Supplementary materials for

A microscopic investigation of the effect of random envelope fluctuations on phoneme-in-noise perception

Alejandro Osses and Léo Varnet

Laboratoire des systèmes perceptifs, Département d'études cognitives, École normale supérieure, PSL University, CNRS, 75005 Paris, France

(Dated: 13 November 2023)

The information contained in this document is supplementary to our main study, that has the same title. As a consequence, the explanations in these supplementary materials may not be self explanatory. In such cases, the reader is referred to the main text. Additionally, the raw data and post-processed data for our study can be either recreated using the fastACI toolbox (Osses and Varnet, 2023) or can be retrieved online (Osses and Varnet, 2022b). All figures from the main paper and from these supplementary materials can be recreated using fastACI.

Pages: 1–6

I. PARTICIPANTS' DETAILS

Twelve participants took part in our study aged between 22 and 43 years old. Further details of the participants are given in Table I. We characterized their hearing status by measuring audiometric thresholds and their performance in a speech-in-noise test, whose details are given next. While the hearing thresholds were used as the only inclusion criterion, the speech-in-noise thresholds were planned to give an indication of the participants' supra-threshold hearing status, to be used as referential data for the design of future studies.

TABLE I. Participants' details. The age is expressed in years at the time of testing. Participants S06 and S12 were the two last participants to complete the experimental sessions. Their data were excluded in Sec. III B for the analysis with the preregistered number of $N = 10$.

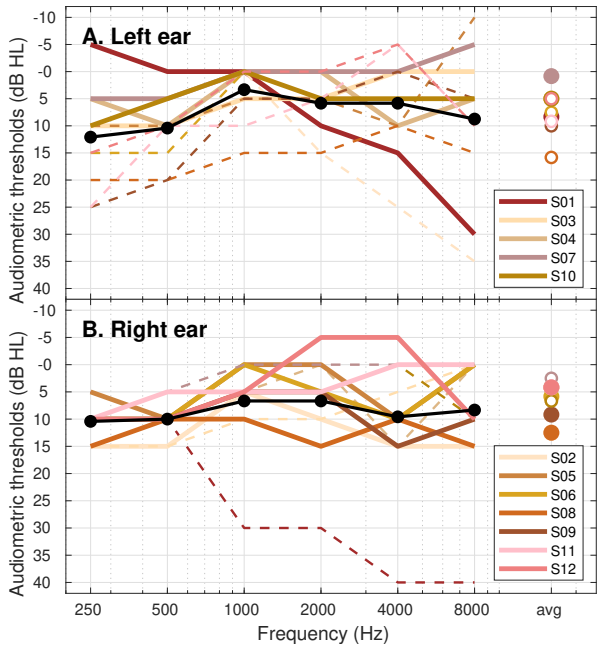
Subject	Age	Gender	Mother tongue	Speaks French
S01	33	M	French	Yes
S02	36	M	Spanish	No
S03	31	F	French	Yes
S04	38	M	French	Yes
S05	24	F	Italian	No
S06	43	M	French	Yes
S07	23	M	French	Yes
S08	27	F	Turkish	Yes
S09	25	M	French	Yes
S10	22	F	French	Yes
S11	36	M	Spanish	No
S12	22	M	French	Yes

A. Audiometric thresholds

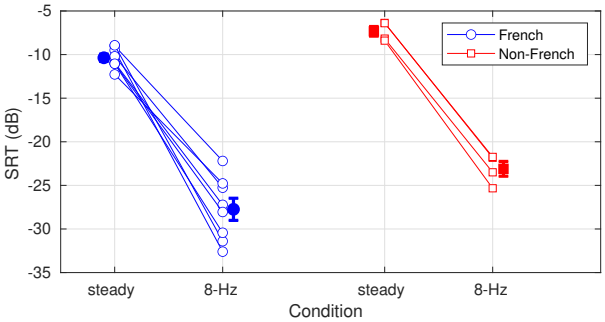
Audibility thresholds were measured using pure-tone audiometry at six frequencies (250, 500, 1000, 2000, 4000, and 8000 Hz) and had average thresholds between 0.8 (S07) and 12.5 dB HL (S02) in their best ear, meeting our inclusion criterion of having thresholds of 20 dB HL or better. The obtained hearing thresholds are shown in Suppl. Fig. 1.

