



HAL
open science

Mapping the spectrotemporal regions influencing perception of French stop consonants in noise

Géraldine Carranante, Charlotte Cany, Pouria Farri, Maria Giavazzi, Léo Varnet

► **To cite this version:**

Géraldine Carranante, Charlotte Cany, Pouria Farri, Maria Giavazzi, Léo Varnet. Mapping the spectrotemporal regions influencing perception of French stop consonants in noise. *Scientific Reports*, 2024, 14 (1), pp.27183. 10.1038/s41598-024-77634-w . hal-04774563

HAL Id: hal-04774563

<https://hal.science/hal-04774563v1>

Submitted on 8 Nov 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License



OPEN Mapping the spectrotemporal regions influencing perception of French stop consonants in noise

Géraldine Carranante¹, Charlotte Cany¹, Pouria Farri¹, Maria Giavazzi² & Léo Varnet¹✉

Understanding how speech sounds are decoded into linguistic units has been a central research challenge over the last century. This study follows a reverse-correlation approach to reveal the acoustic cues listeners use to categorize French stop consonants in noise. Compared to previous methods, this approach ensures an unprecedented level of detail with only minimal theoretical assumptions. Thirty-two participants performed a speech-in-noise discrimination task based on natural /aCa/ utterances, with C = /b/, /d/, /g/, /p/, /t/, or /k/. The trial-by-trial analysis of their confusions enabled us to map the spectrotemporal information they relied on for their decisions. In place-of-articulation contrasts, the results confirmed the critical role of formant consonant-vowel transitions, used by all participants, and, to a lesser extent, vowel-consonant transitions and high-frequency release bursts. Similarly, for voicing contrasts, we validated the prominent role of the voicing bar cue, with some participants also using formant transitions and burst cues. This approach revealed that most listeners use a combination of several cues for each task, with significant variability within the participant group. These insights shed new light on decades-old debates regarding the relative importance of cues for phoneme perception and suggest that research on acoustic cues should not overlook individual variability in speech perception.

Keywords Phoneme discrimination, Reverse correlation, Auditory classification image, Acoustic cues, Stop consonants

This study investigates the acoustic information on which the human auditory system relies to decode a speech signal into phonemes. Bridging the realms of acoustics and linguistics, speech can be seen as a code between specific time-varying spectral patterns, called acoustic cues, on the one hand, and elemental linguistic units, such as phonetic traits, on the other hand. This “acoustics-to-phonetics conversion” forms the first layer of the recognition chain, whose output is then transformed into phonemes, syllables, words, etc., to finally retrieve the original meaning of the utterance. The relationship between the acoustic signal and its decoding into linguistic units has been a key research problem for speech science in the last century. However, as of today, there is still an important open debate on the exact acoustic information extracted and exploited by our perceptual system to recognize and categorize phonemes. In this article, we reveal the spectrotemporal mapping of the acoustic information listeners use to categorize (French) stop consonants in noise by analyzing the confusions produced at the trial-by-trial level. While other methods based on signal reductions have already been used to pursue this goal (see below for a review), this type of approach offers the potential to achieve an unprecedented level of detail in localizing acoustic cues.

Over the course of a century of scientific investigations on the acoustic cues governing stop consonant recognition, different methodologies have been proposed, yielding convergent yet not entirely congruent insights. Historically, Fletcher ran in 1922 the first perceptual experiments in order to identify acoustic information in the speech signal¹. He had participants listen to and repeat nonsense speech syllables, either low-pass or high-pass filtered, with various cutoff frequencies. In this way, he sought to identify the frequency bands which contribute the most to intelligibility. Although this initial approach lacked the precision needed to examine individual acoustic cues, Fletcher’s work laid the foundation for subsequent developments in experimental phonetics. Notably, it led to the important conclusion that the speech signal encompasses multiple cues distributed across independently-processed frequency channels and optimally combined^{2,3}.

Following Fletcher, other teams measured the intelligibility of nonsense syllables at various signal-to-noise ratio (SNR). In particular, the examination of confusion matrices in syllable-in-noise recognition tasks by Miller

¹Laboratoire des systèmes perceptifs, Département d’études cognitives, École normale supérieure, PSL University, CNRS, 75005 Paris, France. ²Laboratoire de Sciences Cognitives et Psycholinguistique, Département d’études cognitives, École normale supérieure, PSL University, CNRS, 75005 Paris, France. ✉email: leo.varnet@cnrs.fr

& Nicely⁴ revealed that perceptual errors were not randomly distributed. Instead, these errors clustered within confusion groups aligned with primary phonological features identified by phoneticians. For instance, stop consonants in French are organised along two main features: voicing, distinguishing between voiced ([b], [d], [g]) and unvoiced ([p], [t], [k]) stops, and place of articulation, distinguishing between labial ([b], [p]), coronal ([d], [t]) and dorsal ([g], [k]) stops. A careful examination of confusion matrices as a function of SNR led to important findings about the acoustic cues underlying the recognition of these consonants^{5,6}. Notably, the dental stop [t] exhibited more robustness to noise compared to its labial and coronal counterpart [p] and [k], attributed to the presence of a high-frequency burst cue at the consonant onset. Singh & Allen⁶ argued that the recognition of stop consonants in noise generally depends on the detection of a single necessary and sufficient cue, resulting in an overall binary error pattern (cue detected vs. cue not detected). Exceptions to this rule-of-thumb included the labial stop [b], which seemed to lack a single noise-robust cue, and instances of poorly-articulated utterances introducing conflicting cues into the recognition process.

