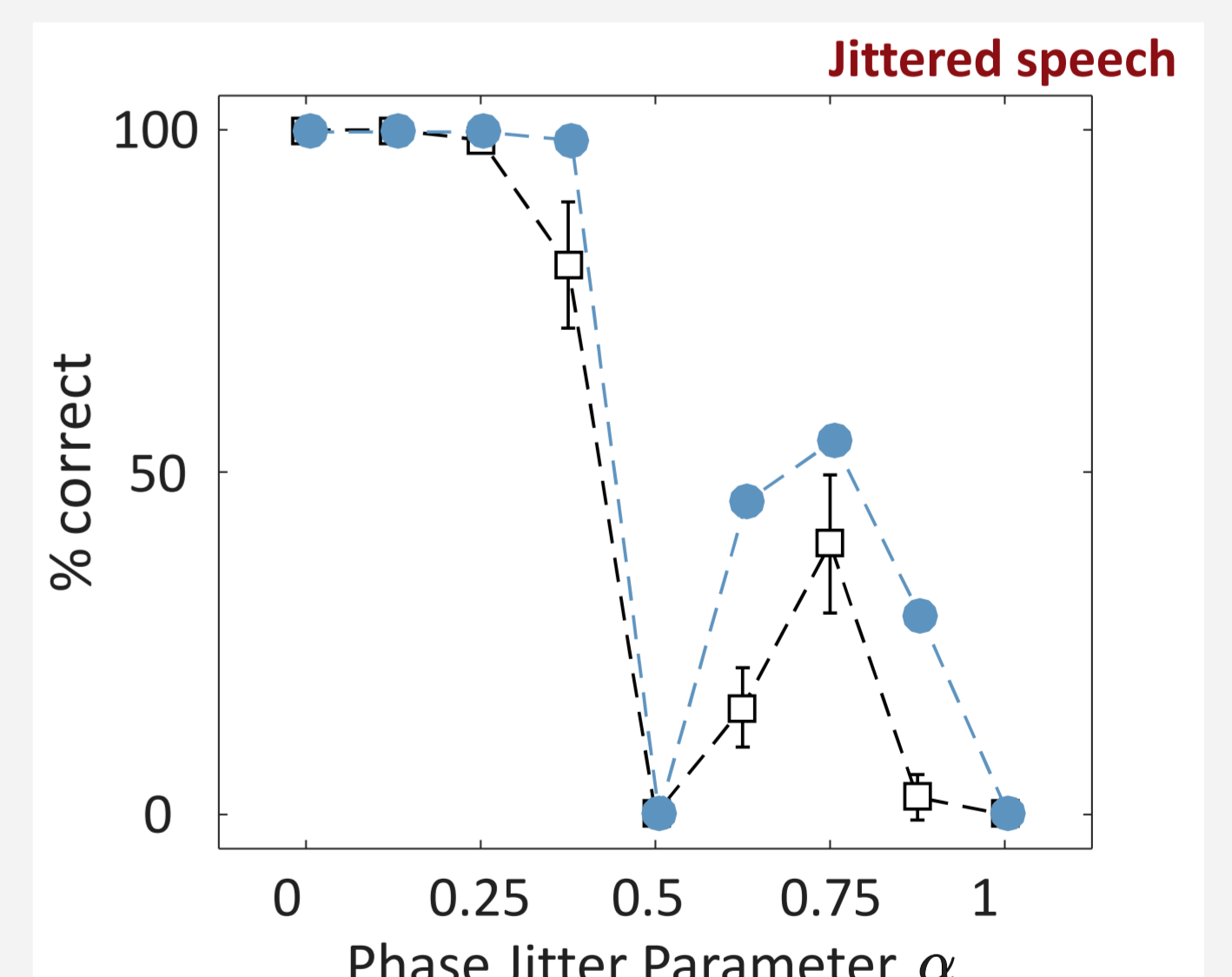


Fig 9. Intelligibility scores for noisy speech with phase jitter. Human data from [5]. MAE = 10.2 %.

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. 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

[1] Relaño-Iborra et al. J. Acoust. Soc. Amer. 2016. 140(4):2670-2679.
[4] Jepsen, et al. J. Acoust. Soc. Amer. 2008 124(1):422-438.

[2] Jørgensen et al. J. Acoust. Soc. Amer. 2013. 134(1):436-446.
[5] Chabot-Leclerc, et al. J. Acoust. Soc. Amer. 2014. 135(6):3502-12.

[3] Taal et al. IEEE Trans. Audio Speech Lang. Process. 2011. 19(7):2125-2136.
[6] Jepsen & Dau. J. Acoust. Soc. Amer. 2011. 129(1):262-28.