B. Intellitest

The participants' supra-threshold hearing status was measured using the Intellitest speech-in-noise test (Gnansia *et al.*, 2014). The Intellitest is a closed-set speech material of 16 words of the structure VCVCV containing three takes of each word (total of 48 samples). The dataset was split into three single lists of 16 non-repeated words. Three single lists were evaluated twice, either using a speech-shaped noise (SSN), or using an 8-Hz amplitude modulated version of the SSN. In these experiments the speech level was adjusted targeting a 50% score. The threshold estimate for each noise condition obtained from the median of the three single list runs in each condition are shown in Suppl. Fig. 2. In this figure we grouped the participants into native French speakers ($N = 8$, blue traces, "French") and the rest of the participants ($N = 4$, red traces, "Non-French"). The speech reception thresholds (SRTs) for the French group had median thresholds of -10.4 and -27.7 dB in the steady-noise and 8-Hz AM noise, respectively. The threshold using the modulated masker was 17.3 dB lower (better) than the threshold in the steady-noise condition. The results for non-French group were -7.3 and -23.1 dB in the steady-noise and 8-Hz AM noise, respectively. These thresholds were higher (worse) than the thresholds obtained for the French speakers, by 3.1 and 4.6 dB for the two noise conditions. The difference between performance in steady and modulated noise was 15.8 dB.



SUPPL. FIG. 1. (Color online) Audiograms for all participants. Left and right ear thresholds are shown in Panels **A** and **B**, respectively. The participant's best-ear thresholds are connected by continuous traces, and the subject ID is indicated in the corresponding panel legend. Average thresholds across participants are indicated by the black traces and the average audiometric threshold for all tested frequencies between 250 and 8000 Hz are indicated by the right-most markers (filled symbols are used when those averages are from the participant's best ear).



SUPPL. FIG. 2. (Color online) Results for the evaluation of the Intellitest speech material using a steady SSN background noise or an 8-Hz 100% amplitude-modulated version of it. We present separately the results for French speakers (blue) and non-French speakers (red). The filled symbols indicate the group mean thresholds and the error bars represent one SEM.

II. RETRIEVING THE SOUND STIMULI

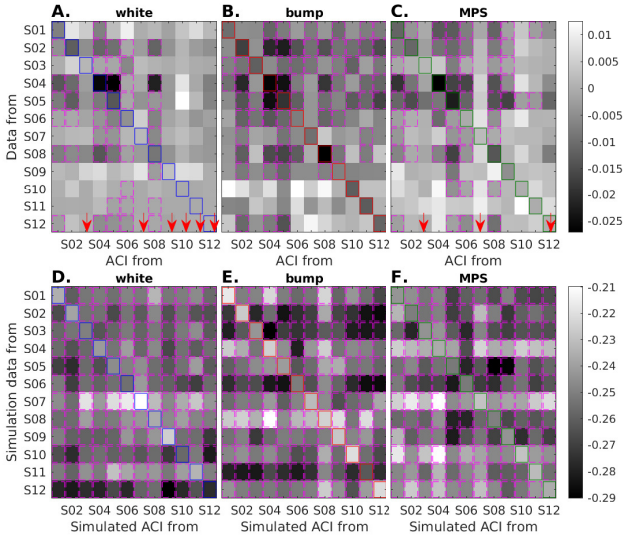
The thirty-six sets of noises and the two speech samples (/aba/ and /ada/) used in the experiments can be retrieved either from Zenodo (Osses and Varnet, 2022b) or using our in-house fastACI toolbox. To retrieve the sounds using the toolbox, the script `publ_osses2022b_preregistration_0_init_participants.m` needs to be run. Note that for recreating the MPS noises, the PhaseRet toolbox (Průša, 2017) needs to be installed and compiled. No extra dependencies are required to reproduce the white and bump noises.