In the 1950s, Liberman, Delattre, Cooper, and other researchers at the Haskins Laboratories initiated a series of landmark studies on the acoustic cues of stop consonants leveraging one of the earliest speech synthesizers. Based on the spectrographs of natural speech recordings, it was postulated that two cues could be both necessary and sufficient for the perception of stop consonants: the release burst (high frequency energy at the onset of the consonant) and the consonant-vowel transition, which corresponds to the movement of the formants from the articulation of the consonant to the steady state of the vowel (henceforth CV transition). Two experiments were carried out to test this hypothesis. In the first experiment, simplified synthetic speech stimuli were created, composed of a stable vowel part (two formants F1 and F2 at fixed positions) preceded by a burst at varying frequencies⁷. Identification scores measured on a group of listeners revealed that while the burst frequency influenced the perceived consonant, the relationship was not straightforward. In some cases the distribution of responses was even found to be bimodal, with the same percept emerging for separate ranges of frequencies. The second experiment dropped the burst and examined the effect of CV formant transitions alone⁸. The speech sounds were again synthesized from two formants (F1 and F2). However, this time, the F2 was designed with various transition types, from a rising F2 to a falling F2 with regularly-spaced intermediate steps. The F1 was designed to show either a large or reduced transition to evoke the perception of voiced or unvoiced consonants, respectively. In general, results on the F2 transition indicated that stop consonants were ordered from [b] to [d] to [g] along the voiced continuum, from rising to falling F2, and from [p] to [t] to [k] along the unvoiced continuum. However, it again seemed impossible to find a direct correspondence between phonemes and F2 onset frequency. This has led the authors to suggest that phonetic decision could be based on the combination of several cues. In an attempt to summarize the results obtained by the early synthetic speech experiments carried out at the Haskins laboratories, Delattre⁹ listed three possible acoustic cues for place perception, and six acoustic cues for voicing perception. Later technical developments of speech resynthesis algorithms allowed for a refinement of the “synthetic speech continuum approach” in terms of naturalness of the stimuli. Cue-trading experiments¹⁰ and 2-dimensional continuum experiments¹¹ have also considered the orthogonal manipulation of a primary and a secondary cue, for example formant transition and release burst for [b]-[d] categorization, or voice-onset time (VOT) and f₀ onset for [d]-[t] categorization^{12,13}.

Two years after Liberman et al.'s pioneering synthetic speech experiment, a different approach was proposed to answer the question of whether the burst or the formant transitions was the most important cue for the perception of stops. Contrary to the speech synthesis approach, this method employed natural speech utterances as stimuli. Segments of the sounds were selectively cut out and presented to the participants either in isolation or combined with different contexts. By varying the portion of the signal that was removed, it became possible to locate the temporal position of the necessary and sufficient cues. Using this “gating” approach on unvoiced stops in CV sequences, Schatz demonstrated that the burst cue does not provide sufficient information to identify [k] in all vowel contexts¹⁴, confirming Liberman et al.⁷ conclusion that both the burst and formant cues are required for correct identification of the sounds. Similar phenomena were observed for other stops and other vocalic contexts^{15–18}, although other authors have claimed that the burst may actually be a sufficient cue^{19,20}. More generally, an important insight from the gating experiments is that, in some cases, recognition scores remained above chance even when the consonant segment was entirely removed, indicating that listeners can make use of the coarticulation cues information located in adjacent segments^{16,21,22}.

Recently, Jont Allen and his research group integrated the filtering, gating and masking approaches into a unified framework termed “three-dimensional deep search”. By combining the recognition scores obtained along the frequency, temporal, and intensity dimensions on individual utterances, they were able to identify necessary and sufficient cues with an unprecedented level of precision²³. In a first study, they observed that the high-frequency burst energy played a necessary role in the recognition of [t]: intelligibility scores for that consonant rapidly dropped when this acoustic feature was filtered out, masked by noise or gated out²⁴. Furthermore, the audibility of this cue was shown to directly predict the robustness of consonant [t] in white or speech-shaped noise. This approach was then generalized to all stop consonants²³. The burst, characterized by its frequency distribution and delay to the onset of voicing, was found to be necessary and sufficient for the identification of all consonants. However, in the case of [d], [g], [b], and [p], the CV F2 transition cue appeared to be an important cue, necessary to achieve perfect intelligibility. The conclusion that stops are primarily defined by the characteristics of their burst was later confirmed by selectively removing this cue and showing that this manipulation could shift perception from one consonant to the other²⁵.

