Predicting Speech Intelligibility based on Fluctuations in Simulated Auditory-Nerve Responses

Zaar, Johannes; Scheidiger, Christoph; Carney, Laurel H.; Dau, Torsten

Publication date:
2018

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

Predicting Speech Intelligibility based on Fluctuations in Simulated Auditory-Nerve Responses

Johannes Zaar[1,a], Christoph Scheidiger[1], Laurel H. Carney[2], and Torsten Dau[1]

[1]Hearing Systems group, Department of Electrical Engineering, Technical University of Denmark
[2]Departments of Biomedical Engineering, and Neurobiology and Anatomy, University of Rochester

INTRODUCTION

Various speech intelligibility (SI) models have been proposed to predict the ability of normal-hearing (NH) listeners to understand speech in adverse listening conditions. However, most current SI models are based on a strongly simplified linear simulation of the highly non-linear auditory periphery, which limits their ability to predict effects of hearing impairment on SI. At the same time, the models' decision stages typically interact strongly with the type of auditory front-end processing applied.

Jørgensen *et al.* (2013) proposed a powerful SI model termed "multi-resolution speech-based Envelope Power Spectrum Model" (mr-sEPSM). Using a modulation filterbank, the mr-sEPSM calculates the signal-to-noise ratio in the envelope domain (SNR_{ENV}) from the noisy speech (S+N) and to the noise alone (N) signals (see Fig. 1). It was shown to account for speech reception thresholds (SRTs) in NH listeners in a large range of acoustical conditions.

In order to also account for speech intelligibility in hearing-impaired (HI) listeners, the present study attempted to incorporate a non-linear auditory-nerve (AN) model (Zilany *et al.*, 2014; see Fig. 2) in the framework of the mr-sEPSM. Two approaches were considered:

I. **Auditory-nerve model & SNR_{ENV} back end:** The linear front end of the mr-sEPSM was replaced by the non-linear AN model, keeping the SNR_{env}-based decision process from the original mr-sEPSM.
II. **Auditory-nerve model & IC coding back end:** The same input signals and AN front end as before were used in combination with a decision process inspired by the assumption of across-CF contrast evaluation after modulation analysis in the inferior colliculus (IC; Carney *et al.*, 2015).

mr-sEPSM

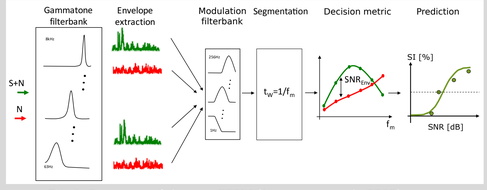


Fig. 1: Structure of the mr-sEPSM (Jørgensen *et al.*, 2013)

Auditory-nerve model

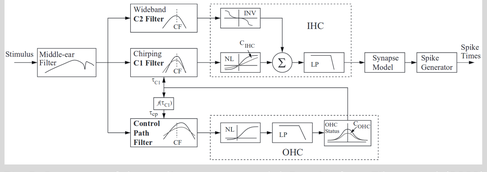


Fig. 2: Structure of the auditory-nerve model. Reprint from Zilany *et al.* (2009).

EXPERIMENTAL REFERENCE DATA

- NH reference data: SRT data obtained in **9 Danish NH listeners** (Christiansen and Dau, 2012). Speech level: 65 dB SPL.
- HI reference data: SRT data obtained in **13 Danish HI listeners** (Christiansen and Dau, 2012). Speech level: 80 dB SPL.
- Speech: Danish five-word sentences from the **CLUE corpus** (Nielsen and Dau, 2009), spoken by a male talker (F_0 = 119 Hz).
- Three additive-noise conditions using the following noise types:
  1) **SSN**: Steady-state speech-shaped noise
  2) **SAM**: Sinusoidally amplitude-modulated speech-shaped noise
  3) **ISTS**: International speech test signal (Holube *et al.*, 2010)

AUDITORY-NERVE MODEL & SNR_{ENV} BACK END – NH CONFIGURATION

The gammatone filterbank and envelope extraction stages of the original mr-sEPSM (Fig. 1) were replaced by the peristimulus time histograms (PSTH) obtained from the firing rates of the AN model at 21 CFs (log-spaced between 125 Hz and 8 kHz).

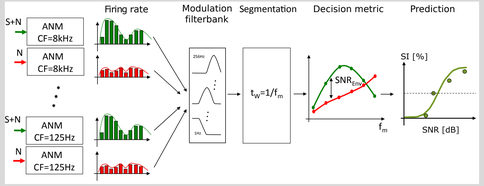


Fig. 3: Structure of the considered model

NH predictions in "linear" operation mode

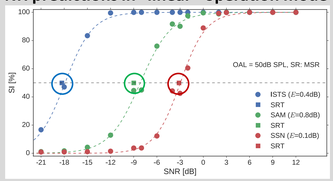


Fig. 4: Predictions obtained in NH configuration with a speech level of 50 dB SPL and only medium spontaneous rate (MSR) fibers. Fitting condition: SSN.