Once generated, the noises will be stored in separate folders named NoiseStim-white, NoiseStim-bump, and NoiseStim-MPS, each of them containing 4000 waveforms using a numbered labeling (`Noise_00001.wav`–`Noise_04000.wav`).

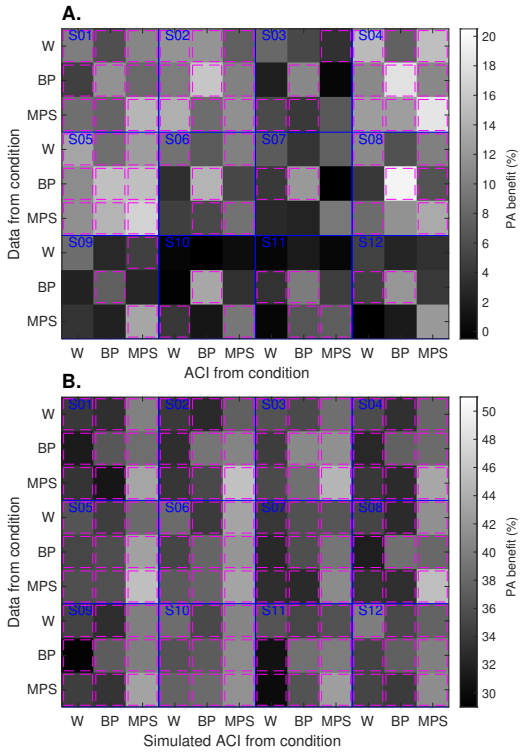
III. COMPLEMENTARY INFORMATION TO THE DATA ANALYSIS

A. Analysis using the data of all participants: Extra figures

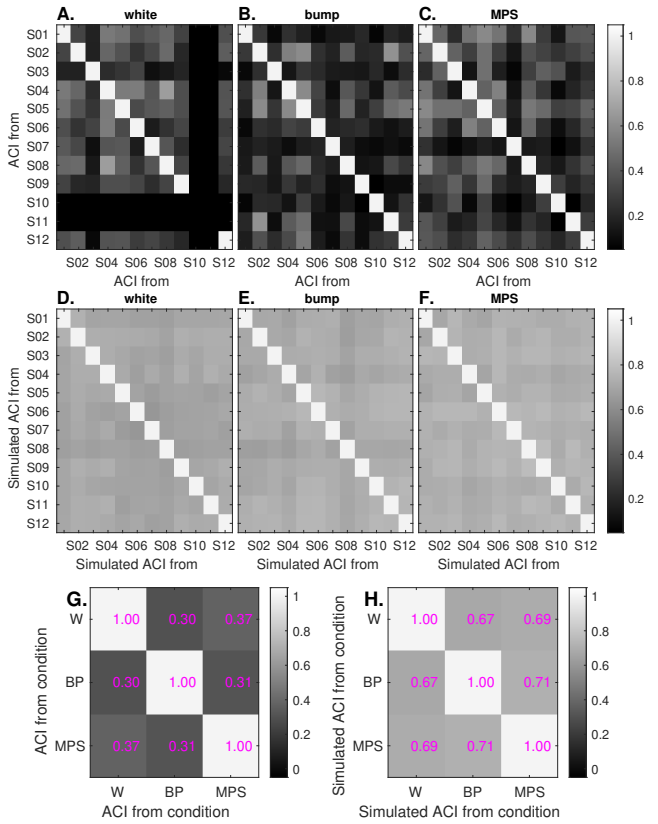
Figure 3 is complementary to main Fig. 8 and contains the cross-prediction values using the deviance per trial (CVD_t). This information had been omitted in main Fig. 8. The panels (**A**–**C**) show the CVD_t values obtained from the experimental data, whereas the bottom panels (**D**–**F**) show the corresponding values for the artificial listener. In this case, the insets 'S01' to 'S02' indicate the set of waveforms used to run the fixed normal-hearing auditory model (Sec. IV).