As detailed in the previous paragraphs, the extensive exploration of stop consonant perception using various approaches yielded important findings. In particular, perceptual experiments have revealed that the recognition of specific phonemes relies on the detection of fine time-frequency features, while large portions of the speech signal are not information-bearing. Notably, place-of-articulation cues may include (depending on the specific consonant, vocalic context, and parameters of the experiment): frequency of the release burst,

F2 CV transition, F3 CV transition, as well as VC transition cues^{9,26}. As for the perceptual cues to voicing, they include the intervocalic interval duration, the duration of the preceding vowel, burst strength and location, f_0 and F1 transitions, and the presence of a voice bar^{13,27,28}. Despite the overarching consensus on the existence and importance of specific acoustic cues for consonant perception, disagreement persists regarding the relative weighting and exact roles of these cues in the recognition process. Regarding place perception, for instance, some researchers have argued that F2 is the primary cue, complemented by a secondary burst cue^{7,8}: although either the release burst or the CV formant transitions alone are sufficient cues to place perception¹⁷, the latter have been shown to play a more dominant role^{29,30}. On the contrary, other authors have claimed that the burst is the dominant cue: according to their results, this cue is context-independent¹⁹, used in isolation in [t] and [k]^{23,24} and necessary for the correct perception of all stops^{6,23,25}.

These divergences in the results of experimental studies may, in part, be due to methodological limitations. Most of the aforementioned experiments used “reduced speech” stimuli, i.e. speech sounds that have been drastically modified to reduce the number of cues they contain. Whether achieved through filtering, truncation or synthesis, this approach makes it possible for the experimenter to manipulate acoustic cues independently, or to isolate specific cues. However, a notable drawback is the often low quality of the reduced stimuli, because of the limited number of cues or because these cues do not co-vary in the same way as in natural speech. It is therefore possible that the cues identified in these experiments are specific to the type of stimuli used, and that they would not generalize to natural speech comprehension. Furthermore, these techniques make it difficult to study several cues at once, which may also explain why different research teams have converged to different cues. One major step towards a more unified approach has been taken by Jont Allen and his team who combined different forms of reduced speech (filtered speech and truncated speech) with a masking experiment²³. Nevertheless, combining the results of these three types of experiments into a general picture requires a difficult and mostly qualitative interpretation step.

In the present study, we investigated the question of the cues to stop consonant perception using the Auditory Classification Images (ACI) approach, a recently-developed technique for revealing the acoustic cues involved in auditory categorization tasks. Unlike previous approaches, the ACI is based not on reduced speech, but on natural speech stimuli embedded in a high level of background noise with randomly added bursts of energy (referred to as “bump noise”). The rationale behind this approach is to randomly add bursts of noise onto natural speech recordings and to identify which burst locations influence perception the most. The outcomes are summarized into an ACI, a time-frequency map of the impact of burst on phonetic decision. This map therefore indicates which spectrotemporal regions bear information for the task, i.e., corresponds to an acoustic cue. The ACI approach has already been successfully applied to a /aba/-/ada/ categorization task^{31,32}, revealing multiple cues for place-of-articulation perception, including the F2 and F1 CV transitions, the release burst, as well as VC transition cues in the initial [a]. Our team is currently pursuing a research program aimed at extending this approach to all French stops, in order to produce a comprehensive list of cues used for place-of-articulation and voicing perception in plosive consonants. This study represents a first step towards that goal, by focusing on the vocalic context [a] and the seven contrasts [b]-[d], [d]-[g], [p]-[t], [t]-[k], [b]-[p], [d]-[t] and [g]-[k]. These contrasts differ by a single phonological trait (place-of-articulation or voicing), and only involve adjacent places of articulation (labial and alveolar, or alveolar and velar).

The research hypothesis for this study were preregistered (<https://osf.io/fqejt>). They can be summarized as follows:

Hypothesis 1 Using the specific set of noises used during the experiments and the derived ACIs, we can predict the trial-by-trial responses (e.g. /ada/ vs. /aga/) of each participant with an accuracy that is significantly above chance.

Hypothesis 2 The location of the cues revealed by the ACI should match the predictions from the phonetic literature, described in the previous sections, since both are related to the availability of perceptual information to a voicing or place contrast. We thus expect to find significant perceptual weights on the same spectro-temporal regions.

Hypothesis 3 In each phonetic contrast considered, the ACIs will be globally similar for all participants. In line with the previous ACI studies on /aba/-/ada/ categorization, we expect that the primary cues will be present in every individual while secondary cues may or may not be used by a given listener.

A fourth hypothesis regarding the performances of a perceptual model of the human auditory system in the task was also preregistered. As this analysis is not central to our main conclusions, it will be detailed in the Supplementary materials.

Methods

All stimuli and procedures described in this section were preregistered (<https://osf.io/fqejt>). As described in the supplementary materials, the experiment can be replicated within the fastACI toolbox³³ under the name speechACI_Logatome and all analyses and figures can be reproduced using the script `publ_carranante2024_figs` and the data available as a Zenodo repository (<https://doi.org/10.5281/zenodo.11507060>).

The same methodology as in Osses et al.³¹ was followed, except that 6 new phonetic contrasts and a single noise type (“bump noise”) were used. The readers are referred to this initial publication for a more detailed description of the protocol and analyses. Some of the data presented in the /aba/-/ada/ condition was already reported and analysed in the initial study³¹.

The study was approved by the Comité d’Ethique de la Recherche Paris-Université, and conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Stimuli

Target sounds

The experiment comprises 7 phoneme-categorization conditions corresponding to different pairs of stimuli: /aba/-/ada/ (condition ABDA), /ada/-/aga/ (condition ADGA), /apa/-/ata/ (condition APTA), /ata/-/aka/ (condition AKTA), /ada/-/ata/ (condition ADTA), /aba/-/apa/ (condition ABPA), /aga/-/aka/ (condition AGKA).

