Extending a computational model of auditory processing towards speech intelligibility prediction

HELIA RELAÑO-IBORRA*, JOHANNES ZAAR, AND TORSTEN DAU

Hearing Systems, Department of Electrical Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark

A speech intelligibility model is presented based on the computational auditory signal processing and perception model (CASP; Jepsen *et al.*, 2008). CASP has previously been shown to successfully predict psychoacoustic data obtained in normal-hearing (NH) listeners in a wide range of listening conditions. Moreover, CASP can be parametrized to account for data from individual hearing-impaired listeners (Jepsen and Dau, 2011). In this study, the CASP model was investigated as a predictor of speech intelligibility measured in NH listeners in conditions of additive noise, phase jitter, spectral subtraction and ideal binary mask processing.

INTRODUCTION

Computational models of the auditory system are a powerful tool to investigate the ability of humans to hear, process and encode acoustic stimuli. These models provide information about the mechanisms involved in the perception of acoustic signals. Moreover, they can provide insights about the effects of hearing loss in the impaired system. Recently, a model termed correlation-based speech-based Envelope Power Spectrum Model (sEPSM$^{\text{corr}}$; Relaño-Iborra *et al.*, 2016) was presented, which employs the auditory processing of the multi-resolution speech-based Envelope Power Spectrum Model (mr-sEPSM; Jørgensen *et al.*,2013) and combines it with the correlation back end of the Short-Time Objective Intelligibility measure (STOI; Taal *et al.*, 2011). The sEPSM$^{\text{corr}}$ was shown to accurately predict NH data for a broad range of listening conditions, e.g., additive noise, phase jitter and ideal binary mask processing. The main idea behind the sEPSM$^{\text{corr}}$ is that the correlation between the envelope representations of the clean speech and the degraded speech is a strong predictor of intelligibility. However, recent studies have shown that the mr-sEPSM preprocessing is limited with respect to predicting intelligibility data from hearing-impaired (HI) listeners (Scheidiger *et al.*, 2017). Specifically, while sensitivity loss and loss of frequency selectivity can functionally be incorporated, the crucial level-dependent effects and nonlinearities that are typically strongly affected by hearing loss cannot be successfully simulated using this framework.

The finding from the Relaño-Iborra *et al.* (2016) study that the correlation between the clean and degraded speech in the modulation power domain can be a reliable predictor of intelligibility was further investigated here using a more realistic auditory

*Corresponding author: heliaib@elektro.dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 6: Adaptive Processes in Hearing, August 2017, Nyborg, Denmark. Edited by S. Santurette, T. Dau, J. C.-Dalsgaard, L. Tranebjærg, T. Andersen, and T. Poulsen. The Danavox Jubilee Foundation, 2017.
 ISBN: 978-87-990013-6-1.

preprocessing front end. In particular, the front end of the computational auditory signal processing and perception model (CASP; Jepsen *et al.*, 2008) was considered. CASP has been shown to successfully predict psychoacoustic data of normal-hearing (NH) listeners obtained in conditions of, e.g., spectral masking, amplitude-modulation detection, and forward masking. Furthermore, the model can be adapted to account for data obtained in individual HI listeners in different behavioural experiments (Jepsen and Dau, 2011).

In this study, the CASP model was extended to investigate its potential use as a predictor of speech intelligibility data. In order to adapt CASP to function as a speech intelligibility prediction model, the speech-based CASP (sCASP) introduces modifications in the model's back-end processing and decision metric. The model was validated as a predictor of intelligibility of Danish sentences measured in NH listeners in conditions of additive noise, phase jitter, spectral subtraction and ideal binary mask processing.

THE sCASP MODEL

General structure

The proposed sCASP implementation maintains most of the structure of the original CASP model, albeit with some minor changes required due to the use of speech stimuli. The model receives the unprocessed clean speech and the noisy or degraded speech mixture as inputs (i.e., it has a-priori knowledge of the speech signal). Both inputs are processed through outer- and middle-ear filtering, a nonlinear auditory filterbank, envelope extraction, expansion, adaptation loops, a modulation filterbank, and a second-order envelope extraction for modulation channels above 10 Hz. The internal representations produced at the output of these stages are analyzed using a correlation-based back end. Figure 1 shows a diagram of the main model stages.

