Consonant-in-noise discrimination using an auditory model with different speech-based decision devices

Alejandro Osses Vecchi, Léo Varnet

Laboratoire des systèmes perceptifs, ENS, PSL University, Paris, France, Email: {alejandro.osses, leo.varnet}@ens.psl.eu

Abstract

This study presents insights into the discrimination of two consonants presented in vowel-consonant-vowel (VCV) words embedded in speech-shaped noise (SSN) by adopting an auditory model that uses a modulation filter bank front-end followed by either of two speech back-end decision modules from the literature. These decision modules have been validated in the past for the discrimination of sentences in closed- and open-sets. Our analysis is focused on the discrimination cues available to the model, evaluating whether these cues might be further used to simulate listener-dependent performance. For that purpose we will rely on a reverse correlation approach by comparing the noise representations that lead to the choice of one or the other consonant.

Introduction

The method of Auditory Classification Images (ACIs) is an experimental paradigm developed to estimate the time-frequency (T-F) cues that different listeners use to discriminate between two phonemes [1]. The method relies on a reverse-correlation approach comparing the precise noise realisations that lead to correct and incorrect responses during the experimental sessions. The method has proven to be a reliable tool to characterise the listening strategies of a participant.

With the goal of gaining more insights into the T-F cues related to the different listening strategies, we adopted an auditory model to simulate ACIs. We chose the modulation-filter-bank (MFB) model [2], which roughly approximates the hearing processing from the outer ear up to the inferior colliculus [3]. This was our first attempt in deriving ACIs from simulated phoneme discrimination data, for which either of two decision back-end modules were attached to the model. With this approach, we simulated the performance of an average human listener that could use one of two possible decision strategies.

Methods: Experiment

In this study we present new data collected from two participants for the discrimination of the words /aba/ and /ada/, that are aligned on syllable onsets, having a total duration of 0.84 s (including initial and final silent sections of 0.075 s), and uttered by a female speaker as in [1], but embedded in SSNs, i.e., using Gaussian noises that have a long-term averaged spectrum matched to the spectrum of a female speaker. The experiment consisted in 5000 presentations of /aba/ or /ada/ (2500 each), where each trial consisted of one speech-in-noise interval to which the participants had to indicate one of two possible answers (/aba/ or /ada/). This means that the task is implemented as a one-interval two-alternative forced-choice (1-I, 2-AFC) experiment. In the experiment, the level of the noises was fixed at 65 dB SPL and the signal-to-noise ratio (SNR) was adjusted on a trial-by-trial basis to track the speech level at which the participants reached a 70.7%-correct score using a weighted one-up one-down method [4] with unequal step sizes (2.41 and 1 dB for the up- and down-steps, respectively). Hence, the level of the speech samples was 65 dB SPL for an SNR=0 dB, or lower for lower SNRs. All the noises were stored in the test computer together with the corresponding participants' responses for a later processing using the ACI method. The sounds were presented to the participants using headphones. To avoid the participants' fatigue, the experiments were organised in 12 sessions of 400 trials and a last session of 200 trials. This experiment is implemented in the `speechACI_varnet2013` script of the fastACI toolbox [5].

Methods: Auditory model

The paradigm described in the previous section was also used in the simulations. The only difference is the use of an artificial listener, meaning that the sounds were delivered monaurally to the auditory model (`osses2021.m` routine from the AMT Toolbox [6]). The internal representations correspond to the three-dimensional outputs (time, frequency, modulation frequency) of the MFB model (Stage 5 from [2]), expressed in model units (MU), that were used as input to two decision back-ends (details below). In both decision schemes, two templates were derived, one for each word, and no internal noise was employed. The model conducted the simulations in thirteen sessions (as in the experiments), and a new pair of templates was derived at the beginning of each session.

Template derivation: For each of the sounds (/aba/ or /ada/) the next steps were followed. Ten noises were randomly generated, and were added to the corresponding speech sample at 59 dB (SNR of -6 dB). The 10 simulated internal representations were arithmetically averaged and used as "a template," i.e., an expected internal representation that should lead the artificial listener to a correct discrimination. For convenience the template of the words /aba/ and /ada/ are labelled as T_1 and T_2, respectively. Finally, the templates were scaled to jointly meet the condition of unit energy (Eq. 2 in [2]).