We used 6 productions of speech consonant-vowel-consonant (CVC) pseudowords from one single male speaker, taken from the OLLO speech corpus³⁴ ([aba]: S43M_L007_V6_M1_N2_CS0; [ada]: S43M_L001_V6_M1_N1_CS0; [aga]: S43M_L003_V6_M1_N2_CS0; [apa]: S43M_L008_V6_M1_N1_CS0; [ata]: S43M_L002_V1_M1_N1_CS0; [aka]: S43M_L004_V1_M1_N1). These speech samples were preprocessed to align the time position of the vowel-consonant transitions, to equalize their energy per syllable, and to have the same total duration. The stored sounds have a duration of 0.86 s, and a sampling frequency of 16000 Hz. The target speech sounds are shown in Fig. 1. The use of a single utterance for each target consonant is addressed in the discussion and further justified in the Supplementary Materials.

Background noise

This experiment is based on the “bump noise” condition described in Osses et al.³¹ This type of masker has a white-noise-like long-term spectrum, with enhanced temporal fluctuations. The bump noise instances were generated by introducing 30 randomly located Gaussian-shaped bumps into a white noise instance. The bumps had a temporal width of $\sigma_t = 0.02$ sec, a spectral width of $\sigma_f = 0.5$ ERB, with amplitudes emphasized by a maximum of 10 dB.

For each participant, a new set of 4000 noises was generated, with the same sampling frequency and duration as the target sounds. In each trial, the noise was presented at a level of 65 dB SPL and the target sound was adjusted in level depending on the targeted SNR. The resulting noisy speech sounds were presented diotically via headphones. A small roving in the presentation level between -2.5 and $+2.5$ dB was applied to discourage the use of level cues.

Experimental procedure

Each dataset consisted in a total of 4000 observations collected for one participant in one condition. The experiment was divided in 10 blocks of 400 trials each (≈ 15 minutes per block). Each trial consisted in a one-interval two-alternatives forced choice (“yes/no paradigm”): the participant was presented with one of the two noisy targets and had to categorize it as one target pseudoword or the other by pressing button 1 or 2. A

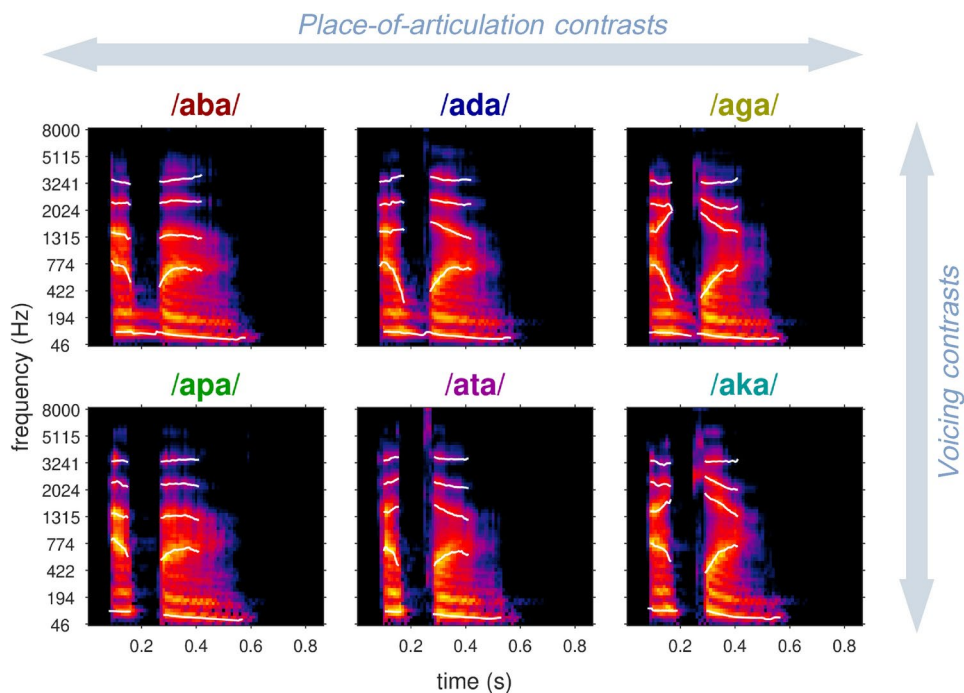


Fig. 1. Time-frequency representations of the six targets used in the experiment, organized by distinctive features. The time and frequency resolutions are the same as those used for the analysis. Warmer regions indicate higher amplitudes in a logarithmic scale. The white traces indicate the fundamental frequency (f_0) and formant trajectories (F1-F4).

total of 2000 of each of the two targets were presented. The order of the trials was randomized across the whole experiment.

After a correct or incorrect response, the SNR of the next target word was decreased or increased, respectively, following a weighted one-up one-down weighted adaptation rule targeting the 70.7% correct point of the psychometric function. Following each trial, participants received feedback on the correctness of their answer and on their average response bias over the last 100 trials. They were explicitly instructed to minimize their response bias as much as possible, as the ACI method is more sensitive for participants who are less biased.