Modelling of the auditory preprocessing

The first stage is an outer- and middle-ear filtering stage implemented as two finite impulse response filters as in Lopez-Poveda and Meddis (2001); the output of this stage can be related to the peak velocity of vibration at the stapes as a function of frequency. Afterwards, the inputs pass through the dual-resonance nonlinear filterbank (DRNL; Lopez-Poveda and Meddis, 2001). Within this auditory filterbank, the signals are processed in two independent parallel paths, where the linear path applies a linear gain, a cascade of gammatone filters and a lowpass filter, and the nonlinear path applies a cascade of gammatone filters and a broken-stick nonlinearity followed by another cascade of gammatone filters and a lowpass filter. The summed signal of the two paths includes the effects of the nonlinear basilar-membrane processing, which accounts for level-dependent compression and auditory-filter tuning. This is followed by an envelope extraction stage, realized by half-wave rectification and second order low-pass filtering (f_c = 1 kHz). The envelopes are then expanded quadratically into an intensity-like representation. Afterwards, effects of adaptation are modelled using

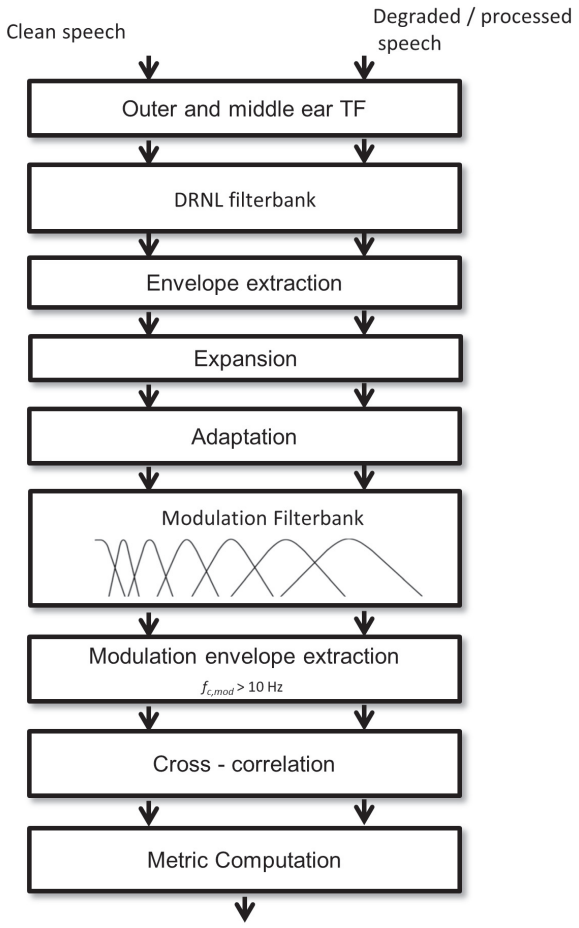

Fig. 1: Modelling stages of the speech-based CASP model.

a chain of five feedback loops (Dau *et al.*, 1996). Finally, modulation processing is included in the model using a bank of frequency-shifted first-order low-pass filters (i.e, they act as band-pass filters) in parallel with a second-order low-pass filter. For modulation filters centered below 10 Hz, the real part is considered, and for modulation filters above 10 Hz, the absolute ouput is considered, in order to account for decrease of a modulation phase sensitivity (Dau, 1996). For more details, reference is made to Jepsen *et al.* (2008) and Jepsen and Dau (2011).

Back end and decision metric

The resulting three-dimensional internal representations (as a function of time, audio frequency and modulation frequency) are analysed by cross-correlating the time signals obtained in each combination of modulation and auditory channel. The cross-correlation is performed in short time windows in a similar way as in the sEPSM$^{\text{corr}}$ model (Relaño-Iborra *et al.*, 2016), with the window length defined by the inverse of

the modulation frequency. In order to obtain a unique model output for each pair of input signals, the correlation values are averaged across time, audio-frequency and modulation channel. This differs from the calculation in the sEPSM$^{\text{corr}}$, since it does not require the summation of correlation values across time windows (Relaño-Iborra *et al.*, 2016, Eq. 3) but instead averages the correlation values across time. The sCASP back end also differs from the original CASP model, where (i) decisions are based on the correlation of the normalized difference between the internal representation of the masker plus a suprathreshold signal (considered as template) and that of the masker alone (Dau *et al.*, 1996) and (ii) no short-term processing is applied.

METHODS

Test conditions

The model was validated in four different listening conditions: speech in the presence of additive interferers, noisy speech under reverberation, phase jitter and ideal binary mask (IBM) processing. For the latter, the Dantale II corpus (Wagener *et al.*, 2003) was used, and for all other conditions the CLUE corpus was used (Nielsen and Dau, 2009).