Decision back-end 1 (from Osses & Kohlrausch [2]): The internal representation of the current interval R_c is

compared with T_1 and T_2, where the artificial listener, i.e., the model, performs the cross-correlation at lag 0. The artificial listener indicates the option /ada/ if $R_c \cdot T_1 \geq R_c \cdot T_2 + K$ or the option /ada/ if $R_c \cdot T_1 < R_c \cdot T_2 + K$. More formally:

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

where K is a constant that can be used to bias the model choice from /aba/ towards /ada/. $K = 0$ represents the exact decision as used in [2]. Different values of K were used in our simulations.

Decision back-end 2 (from Relaño-Iborra *et al.* [7]): The representation R_c is correlated with T_1 and T_2 using the two-dimensional Pearson correlation (function corr2 from MATLAB) across the dimensions of time and frequency and hence assuming an independent processing for each of the (up to) 12 modulation filters in the model. The correlation values r are obtained in time windows defined by the centre frequencies of the modulation filters $\mathrm{m}f_c$ ($\text{window}_{\text{length}}=1/\mathrm{m}f_c$), where first the negative r values are set to zero and then are arithmetically averaged for each filter. The obtained twelve r values are again arithmetically averaged resulting in one single Pearson correlation for each template comparison [7]. For the comparisons between R_c and T_1 and T_2 we adopt the nomenclature of r_1 and r_2 for the corresponding obtained correlation value. The artificial listener indicates the option /aba/ if $r_1 \geq r_2$ and /ada/ otherwise.

Methods: Assessment of ACIs

The ACIs were obtained using the fastACI_getACI script of the fastACI toolbox that requires as input the experiment information, the set of noises used by each (human or artificial) listener, and the collected listener's responses. In line with the original method (e.g., [1]), only the waveforms of the background noises were used to obtain the ACIs. But in contrast to previous work, the T-F representations were obtained using the toolbox option TF_type='gammatone' which decomposes the noise representations into frequency bands using the same Gammatone filter bank and simplified inner-hair-cell envelope extractor as in the auditory model (osses2021.m with only Stages 2 and 3 being enabled), but adopting a frequency spacing of 0.5 Equivalent Rectangular Bandwidths (ERB) between 40 and 8000 Hz, resulting in 64 frequency bands. The outputs of the envelope extractor that have a default sampling frequency of 16 kHz are "downsampled" by taking average amplitudes within 1-ms windows for each frequency channel, obtaining a representation sampled at 1000 Hz. The resulting T-F representation for each 0.84-s long sound is a 64×84 matrix (64 bands times 84 time bins). The 5000 T-F noise representations ($5000 \times 64 \times 84$ matrix) of each participant are then used as input to a generalised linear model (GLM) to fit that matrix to the corresponding participant responses (5000×1 matrix). For this purpose the fastACI toolbox with the option glmfct='lasso' is used. This option uses the regularisation function lasso.m from MATLAB, that looks for the fit that minimises the cross-validation deviance with respect to the input data. The Lasso function receives the T-F representations after a Laplacian Pyramid decomposition, to provide an a-priori smooth weighting across the time and frequency dimensions of the input matrix (e.g., [8]). Finally, the reconstructed 64×84 matrix of weights providing the best fit between the input (T-F representation of the noise) and output (vector of participant's response) is selected as the final ACI. We used the same procedure for the human and artificial listeners.

Results

Experimental results: The experimental results for participants S01 and S02 are shown in Fig. 1.