The actual data collection was preceded by a short training to make sure that the participant correctly understood the task. This “warm-up” session, of a duration left to the participant’s appreciation, was similar to the main experiment except for three additional buttons allowing repeating the noisy speech stimuli or listening to the target sounds in silence. The results of this training was excluded from any further analysis.

Participants

For each condition, the data from seven participants was collected, yielding a total of 49 datasets. Participants were native French speakers aged 18 or more, with at least one normal hearing ear. This was evaluated by an audiometric test carried during the first test session (see Supplementary Fig. S1). Only participants with normal-hearing thresholds, i.e., with pure-tone average thresholds equal to or lower (better) than 20 dB HL in their best ear, were retained for the main experiment. Additionally, the results of the first experimental session were used to check that participants are able to perform the task. Participants who did not reach a SNR threshold of -11.5 dB or less in the first session were offered to make a second, then a third try. If they failed again to reach the expected minimum threshold, they were rejected from the analysis. All participants provided written informed consent before participating in the experiment. They were paid on an hourly basis (10 EUR/h).

The data was already collected for 6 participants in the ABDA condition and published in a previous article³¹. Since the primary focus of this investigation was not to compare conditions but rather to examine each condition separately, participants were permitted to perform multiple conditions if desired. Three authors of the study (LV, GC and CC) were also included as participants (S01, S13 and S25, respectively).

Analysis

Performance assessment

Various behavioral metrics were computed from each set of data to assess the overall effect of noise on the listeners’ performance, including correct response rate, SNR threshold, discriminability index and criterion. SNR thresholds were calculated as the median of all reversals in each block of 400 trials, excluding the first four reversals.

Auditory classification images

Each dataset (one participant in one condition) was processed to obtain one individual ACI.

Two criteria were used to select the trials included in the ACI calculation. In each experimental block of 400 trials, the segment corresponding to the initial convergence of the psychophysical staircase was excluded by rejecting the trials up to the fourth reversal. Subsequently, a number of trials with the most extreme SNRs were removed to balance the occurrence of the two alternative responses in the dataset. These criteria were implemented in Osses et al.³¹ to improve the statistical power of the method. They led to the rejection of ~ 10% of the trials. The remaining participant’s responses were then processed to obtain an individual ACI.

The time-frequency representation of the noise presented in trial i was obtained using a Gammatone auditory filterbank and denoted \underline{N}_i . The corresponding response of the listener (“target 1” or “target 2”) is denoted r_i . The ACI is a time-frequency matrix of “decision weights” used by the listener to identify the presented phoneme. This matrix of weights, denoted \underline{ACI} , was estimated through a Generalized Linear Model, relating the exact acoustic content of the noise in a given trial to the corresponding response of the participant, according to Equation 1:

$$P(r_i = \text{target1}) = \Phi(\underline{N}_i \circ \underline{ACI} + c) \quad (1)$$

where \circ correspond to the element-wise matrix product, and c is an additional model weight representing the participant’s general bias on the considered dataset (by construction, this weight is close to zero). Following the method outlined by Osses et al.³¹, this statistical model was adjusted with a lasso regularization on a Gaussian-pyramid basis. This fitting approach is implemented in the fastACI toolbox under the name `l1glm`.

The above-described analysis pipeline resulted in an individual ACI for each dataset. The significance of each ACI was assessed based on the accuracy of their predictions. In the case of generalized linear models, accuracy is typically reported in terms of deviance, which intuitively measures the deviation from a perfect fit. Here, we report the “mean deviance” obtained following a 10-fold cross-validation procedure (out-of-sample predictive performance of the fitted statistical model described in Equation 1). This deviance is expressed relative to the mean cross-validated deviance of a “zero-weight ACI”, following the same procedure as Osses et al.³¹ Therefore, any deviance value significantly smaller than zero reflects predictions significantly better than chance. This prediction performance metrics was first obtained for each individual ACI (within-subject deviance), then also in cross-predictions between participants engaged in the same condition (between-subject deviance). For this purpose, the ACI of each participant was used to predict the data of each other participant in the same condition. These between-subject predictions offer a way to quantify the similarity among listening strategies within a condition.

The focus of this study being on the exact composition of the ACIs, we need to assess the significance of each perceptual weight. The lasso regression approach can already be seen as a form of “variable selection”

where only the relevant variables are associated with non-zero weights. However, we want to confirm this using a significance level threshold. This was done using a randomization approach: a level of weights corresponding to the null hypothesis of no effect was derived by randomizing the responses of the participants and deriving an ACI corresponding to these random responses. This process was iterated 1000 times to obtain a distribution of weights under the null hypothesis. Then the amplitude of the measured weights was compared to this null distribution to determine if they had less than 5% of chance to occur by chance.

Given the small sample size in each condition, no statistical test was performed on the ACIs at the group level. This approach focusing on individual results is in line with most of the reverse correlation literature³⁵, and with the experimental phonetics literature³⁶.

Results

In this experiment, six recordings from one single speaker were used in seven phoneme-categorization conditions corresponding to different pairs of stimuli: /aba/-/ada/, /ada/-/aga/, /apa/-/ata/, /ata/-/aka/, /ada/-/ata/, /aba/-/apa/, /aga/-/aka/. Each condition was completed by 7 participants, resulting in 49 sets of behavioral data.