SUPPL. FIG. 3. (Color online) **A**–**C**: Between-subject cross-prediction matrices for the three conditions using CVD_t. The matrices contain the deviance benefit plus 1.64 SEM (ΔCVD_t+1.64 SEM). When this quantity is less than 0, the cross predictions using the ACIs from the participants indicated in the abscissa are able to predict significantly above chance the data from the participants indicated in the ordinate. Those cases are enclosed in pink boxes. The main diagonals are enclosed in colored squares and correspond to the same auto-prediction values that are shown as open markers in main Fig. 7**A**. The red arrows indicate the ACIs that did not achieve significant auto predictions. In such a case, the significance of the cross predictions was not evaluated. **D**–**F**: Same as the top panels, but using the simulated ACIs derived from the artificial listener.



SUPPL. FIG. 4. (Color online) Between-noise cross-prediction (3-by-3) matrices for (**A**) each participant and for (**B**) each artificial listener, using ΔPA. The pink boxes indicate cross predictions that provided significant better-than-chance ΔPA values. In each 3-by-3 matrix, the main diagonal takes values that were overall higher than those from the off-diagonals.



SUPPL. FIG. 5. Pearson correlation values for ACIs between participants obtained from the experimental data (top panels, **A**–**C**) or obtained from simulations (middle panels, **D**–**F**). The correlation values in the bottom row were obtained from the ACIs between noise conditions for each participant, and then the values were averaged across participants. All matrices in this figure are symmetric with respect to their diagonal.

Figure 4 is complementary to main Fig. 9 and shows the individual 3-by-3 matrices for each participant for the cross predictions between noises. The top and bottom panels show ΔPA values derived from the experimental data and from the artificial listener, respectively. The arithmetic average of ΔPA values across participants, corresponds to the 3-by-3 matrix presented in main Fig. 9.

Figure 5 (top panels, **A**–**C**) shows the correlations across ACIs (from main Fig. 6). Supplementary Fig. 5 (middle panels, **D**–**F**) shows the correlations across simulated ACIs (from main Fig. 10 and Suppl. Fig. 6). The global results in both rows of panels is similar to the results obtained using the ΔPA metric: The correlations across experimental ACIs (top panels) are lower than the correlations across simulated ACIs obtained from the simulations. The off-diagonal correlations are 0.33, 0.20, 0.29 for white, bump, and MPS noises in the top panels. The corresponding values in the middle panels are 0.70, 0.70, and 0.74. The results in the bottom panels indicate the average results for the Pearson correlations within participant but between noises, comparable to main Fig. 9. In agreement with main Fig. 9, the off-diagonal correlations had an average of 0.33 (Panel **G**, experimental data) and 0.69 (Panel **H**, simulation data).

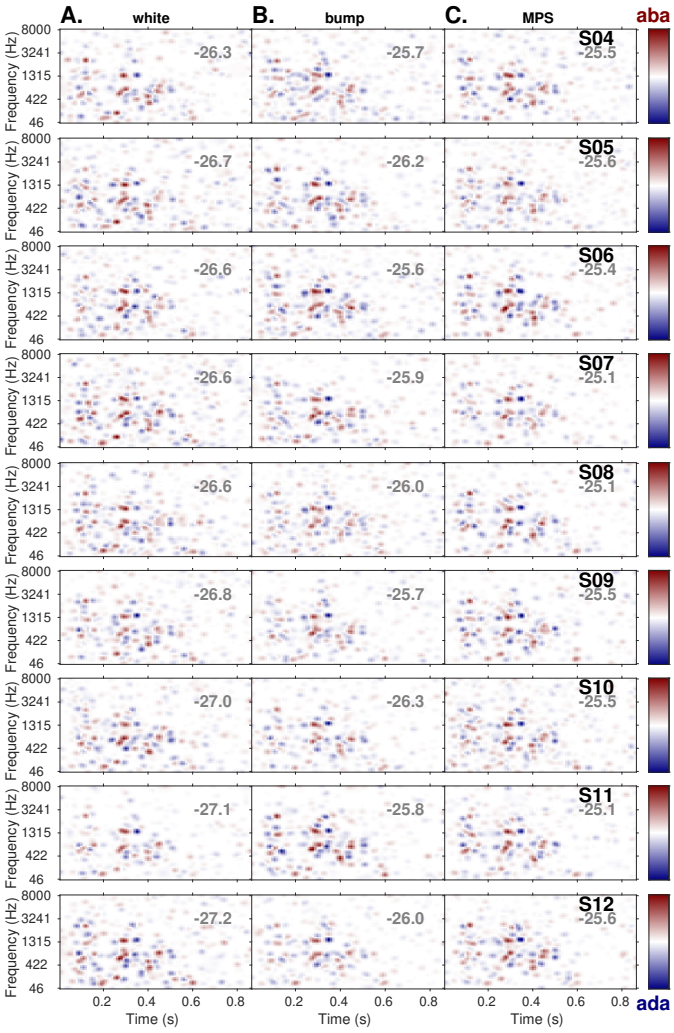
B. Analysis using the data of ten participants