Three additive noises were considered for the first experiment: (i) speech-shaped noise ('SSN'), (ii) an 8-Hz sinusoidally amplitude-modulated SSN with a modulation depth of 1 ('SAM'), and (iii) the speechlike but non-semantic, international speech test signal ('ISTS'; Holube *et al.*, 2010). Signal-to-noise ratios (SNRs) ranging from -27 to 3 dB with a step size of 3 dB were used. Model predictions were compared to human data obtained under the same conditions by Jørgensen *et al.* (2013).

Phase jitter was applied to sentences mixed with SSN at a fixed SNR of 5 dB as follows:

$$r(t) = \mathrm{Re}\{s(t)e^{j\Theta(t)}\} = s(t)\cos(\Theta(t))) \qquad \text{(Eq. 1)}$$

where $s(t)$ represents the non-processed mixture, $r(t)$ the resulting jittered stimulus and $\Theta(t)$ denotes a random process with a uniform probability distribution between $[0, 2\alpha\pi]$ with α ranging between 0 and 1 (Elhilali *et al.*, 2003). The simulations were compared to the data obtained in Chabot-Leclerc *et al.* (2014).

For the spectral subtraction experiment, the sentences were mixed with SSN at SNRs from -9 to 9 dB, in 3 dB steps. Spectral subtraction was applied to each mixture following:

$$\widehat{S(f)} = \sqrt{P_Y(f) - \kappa\widehat{P_N}(f)} \qquad \text{(Eq. 2)}$$

where $\widehat{S}$ is the enhanced magnitude spectrum of the noisy mixture after spectral subtraction. $\widehat{P_N}$ and P_Y are the averaged power spectra of the noise alone and the original speech-plus-noise mixture, respectively. Values for the over-subtraction factor, κ, of 0, 0.5, 1, 2, 4, and 8 were considered. The simulations were compared with data collected by Jørgensen and Dau (2011).

Finally, IBMs where applied to two SNR mixtures(corresponding to 20% and 50% understanding) of Dantale II sentences with four different interferers: SSN, car-cabin noise ('Car'), noise produced by bottles on a conveyor belt ('Bottle'), and two people speaking in a cafeteria ('Café'). Different IBMs were built as follows:

$$\mathrm{IBM}(t,f) = \begin{cases} 1 & \text{if } \mathrm{SNR}(t,f) > \mathrm{LC} \\ 0 & \text{otherwise} \end{cases} \qquad \text{(Eq. 3)}$$

where LC corresponds to the local criterion, which defines the density of the mask. Eight different values of LC were considered, and discussed here in terms of the relative criterion defined as $RC = LC - RC$ as in the reference study of Kjems *et al.* (2009).

Model fitting

The correlation-based output of the proposed model is monotonically related to the SNR of the input mixture. In order to convert the model output to intelligibility scores, a fitting condition is required. The transformation is performed by applying a logistic function to the model outcome χ:

$$\Phi(\chi) = \frac{100}{1 + e^{a\chi + b}} \qquad \text{(Eq. 4)}$$

where a and b are free parameters adjusted to map the model output to intelligibility scores in the fitting condition. The model was calibrated twice in the present study, once per speech material, using SSN at different SNRs. Thus, the mapping accounts for the intelligibility of the speech material but implies no a-priori knowledge about the degradations tested, other than the degradation induced by the SSN.

RESULTS AND DISCUSSION

Figure 2 shows the human data, in open squares, and the corresponding model predictions, in gray circles, for the four conditions under consideration. The accuracy of the model predictions was measured in terms of their Pearson's correlation and mean average error (MAE) with the human data.

The model can account for the changes in intelligibility reported by the listeners for speech in the presence of different additive noises, as seen in panel (a), with $\rho = 0.99$ and MAE $= 1.5$ dB. Regarding the non-linear conditions, i.e., spectral subtraction, phase jitter and IBM (panels b, c and d), the model can account fairly well for the data with $\rho = 0.78$ and MAE $= 1.2$ dB, $\rho = 0.96$ and MAE $= 8.5\%$ and $\rho = 0.78$ and MAE $= 13\%$, respectively.