The phonemic contrasts were tested using an adaptive procedure targeting the 70.7% point of the psychometric function. As a result, the average SNR was not the same across the different conditions, reflecting the fact that some discrimination tasks were more difficult than others. Participants performed particularly well in the [g]-[k] contrast (mean SNR = -17.1 dB) and the [d]-[t] contrast (mean SNR = -15.0 dB) while the other conditions resulted in SNRs between -12.9 dB and -13.9 dB (see Supplementary Fig. S2). An analysis of variance carried on the SNR thresholds per experimental block confirmed that there was no strong learning effect occurring over the course of the experiment (see Supplementary Analysis). In addition, the performances of the participants in the different conditions were characterized in terms of sensitivity and criterion as a function of SNR. These preregistered analyses are reported in the Supplementary Materials.

One ACI was obtained for each participant and each condition. Each of the 49 individual ACI therefore corresponded to a GLM adjusted to a separate dataset (see Supplementary Materials). Figure 2 presents the averaged ACIs per condition, with the formant and f_0 trajectories superimposed in order to facilitate interpretation.

For each individual ACI, the quality of the fit was measured using the cross-validated within-subject deviance metrics (depicted as blue dots in Fig. 3) expressed relative to a baseline corresponding to a “null” ACI where all weights are set to zero (dotted line). Remarkably, only two out of the 49 ACIs failed to produce predictions significantly better than chance. Additionally, cross-validated deviance was computed across participants (between-subject deviance) to assess the similarity between individual ACIs within each condition. If two participants use similar listening strategies, the data from one participant should be predicted equally well using either their own ACI or the ACI of the other participant^{31,37}. On the contrary, any significant gap between the within-subject and between-subject accuracy indicates that the weighting patterns within the considered ACIs are substantially different. The red dots in each panel of Fig. 3 represent how well the data from this participant is predicted by the other participants’ ACIs, on average. This analysis demonstrates that a majority of participants (33 out of 49 datasets) exhibited a between-subject deviance significantly higher (i.e., worse) than their within-subject deviance. This indicates that, within each condition, the participants employed listening strategies that differ significantly from one another overall.

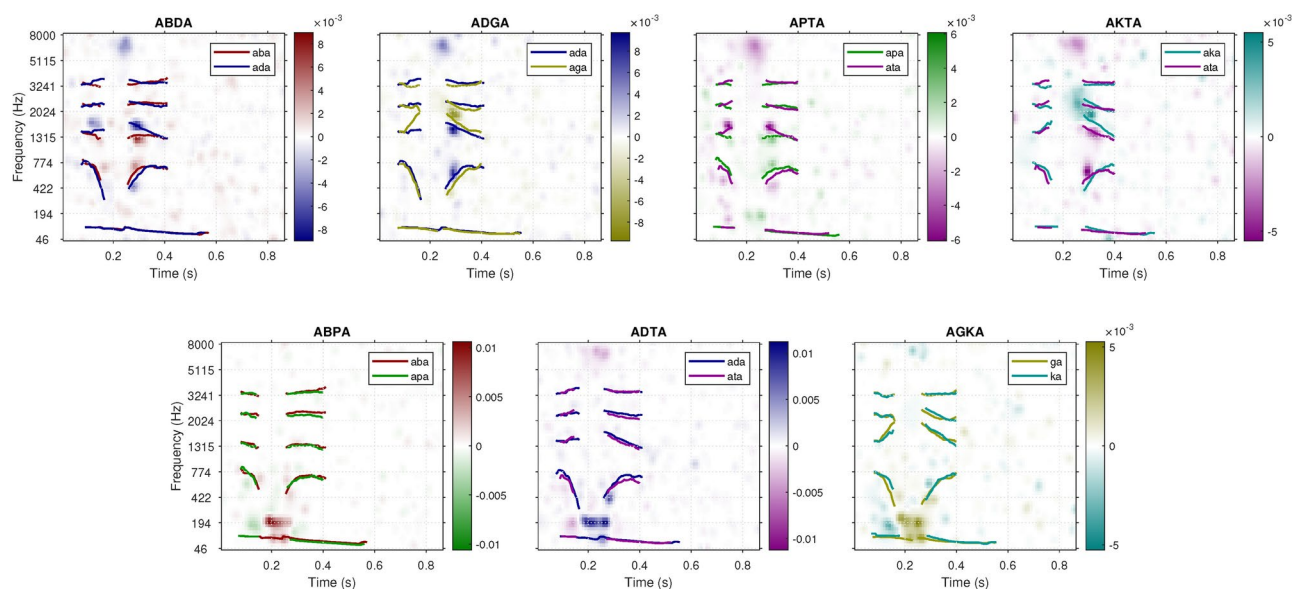


Fig. 2. Mean ACIs for all conditions. Top line: place of articulation contrasts. Bottom line: voicing contrasts. The colored lines correspond to the formant and f_0 trajectories for the two targets.