Here we replicated the reported mixed ANOVAs and the assessment of group averaged performance excluding the data of the two last completed participants (S06 and S12), i.e., only using data from the preregistered number of participants ($N = 10$).

Behavioral performance (as in main Sec. III A):

Two-way mixed ANOVA on SNR: This analysis was run to test the learning effect on SNR (comparable results as with $N = 12$). There was a significant effect of the factors masker ($F(2, 287) = 15.82$, $p < 0.001$) and test block ($F(1, 287) = 33.06$, $p < 0.001$). As with $N = 12$, a post-hoc analysis confirmed that the effect of masker type was due to a difference in the bump-noise condition compared to the other two types of noise, with white and MPS noises having statistically the same SNRs.

Two-way mixed ANOVA on d' (comparable results as with $N = 12$): There was a significant effect of



SUPPL. FIG. 6. ACIs derived from the simulations using the artificial listener for white (column **A**), bump (column **B**), and MPS noises (column **C**) using the set of noises from participants S04–S12 (top to bottom rows), which were not shown in main Fig. 10. The values in gray indicate the corresponding mean simulated SNR threshold expressed in dB.

the factors masker ($F(2,137) = 10.24$, $p < 0.001$) and SNR ($F(1,137) = 788.09$, $p < 0.001$).

Two-way mixed ANOVA on c (comparable results as with $N = 12$): There was a significant effect for the factor SNR ($F(1,137) = 13.12$, $p < 0.001$), but not for the factor masker ($F(2,137) = 1.41$, $p = 0.249$).

Out-of-sample prediction accuracy (as in main Sec. III C): This analysis is comparable to the results shown in main Fig. 7. The group results for the ΔPA metric are 8.4, 13.3, and 12.0%, for the white, bump, and MPS noises, respectively. For the analysis of incorrect trials only, the corresponding values were 12.3, 18.3, and 16.9%.

IV. THE ARTIFICIAL LISTENER

As briefly described in main Sec. II C, an auditory model was used to simulate the performance of an average normal-hearing listener who uses a fixed decision criterion to compare sounds. In this sense, the model is used as an artificial listener.

A. Model description

The model consists of a front-end and a back-end processing. The front-end processing converts an incoming sound waveform into an internal representation, i.e., into a representation that is believed to reflect how sounds are actually transformed along the ascending auditory pathway (e.g., Osses *et al.*, 2022).

a. Front-end processing.

The auditory model accepts monaural input waveforms and delivers a three-dimensional signal in time, audio frequency, and modulation frequency, that are expressed in model units (MU), an arbitrary amplitude unit (e.g., Kohlrausch *et al.*, 1992). Most of the model stages have been previously described in detail (Osses, 2018; Osses and Kohlrausch, 2018, 2021). Here, we provide a short description of each stage, emphasizing some small implementation updates.

Outer- and middle-ear filtering (updated): Two cascaded 512-tap FIR filters are used to produce a combined bandpass frequency response (Osses *et al.*, 2022, their Fig. 3). In contrast to the previous model version (Osses and Kohlrausch, 2021), the middle-ear filter is implemented using the linear-phase version instead of its minimum-phase implementation. A group delay compensation is applied to the filtered signal.

Gammatone filter bank: Set of 31 audio frequency bands with f_c between 86.9 Hz and 7819 Hz, spaced at 1 ERB_N, as described by Hohmann (2002). Only the real part of the complex-valued outputs of the filter bank is used.