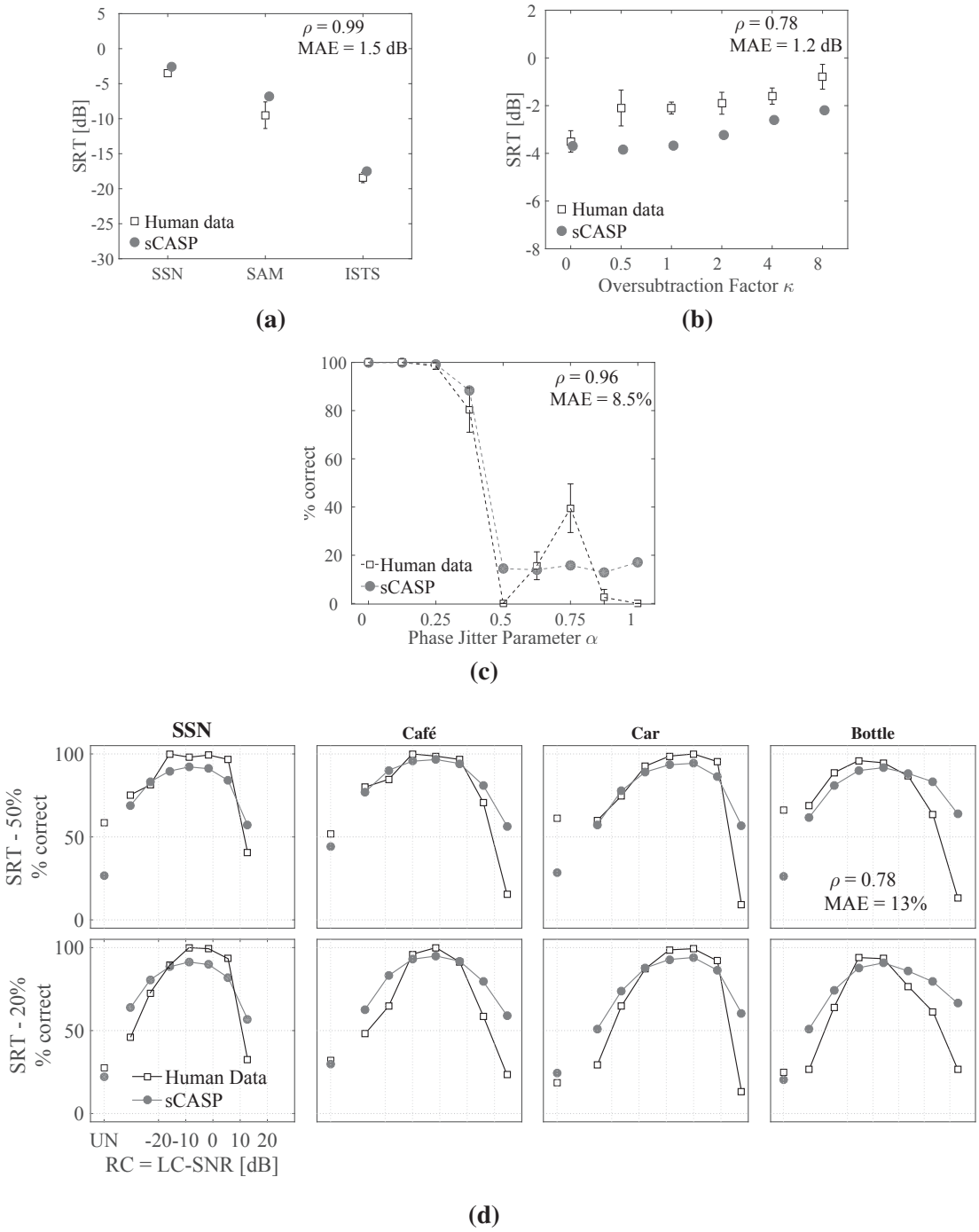

Fig. 2: Panels (a) - (d) show human data (open squares) and model predictions (gray circles) for conditions of speech in the presence of additive noise (a), noisy speech with spectral substraction processing (b), noisy speech distorted with phase jitter (c) and IBM-processed noisy speech (d). Human data from Jørgensen *et al.* (2013), Jørgensen and Dau (2011), Chabot-Leclerc *et al.* (2014) and Kjems *et al.* (2009), respectively.

However, in the phase jitter condition (panel c), it can be observed that the model exhibits some flooring effects, such that it does not predict intelligibility scores below 15% and underestimates the recovery of intelligibility reported by listeners for $\alpha = 0.75$. This might be an indication of the need for an across-frequency analysis as suggested by Chabot-Leclerc *et al.* (2014). Furthermore, although not shown here, the sCASP model does not account for effects of reverberation on speech intelligibility (as also reported for the sEPSM$^{\text{corr}}$).