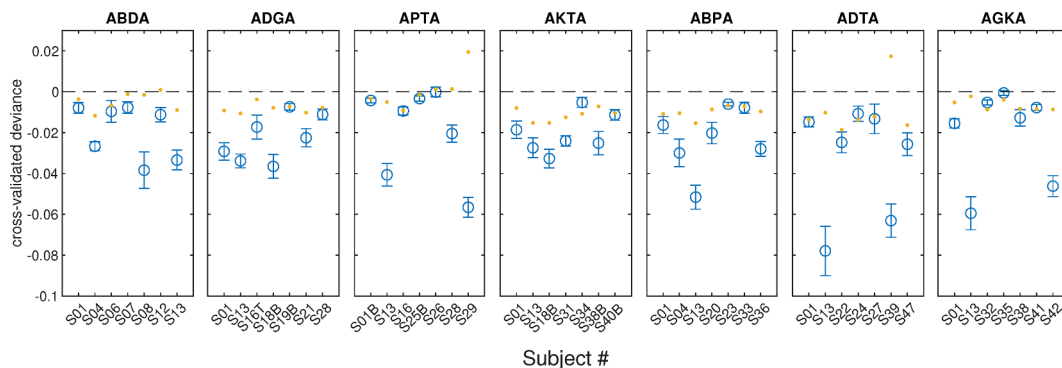


Fig. 3. Cross-validated within-subject deviance per trial, for each individual ACI (blue). Error bars represent the variability across cross-validation folds (± 1.64 SEM). The deviance is expressed relative to a baseline corresponding to the prediction accuracy of a zero-weight ACI (black dotted line). Therefore, negative deviance values indicate an improvement in prediction accuracy, compared to chance. The red dots correspond to the mean between-subject deviance. This value is obtained by predicting each participant’s data using the ACIs of the 6 other participants in the same condition.

| | Place-of-articulation constrasts | | | | Voicing constrasts | | |
|------------------|----------------------------------|------|------|------|--------------------|------|------|
| | ABDA | ADGA | APTA | AKTA | ABPA | ADTA | AGKA |
| CV F2 transition | 7/7 | 7/7 | 7/7 | 7/7 | 0/7 | 1/7 | 0/7 |
| CV F1 transition | 3/7 | 6/7 | 2/7 | 3/7 | 4/7 | 7/7 | 4/7 |
| Burst | 4/7 | 5/7 | 4/7 | 6/7 | 1/7 | 4/7 | 1/7 |
| Voicing bar | 0/7 | 1/7 | 2/7 | 0/7 | 7/7 | 7/7 | 6/7 |
| VC F2 transition | 4/7 | 0/7 | 3/7 | 0/7 | 0/7 | 1/7 | 0/7 |
| VC F1 transition | 1/7 | 0/7 | 3/7 | 0/7 | 3/7 | 2/7 | 2/7 |

Table 1. Overview of the cues found in individual ACIs. Number of ACIs for which a given cue was found significant, in each condition.

Visual inspection revealed that most of the ACIs corresponding to place-of-articulation contrasts showed a clear pattern of positive and negative weights on the F2 onset, organized vertically – indicating that listeners were performing a frequency discrimination. This was confirmed quantitatively as significant weights were found in the region of the F2 onset for all 28 place-of-articulation ACIs. Additionally, weaker weights were found on the F1 onset and on the location of the high-frequency release burst for the coronal stop (> 5 kHz) and the dorsal stop (~2.5 kHz). Furthermore, the labial vs. coronal contrasts appeared to involve some CV transition cues on the initial [a]. The voicing ACIs revealed a simpler pattern, with positive weights concentrated mainly in the region of the f0 during the intervocalic interval, and preceded by negative weights. Weak weights were also found on the F1 onset and on the release burst for /t/ and /k/. The significance of the weights was assessed at the individual level. In Table 1, we report the proportion of ACIs displaying significant weights in a time-frequency regions of interest (ROI) centered on the position of each potential cue. For the sake of simplicity, all ACIs were considered in this analysis, including the two that were not significant. The general pattern for each feature remains largely unchanged when considering only the significant ACIs.

Discussion

The goal of this study is to identify the specific acoustic cues on which individuals rely during phoneme discrimination tasks, with a focus on French stop consonants. By analyzing speech-in-noise recognition data through a reverse-correlation approach, we aim, first, to independently corroborate well-established cues already identified by the scientific community with a method based on natural speech rather than reduced/synthesized speech stimuli. Secondly, we seek to emphasize certain aspects of our findings that could contribute to elucidating and reconciling disagreements within the field.

All participants were able to perform the task at a very low SNR (mean SNR across all participants and conditions = -14.1 dB), with some variability between the easiest discrimination task (AGKA, mean SNR = -17.1 dB) and the most difficult ones (AKTA and ABPA, mean SNR = -12.9 dB). Although the experiment was quite long, no significant improvement in performance was observed over the course of the experiment, allowing us to pull all 4000 trials altogether for the ACI analysis. Furthermore, the three authors who were included as participants did not achieve significantly better performances than the naive listeners. In fact, in most cases, they were not even the top performers in their group, with the exception of S13 in condition ADTA. This comparison suggests that the involvement of experimenters in the sample did not bias the results, as their performance was comparable to that of the other participants.