Half-wave rectification and LPF: The half-wave rectification is followed by a chain of five cascaded first-order IIR filters with $f_{\text{cut-off}}$= 2000 Hz. This chain produces a filter response with a −3-dB point at 770 Hz.

Adaptation loops: This stage approximates the effect of auditory adaptation at the level of the auditory nerve by using five feedback loops based on a resistor-capacitance analogy (full details in Osses and Kohlrausch, 2021, App. B) with time constants τ = 5, 50, 129, 253, and 500 ms. We used the parameter configuration indicated by Osses and Kohlrausch (2021) that uses a limiter factor of 5 instead of 10.

Modulation filter bank (updated): The implementation was mainly based on the filter banks by Osses and Kohlrausch (2021) and Jep. However, (1) the first-order 150-Hz LPF was implemented as an attenuation gain (see, Osses and Kohlrausch, 2021, their Fig. 14C), and (2) the filters were designed using a Q factor of 1, resulting in 7 modulation filters centered at 2.5, 5, 10, 25, 75, 225, and 675 Hz.

b. Back-end stage.

The auditory model performed the same experimental paradigm as each of the twelve participants for the three noise conditions. For the simulations, the same order of noise presentation and the same level roving as the participants was used, but the exact SNR in each interval depended on the specific model responses.

To generate a decision outcome the internal representation of the current trial R_c—the output of the front-end processing—was compared with the /aba/ (T_1) and /ada/ (T_2) template, derived at an arbitrary supra-threshold SNR of −6 dB (i.e., with the speech sample presented at a level of 59 dB SPL). The comparison was based on a cross correlation at lag 0. The artificial listener indicated the option "aba" if $R_c \cdot T_1 \geq R_c \cdot T_2 + K$ or the option "ada" if $R_c \cdot T_1 < R_c \cdot T_2 + K$ (Osses and Varnet, 2021). More formally:

$$\text{response}_{\text{model}} = \begin{cases} \text{"aba"} & \text{if } R_c \cdot T_1 - R_c \cdot T_2 \geq K \\ \text{"ada"} & \text{if } R_c \cdot T_1 - R_c \cdot T_2 < K \end{cases} \quad (1)$$

B. Template assessment: Choice of supra-threshold level and number of averages

The back-end module of the artificial listener was based on template-matching, in our case, based on two "expected signals" or templates, one for target sound /aba/ and one for /ada/. This means that the ongoing experimental trials were compared (by cross correlation) with the two templates, pointing to "aba" or "ada" if the first or second template produced the highest cross-correlation value. This two-template procedure is explained by Osses and Kohlrausch (2021). However, the templates need to be derived at a "supra-threshold level," and needs to be averaged for a number of noisy-target repetitions (Dau *et al.*, 1996). These two parameters are arbitrary.

1. Supra-threshold level

The supra-threshold level was chosen as a level that was assumed to be well above the simulated threshold. We arbitrarily used an $\text{SNR}_{\text{supra}}$=−6 dB, which is approximately 20 dB above the estimated thresholds (see the values in Suppl. Fig. 6, indicated in gray text as insets), between −27.2 and −25.1 dB. This supra-threshold is in line with the results by Derleth and Dau (2000), where they found a maximum shift in simulated thresholds of about 6 dB for a tone-in-noise task using supra-threshold levels 20 dB above the estimated threshold. We confirmed that the simulated ACIs were not significantly affected by the chosen $\text{SNR}_{\text{supra}}$ value. For that purpose, we re-run simulations using an $\text{SNR}_{\text{supra}}$ of −16 dB, following the recommendation by Derleth and Dau (2000) for obtaining stable simulated thresholds. The new simulations produced thresholds that differed by 2 dB or less with respect to our study simulations.

2. Number of noisy-target repetitions

The number of averages used to derive the noisy /aba/ and /ada/ templates was also an arbitrary choice. The normal assumption is that noisy templates derived from a larger number of averages contain less external variability from the background noises. For white-noise maskers, we investigated the effect of deriving templates using N=10, 100, and 1000 averages for the template derivation and there was no significant differences between the simulated ACIs. Although no difference was found for different N values, we decided to adopt $N = 100$, assuming enough external variability in the templates for all noise conditions, i.e., the white noises and for the noises with enhanced envelope fluctuations, the MPS and bump noise conditions.