Overall, these results are very similar to those reported in Relaño-Iborra *et al.* (2016) (see Table 1), despite the changes both in the front-end and the back-end processing. It thus appears that the pre-processing of CASP, which includes an adaptation stage that emphasizes the higher-frequency envelope content, does not require the accumulation process used in the sEPSM$^{\text{corr}}$ back end in order to replicate its performance (i.e., a linear average of the correlation values across time windows suffices). Still, the main finding of Relaño-Iborra *et al.* (2016) holds, namely that the correlation in the modulation domain can account for speech intelligibility.

	sCASP		**sEPSM$^{\text{corr}}$**	
	ρ	MAE	ρ	MAE
Additive Interferers	0.99	1.5 dB	0.97	1.85 dB
Spectral Subtraction	0.78	1.2 dB	0.82	0.6 dB
Phase Jitter	0.96	8.5%	0.97	19.0%
Ideal Binary Mask Processing	0.78	13%	0.79	12.1 %

Table 1: Comparison of the accuracy of the predictions for the proposed sCASP model and the referenced sEPSM$^{\text{corr}}$ (Relaño-Iborra *et al.*, 2016). ρ denotes the Pearson's correlation between human data and model predictions and MAE stands for mean average error.

CONCLUSION

The sCASP model shows promising results in terms of predicting NH intelligibility in a wide range of listening conditions. Combined with the original CASP model's ability to account for individual HI psychoacoustic data, this provides a strong basis for a framework investigating consequences of hearing loss on speech intelligibility.

ACKNOWLEDGEMENTS

This research was supported by the Oticon Centre of Excellence for Hearing and Speech Sciences (CHeSS).

REFERENCES

Chabot-Leclerc, A., Jørgensen, S., and Dau, T. (**2014**). "The role of auditory spectro-temporal modulation filtering and the decision metric for speech intelligibility

prediction," J. Acoust. Soc. Am., **135**, 3502-3512.

Dau, T. (**1996**). *Modeling Auditory Processing of Amplitude Modulation.* Doctoral dissertation. Retrieved from Bibliotheks- und Informationssystem der Universität Oldenburg.

Dau, T., Püschel, D., and Kohlrausch, A. (**1996**). "A quantitative model of the effective signal processing in the auditory system. I. Model structure," J. Acoust. Soc. Am., **99**, 3615-3622.

Elhilali, M., Chi, T., and Shamma, S.A. (**2003**). "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," Speech Commun., **41**, 331-348.

Holube, I., Fredelake, S., Vlaming, M., and Kollmeier, B. (**2010**)."Development and analysis of an International Speech Test Signal (ISTS)," Int. J. Audiol., **49**, 891-903.

Jepsen, M.L., Ewert, S.D., and Dau, T. (**2008**). "A computational model of human auditory signal processing and perception," J. Acoust. Soc. Am., **124**, 422-438.

Jepsen, M.L., and Dau T. (**2011**) "Characterizing auditory processing and perception in individual listeners with sensorineural hearing loss," J. Acoust. Soc. Am., **129**, 262-281.

Jørgensen, S., and Dau, T. (**2011**). "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," J. Acoust. Soc. Am., **130**, 1475-1487.

Jørgensen, S., Ewert, S.D., and Dau, T. (**2013**). "A multi-resolution envelope-power based model for speech intelligibility," J. Acoust. Soc. Am., **134**, 436-446.

Kjems, U., Boldt, J.B., Pedersen, M.S., Lunner, T., and Wang, D.L. (**2009**)."Role of mask pattern in intelligibility of ideal binary-masked noisy speech," J. Acoust. Soc. Am. , **126**, 1415-1426.

Lopez-Poveda, E.A., and Meddis, R. (**2001**). "A human nonlinear cochlear filterbank," J. Acoust. Soc. Am., **110**, 3107-3118.

Nielsen, J.B., and Dau, T. (**2009**). "Development of a Danish speech intelligibility test," Int. J. Audiol., **48**, 729-741.

Relaño-Iborra, H., May, T., Zaar, J., Scheidiger, C., and Dau, T. (**2016**). "Predicting speech intelligibility based on a correlation metric in the envelope power spectrum domain," J. Acoust. Soc. Am., **140**, 2670-2679.

Scheidiger, C., Zaar, J., Swaminathan, J., and Dau, T. (**2017**). "Modeling speech intelligibility based on neural envelopes derived from auditory nerve spike trains". Association for Research in Otolaryngology Mid-Winter Meeting, Baltimore.

Taal, C.H., Hendriks, R.C., Heusdens, R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio Speech Lang. Process., **19**, 2125-2136.

Wagener, K., Josvassen, J.L., and Ardenkjaer, R. (**2003**)."Design, optimization and evaluation of a Danish sentence test in noise," Int. J. Audiol.,**42**, 10-17.

Zilany, M.S.A., and Bruce, I.C. (**2007**). "Predictions of speech intelligibility with a model of the normal and impaired auditory-periphery," 3rd Int. IEEE/EMBS Conf. Neural Eng., **1-2**, 481.