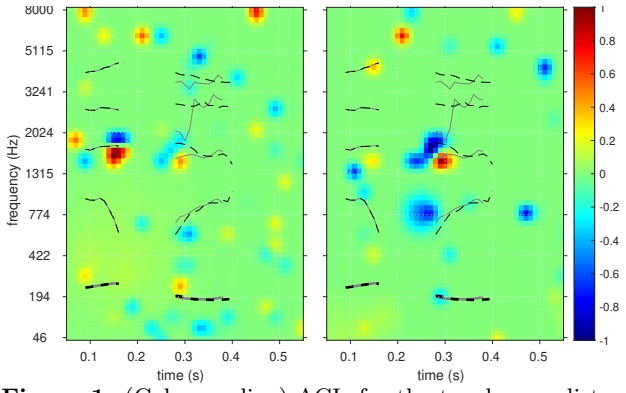
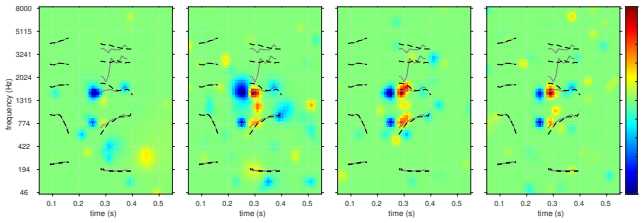


Figure 1: (Colour online) ACIs for the two human listeners S01 (left panel) and S02 (right panel). The red and blue coloured areas represent the T-F regions where the presence of noise dominantly biased the participant's responses towards /aba/ or /ada/, respectively. The green areas represent the regions that were not weighted by the participants. The grey continuous and black dashed lines represent the fundamental frequency and the first four formants for the vowels in /aba/ and /ada/, respectively.

Participant S01 (Fig. 1, left panel) provided more weighting to T-F bins located around the offset of the first /a/ ($t \approx 0.16$ s) using the information just above the second format to favour the choice of /ada/ (blue area) and just below the formant to favour /aba/ (red area). In contrast, participant S02 (Fig. 1, right panel) weighted more the T-F bins located near the onset of the second /a/ ($t \approx 0.25$ s) with cues around the first and second formants to favour the choice of /ada/ (blue areas) and around the second formant to favour /aba/ (smaller red area at t=0.28 s). There are several other small T-F regions that seem to be relevant for both participants above 5000 Hz and also at times later than 0.35 s, but it is possible that these weightings would disappear if more trials were to be collected.

Simulations using decision 1: The simulations using the decision back-end 1 are shown in Fig. 2, where we first obtained the ACI for K=0 (as used in [2]). Due to the bias of the model for this K value (median of 24%, Fig. 4, "0 MU"), favouring the choice of /ada/ (76% of the times), ACIs using several other K values were obtained, as shown in Fig. 2.



(a) $K = -1.55$ MU (b) K=0 MU (c) K=0.39 MU(d) K=0.78 MU

Figure 2: (Colour online) ACIs for the artificial listener using the decision device 1 for different values of K. The colour codes and axis legends are as in Fig. 1.

From these simulations, the model using K=0.39 yielded the lowest response bias (median of 59%, Fig. 4(b), "0.39 MU"), also getting the lowest simulated discrimination threshold (SNR=−16.9 dB, Fig. 4(a)). That ACI is therefore used for the remaining of our analyses and is replotted in the left panel of Fig. 3.

Simulations using decision 2: The simulations using the decision back-end 2 are shown in the right panel of Fig. 3. In contrast to the simulations with the previous detector, the model was nearly unbiased with a 46% of the times choosing /aba/ (Fig. 4(b)) and a simulated threshold SNR=−17.5 dB (Fig. 4(a)).

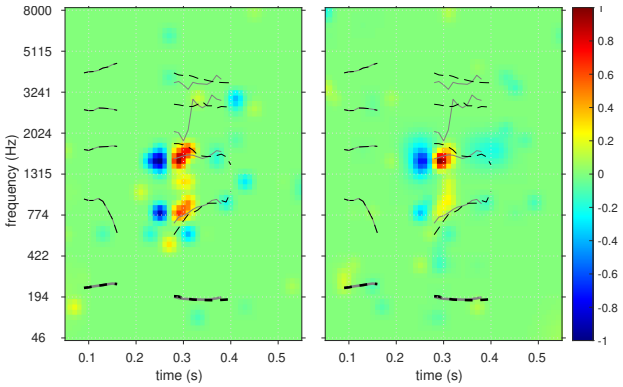


Figure 3: (Colour online) ACIs for the artificial listeners, i.e., the auditory model using the decision device 1 with K=0.39 MU (left panel) and using the decision device 2 (right panel). The colour codes and axis legends are as in Fig. 1.