C. Calibration of the model

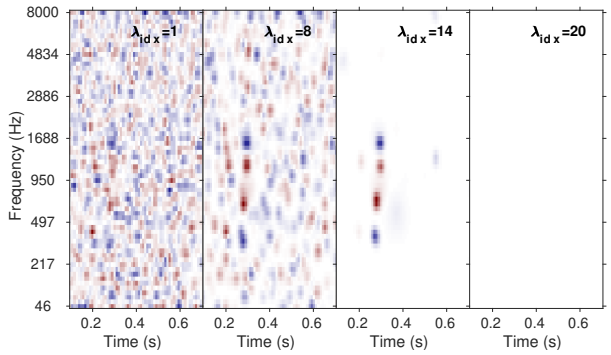
The bias K depended on the exact set of templates and on the type of noise. The use of a $K = 0$ led to biased model responses. For this reason we used K as a free parameter. The fitting of this parameter was performed before the simulation of each new set of noises, using a constant stimulus procedure at a very low speech level arbitrarily set to an SNR of −40 dB (i.e., at a speech level of 25 dB SPL)—a condition that the model should not be able to solve—and we stored the cross-correlation values ($\text{CCV} = R_c \cdot T_1 - R_c \cdot T_2$) for all 4000 trials. The final K value was chosen to be the median of the CCVs.

D. ACIs derived from simulations

The ACIs derived from simulations that used the set of noises of participants S01–S03 were shown in main Fig. 10, in the main text. The remaining simulated ACIs that used the set of noises of participants S04–S12 are shown in Suppl. Fig. 6.

V. ACI FOR DIFFERENT HYPER PARAMETER VALUES

The time-frequency weights in the ACI_k and intercept c_k are obtained during the GLM fitting procedure (see main Sec. II E 3), using the noise vector $\underline{N}_{k,i}$ and the participant's (or artificial listener's) responses. During the 10-fold cross-validation procedure of the lasso regression, different hyperparameter values are evaluated. The intermediate ACIs obtained for four different values of the hyperparameter λ applied to the data of participant S01 in the MPS condition are shown in Suppl. Fig. 7. The right-most ACI, the null ACI, is particularly important for the prediction performance that we used, because the goodness-of-fit metrics of CVD_t and PA (see main Sec. III E 4a) were referenced to that null ACI, whose performance was nearly close to chance. An additional scaling was applied to the PA metric, to correct for guessing, with expected ΔPA values between 0% (performance at chance according to the null ACI) and 100%.



SUPPL. FIG. 7. ACI for participant S01 for the MPS condition using different hyperparameter λ values. During the fitting procedure, the higher the λ value the fitting procedure looks for less and smoother time-frequency cue candidates. The right-most ACI corresponds to the null ACI, where the only non-zero parameter is the intercept c_k. The lambda values in this figure range between $\lambda_1 = 1.1 \cdot 10^{-3}$ and $\lambda_{20} = 0.1$.

VI. RECREATING ALL STUDY FIGURES

All figures from the main text and these supplementary materials can be retrieved using the fastACI script publ_osses2023c_JASA_figs.m. To obtain the figures from the main text:

```
publ_osses2023c_JASA_figs('fig1');
publ_osses2023c_JASA_figs('fig2a');
publ_osses2023c_JASA_figs('fig2b');
publ_osses2023c_JASA_figs('fig3');
publ_osses2023c_JASA_figs('fig4');
publ_osses2023c_JASA_figs('fig5');
publ_osses2023c_JASA_figs('fig6');
publ_osses2023c_JASA_figs('fig7');
publ_osses2023c_JASA_figs('fig8');
publ_osses2023c_JASA_figs('fig8b');
publ_osses2023c_JASA_figs('fig9');
publ_osses2023c_JASA_figs('fig9b');
publ_osses2023c_JASA_figs('fig10');
```