NH predictions in "realistic" operation mode

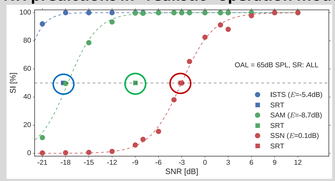
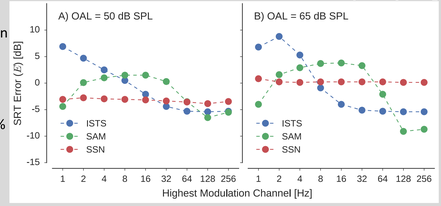


Fig. 5: Predictions obtained in NH configuration with a speech level of 65 dB SPL and all fiber types (60% HSR, 20% MSR, 20% LSR). Fitting condition: SSN.

Prediction error as function of modulation frequency range

Fig. 6: SRT prediction error $E = SRT_{pred} - SRT_{meas}$ as a function of the considered modulation frequency range. Predictions obtained with speech levels of 50 dB SPL (left) and 65 dB SPL (right) using all fiber types (60% HSR, % MSR, 20% LSR). Fitting condition: SSN @ 65 dB SPL.



AUDITORY-NERVE MODEL & IC CODING BACK END – NH AND HI CONFIGURATIONS

This alternative approach for the back end was inspired by the assumption of across-CF contrast evaluation after modulation analysis in the inferior colliculus (IC; Carney *et al.*, 2015), see Fig. 7.

Here, the simulated firing rates were used directly, omitting the spike generation. All fiber types were used (60% HSR, 20% MSR, 20% LSR). The modulation filterbank (see Fig. 3) was replaced by **one single IC filter** (i.e., a modulation bandpass filter) with a center frequency of 125 Hz. The resulting rate pattern was segmented into **20-ms time frames** *k*; the **across-CF correlation** between the noisy-speech and noise-alone representations [*sn(k,CF)* and *n(k,CF)*] was obtained in each time frame. Finally, the correlation coefficients *r(k)* were averaged across time and converted to a distance $1 - r_{avg}$.

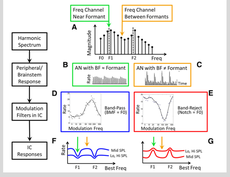


Fig. 7: Hypothesis on vowel coding in the IC. Reprint from Carney *et al.* (2015).

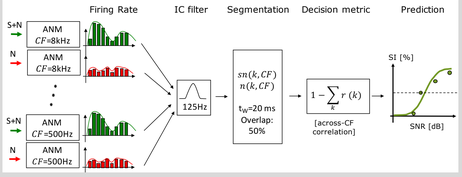


Fig. 8: Structure of the considered model framework.

NH predictions

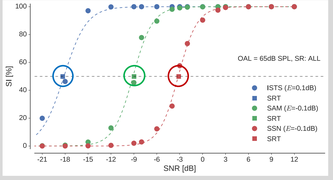


Fig. 9: NH model predictions obtained at 65 dB SPL speech level (as used in experiment). Fitting condition: SSN.

NH predictions: Level effects

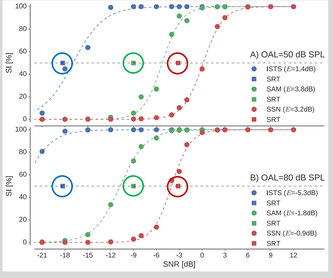


Fig. 10: NH model predictions obtained at speech levels of 50 dB SPL (top) and 80 dB SPL (bottom). Fitting condition: SSN at 65 dB SPL. Note that the behavioral SRTs were measured at 65 dB SPL.

HI prediction examples

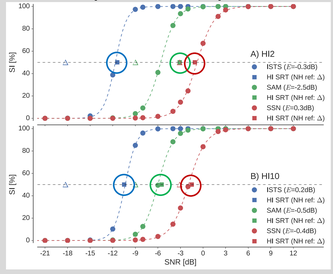


Fig. 11: Model predictions for two HI listeners. Fitting condition: SSN in NH configuration (as in Fig. 9).

Prediction error for all HI subjects

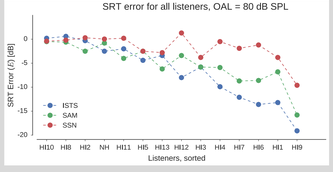


Fig. 12: SRT prediction error $E = SRT_{pred} - SRT_{meas}$ as a function of listener. Fitting condition: SSN in NH configuration (as in Fig. 9).

CONCLUSIONS

- SNR_{ENV}-based model predicted NH SRTs only when unrealistically low speech levels were used (assumption of linear pre-processing)
- Across-CF correlation between "IC-filtered" rate patterns of noisy speech and noise alone yielded
  - Accurate predictions of NH data at realistic speech levels and plausible trends for lower/higher speech levels
  - Promising results for many of the HI listeners

[a] jzaar@elektro.dtu.dk

Jørgensen et al. J. Acoust. Soc. Am. 2013. 134(1): 436-446
Zilany et al. J. Acoust. Soc. Am. 2014. 135(1): 283-286
Carney et al. eNeuro. 2015. 2(4): 1-12
Zilany et al. J. Acoust. Soc. Am. 2009. 126(5): 2390-2412
Christiansen and Dau J. Acoust. Soc. Am. 2012. 132(3): 1655-1666
Nielsen and Dau. Int. J. Audiol. 2009. 48(10): 729-741
Holube et al. Int. J. Audiol. 2010. 49(12): 891-903