Discussion and further work

Comparison between ACIs: Consistent with previous results in white noise [1], the simulated ACIs (Figs. 3) revealed the presence of prominent T-F cues in the region of the second formant (both decision devices) and in the region of the first formant (prominent weight for decision 1, milder for decision 2), but this was only found around the onset of the second /a/. Such region was relevant for participant S02 (Fig. 1, right), but not for participant S01 (Fig. 1, left). It is also interesting to point out that while for the artificial listener the cues were arranged horizontally, i.e., with sequential blue and red regions, for participants S01 and S02, the coloured regions were arranged more vertically. This means that the formant frequency at the onset is more relevant for real participants, and the formant timing is more relevant for the artificial listener.

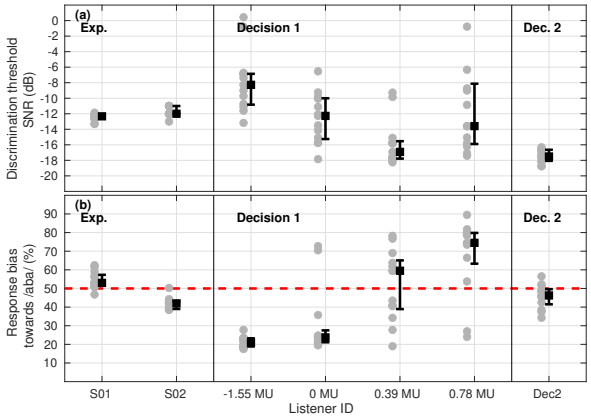


Figure 4: (a) Discrimination thresholds for each human or artificial listener expressed as an SNR (lower is better performance). (b) Listener's bias indicating the number of times (in percentage) a participant chose the option /aba/. A 50% bias (red dashed curve) would indicate no response bias. In both panels, the black markers indicate median and interquartiles across the 13 sessions, and the grey markers indicate the individual values of each test session.

Decision 1, changing the model bias: With the use of different K values we attempted to simulate how the ACIs change for artificial listeners that tended to choose more often one of the two speech samples (/aba/ more often than /ada/ or vice versa). The ACIs for different values of K (Fig. 2, panels b–d) did not change significantly, if we consider that the coloured regions in the plots stayed at the same T-F region. The only exception was for the artificial listener with K=−1.55 MU, that lead to a bias of 21% (i.e., /ada/ was chosen 79% of the times), meaning that the artificial listener was primarily weighting the information in /ada/ and not the one related to /aba/, resulting in an ACI with some blue regions but without red ones.

Use of different decision back-ends: the use of a different decision back-end implies that the same listener –here the auditory model– may adopt different strategies to solve a specific task. Due to the simplicity of our speech task, with only two possible choices, we also tried to reduce the number of potential strategies available to the (human or artificial) listeners. We assume that in this task with two contextless words, cognitive (top-down) processes would play a much less important role in contrast to what it would happen when conducting a sentence-in-noise test. The current method can thus be seen as a way to assess the listening strategies of a listener under a very particular comparison condition, which should be cue-based, or more focused on the bottom-up information available in the sound representations. One problem that remains open is how to control the potential bias in the participant's responses and how this interacts with the high chance level of this task (chance level=50%, for two options).