To obtain the figures from the current document:

```
publ_osses2023c_JASA_figs('fig1_suppl');
publ_osses2023c_JASA_figs('fig2_suppl');
publ_osses2023c_JASA_figs('fig3_suppl');
publ_osses2023c_JASA_figs('fig3b_suppl');
publ_osses2023c_JASA_figs('fig4_suppl');
publ_osses2023c_JASA_figs('fig4b_suppl');
publ_osses2023c_JASA_figs('fig5_suppl');
publ_osses2023c_JASA_figs('fig5b_suppl');
publ_osses2023c_JASA_figs('fig6_suppl');
publ_osses2023c_JASA_figs('fig7_suppl');
```

If you downloaded the raw and post-processed data for this study from Zenodo (Osses and Varnet, 2022b), you need to specify the location of downloaded folders as extra input parameters. If the data are stored in the current working directory of MATLAB, you can use:

```
dir_zenodo=[pwd filesep]; % current directory
flags = {'zenodo', ...
         'dir_zenodo',dir_zenodo};

%% To recreate, e.g., main fig 6:
publ_osses2023c_JASA_figs('fig6',flags{:});
```

If the Zenodo data are located elsewhere, provide a valid directory using the variable dir_zenodo.

REFERENCES

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, doi: 10.1121/1.414959.

Derleth, R., and Dau, T. (**2000**). "On the role of envelope fluctuation processing in spectral masking," J. Acoust. Soc. Am. **108**, 285–296, doi: 10.1121/1.429464.

Gnansia, D., Lazard, D., Léger, A., Fugain, C., Lancelin, D., Meyer, B., and Lorenzi, C. (**2014**). "Role of slow temporal modulations in speech identification for cochlear implant users," Int. J. Audiol. **53**, 48–54, doi: 10.3109/14992027.2013.844367.

Hohmann, V. (**2002**). "Frequency analysis and synthesis using a Gammatone filterbank," Acust. Acta Acust. **88**, 433–442.

Kohlrausch, A., Püschel, D., and Alphei, H. (**1992**). "Temporal resolution and modulation analysis in models of the auditory system," in *The Auditory Processing of Speech*, edited by M. Schouten, **10** (Mouton de Gruyter), Chap. 1, pp. 85–98.

Osses, A. (**2018**). "Prediction of perceptual similarity based on time-domain models of auditory perception," Ph.D. thesis, Technische Universiteit Eindhoven, tel-03871102.

Osses, A., and Kohlrausch, A. (**2018**). "Auditory modelling of the perceptual similarity between piano sounds," Acta Acust. united Ac. **104**, 930–934, doi: 10.3813/AAA.919251.

Osses, A., and Kohlrausch, A. (**2021**). "Perceptual similarity between piano notes: Simulations with a template-based perception model," J. Acoust. Soc. Am. **149**, 3534–3552, doi: 10.1121/10.0004818.

Osses, A., and Varnet, L. (**2021**). "Consonant-in-noise discrimination using an auditory model with different speech-based decision devices," in *Proc. DAGA*, pp. 298–301, hal-03345050.

Osses, A., and Varnet, L. (**2022**a). "A microscopic investigation of the effect of random envelope fluctuations on phoneme-in-noise perception," BioRxiv 1–24, doi: 10.1101/2022.12.27.522040.

Osses, A., and Varnet, L. (**2022**b). "Raw and post-processed data for the microscopic investigation of the effect of random envelope fluctuations on phoneme-in-noise perception" doi: 10.5281/zenodo.7476407.

Osses, A., and Varnet, L. (**2023**). "fastACI toolbox: the MATLAB toolbox for investigating auditory perception using reverse correlation (v1.3)" doi: 10.5281/zenodo.7888588.

Osses, A., Varnet, L., Carney, L., Dau, T., Bruce, I., Verhulst, S., and Majdak, P. (**2022**). "A comparative study of eight human auditory models of monaural processing," Acta Acust. **6**, 17, doi: 10.1051/aacus/2022008.

Průša, Z. (**2017**). "The phase retrieval toolbox," in *AES International Conference on Semantic Audio*, Erlangen, Germany.