Further use of auditory models: In this short contribution we derived ACIs from simulated discrimination data using an auditory model (`osses2021.m` [3, 6]), in a first attempt to understand how much of the T-F cues

measured in human participants for the comparison of two speech sounds (words /aba/ and /ada/) could be explained using this auditory model. The auditory model approximates the signal processing of a normal-hearing listener and has been used previously to simulate a number of psychoacoustic tasks. From the two adopted decision back-ends, Decision 2 (based on [7]) was previously validated in the evaluation of sentences in noise (using the CLUE and the DANTALE II Danish speech corpora) but adopting an auditory model with a non-linear cochlear processing [7]. The study in [7] built upon and extended the work presented in [9], evaluated with the same set of speech materials. On the contrary, Decision 1 was recently used in a similarity comparison between piano sounds, a task that was simulated using two templates, and shares some resemblance with the speech processor originally described in [10]. Although we did not pose major efforts in individualising the parameters of the auditory model, such a parameter customisation would represent one of the natural further steps in our research. We foresee to do this in two ways: (1) Adopting an memory-like internal noise to switch the focus from later to earlier segments of the speech sounds, and (2) to apply customised settings for the peripheral model to account for aspects such as elevated hearing thresholds. Regarding the first point, the use of a backward-increasing additive noise, such as that suggested in [11], may help to get a model cue-weighting towards the first vowel-consonant transition, as we observed for S02. On the other hand, the use of customised peripheral parameters may help to understand how much shift in the ACIs can be expected if extra information about the participants (e.g., hearing thresholds) is used as input to the model.

To conclude this contribution, we would like to summarise our observations and further steps:

- The ACI methodology can be used to assess the listening strategy of both human (as also previously shown, e.g., in [1]) and artificial listeners in speech shaped noises (new in this study).
- Similar to human participants, the artificial listener was able to find relevant (and stable) discrimination cues for the discrimination of two VCV words in the transition of the consonant with the second vowel, at around the first (Decision 1) and second formant (Decisions 1 and 2).
- The artificial listener was not able to find any T-F cue in the transition of the first vowel and the consonant as we observed for participant S01. The understanding of this different listening strategy remains a major challenge for future modelling work. We suggested one possible workaround for this, using an "additive memory noise" in future simulations.
- Changing the model bias (K) in Decision 1 did not change significantly the simulated ACIs (Fig. 2).
- We adopted two decision back-ends that do not include strong assumptions about the potential top-down processing of participants. However, an implicit assumption of using a template-matching approach is that participants can extract the speech information from the speech-in-noise representations.
- We did not adopt customised auditory model configurations, meaning that we did not focus in the individualisation of our simulations. This represents one of our further steps.

Acknowledgements

This research was supported by the French National Research Agency (Grant ANR-17-EURE-0017).

References

[1] Varnet, Knoblauch, Meunier, & Hoen: "Using auditory classification images for the identification of fine acoustic cues used in speech perception." Front. Hum. Neurosci. 7 (2013), 1–12

[2] Osses Vecchi & Kohlrausch: "Perceptual similarity between piano notes: Simulations with a template-based perception model." J. Acoust. Soc. Am. 149 (2021), 3534–3552

[3] Osses Vecchi, Varnet, Carney, Dau, Bruce, Verhulst, & Majdak: "A comparative study of eight human auditory models of monaural processing." ArXiv id: 2107.01753 (2021)

[4] Kaernbach: "Simple adaptive testing with the weighted up-down method." Percept. Psychophys. 49 (1991), 227–229

[5] Osses Vecchi & Varnet. "fastACI toolbox: Exploring phoneme representations and their adaptability using fast Auditory Classification Images," Github commit b18d919 (Last accessed on 25 August 2021).

[6] Majdak, Hollomey, & Baumgartner: "AMT 1.0: the toolbox for reproducible research in auditory modeling." (2021, Submitted to Acta Acustica)

[7] Relaño-Iborra, Zaar, & Dau: "A speech-based computational auditory signal processing and perception model." J. Acoust. Soc. Am. 146 (2019), 3306–3317

[8] Mineault, Barthelmé, & Pack: "Improved classification images with sparse priors in a smooth basis." J. Vis. 146 (2009), 1–24

[9] Jørgensen & Dau: "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing." J. Acoust. Soc. Am. 130 (2011), 1475–1487

[10] Holube & Kollmeier: "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model." J. Acoust. Soc. Am. 100 (1996), 1703–1716

[11] Wallaert, Moore, Ewert, & Lorenzi: "Sensorineural hearing loss enhances auditory sensitivity and temporal integration for amplitude modulation." J. Acoust. Soc. Am. 141 (2017), 971–980