For each dataset, corresponding to one participant in one condition, an ACI was derived by fitting a GLM on the trial-by-trial association between a particular noise instance and the corresponding behavioral response of the participant (e.g. /ada/ vs. /aga/). The significance of each ACI can be measured through the prediction accuracy of the statistical model³¹. Prediction accuracy was significantly above chance in 47 out of the 49 ACIs (blue circles in Fig. 3), confirming that the model successfully captured some aspects of the listeners' strategy (preregistered hypothesis 1). As the model gives a reliable account of the participants' pattern of behavior in the task, it is therefore legitimate to explore the model weights as an insight of the participant's perceptual strategy.

It is important to note, however, that not all individual ACIs exhibit identical weight patterns. In order to assess the homogeneity of listening strategies (preregistered hypothesis 3), we measured the between-subject accuracy between individual ACIs within each contrast. These findings reveal substantial between-participant differences, indicating that the listening strategies of the majority of participants (33 out of 44) are significantly distinct from one another. For example, S39 appears to rely mostly on high-frequency information (~ 8 kHz) contrary to other participants in the APTA condition (see Supplementary Fig. S3).

This variability cannot be explained by obvious differences in hearing threshold or linguistic background as we have selected a homogeneous group of normal-hearing native french speakers. Further study is needed to investigate whether this variability may arise from more subtle differences in hearing thresholds or second languages. Inter-individual variability on perceptual cues in phonological perception has already been documented^{38,39}. For example, it has been shown that there is an individual variability in the reliance on the VOT and f_0 onset cues for voicing perception in English^{13,40} and for stop perception in Korean⁴¹. In the next section, we describe the general shape of the ACIs in each contrast. The implications of the interindividual variability for the conclusions will be examined in the rest of the discussion.

Acoustic cues of French stop consonants

For place-of-articulation contrasts (ABDA, ADGA, APTA, AKTA, top panels of Fig. 2), the results are in general agreement with those obtained in prior ACI studies focusing on /aba/-/ada/^{31,32,42} and /alda/-/alga/-/arda/-/arga/^{37,43,44} categorization tasks. The ACIs show a clear pattern of positive and negative weights organized vertically and matching the location of the CV F2 transition for the corresponding targets (/aba/ and /apa/ at ~ 1300 Hz, /ada/ and /ata/ at ~ 1700 Hz, /aga/ and /aka/ at ~ 2000 Hz). The organisation of these weights indicates that when noise energy is present at the time-frequency location of one of the two targets' CV F2 transition, it influences perception towards that target. This is a marker of the role of the CV F2 transition for place-of-articulation contrast perception, as already shown by the literature on stop consonants using synthetic speech^{7,8,45}. In some, but not all, place-of-articulation ACIs, a similar pattern of weights is visible around the CV F1 onset. The role of the CV F1 transitions in stop consonant perception has been noted in the literature, albeit less frequently^{45–48}. In our study, this information was used by half of our participants (14 out of 28), therefore confirming that this cue may play a role in a context of speech-in-noise comprehension. As suggested by Ohde and Stevens, it could be the case that the perceptual property critical for place perception is the relationship between F2 and F1 at syllable onset⁴⁶. Symmetrically, positive and negative weights were found on the initial /a/ segment in the labial vs. coronal contrasts (ABDA and APTA), corresponding to the VC F2 transition (for 7 out of the 14 participants) and VC F1 transition (for 4 participant only). VC cues are generally weighted less heavily than the corresponding CV cues, consistent with the notion that the former play a minor role in place perception²². In the coronal vs. dorsal contrasts (ADGA and ATKA) these cues are absent, possibly because VC formant transition frequencies show a larger overlap (and therefore a less robust encoding) for back places of articulation⁴⁹. Finally, most place-of-articulation ACIs also reveal a high-frequency cluster of weights at syllable onset, cueing for the detection of [d] and [t]. This indicates that a majority of participants rely on the detection of a burst of energy at the release of the consonant to identify dental plosives, as highlighted by numerous prior studies²³. There are still ongoing discussions to arbitrate whether F2-transition cues and burst cues should be considered necessary and sufficient cues^{23,45,47}. We go into more detail about this matter in the next section.

For voicing contrasts (ABPA, ADTA, AGKA), 20 out of 21 ACIs show weights in the region of the f_0 and the first harmonic (~ 100 and 200 Hz, respectively), during the intervocalic interval. This low-frequency region corresponds to the voicing bar, which is characteristic of voiced sounds. It is therefore expected to find it as a cue for identification of the voiced stop consonant ([b], [d] and [g]). Additionally, some of the voicing ACIs display weights organized vertically on the onset of F1, indicating a CV F1 transition cue, and weights in the high frequencies for [t] and [k] correspond to release burst cues. Both of these cues have also already been mentioned by the literature on voiced stop consonants^{26,27,50,51}.

As detailed above, the acoustic cues revealed by the ACI method closely align with the phonetic literature. The present study therefore confirms the role of these cues for perception of natural (i.e. non-manipulated) speech recordings in noise. However, some previously noted acoustic cues are absent from our data. This is particularly striking in the case of voicing: while some authors have catalogued no less than 16 cues distinguishing between [b] and [p]²⁷, our ACIs for the ABPA contrast revealed only one cue shared by all participants. In particular, the f_0 transitions do not seem to be used by any of 21 listeners engaged in the voicing discrimination task. Similarly, the F3 transition and the voice onset time, highlighted as cueing for place perception in past studies^{7,8,45}, do not correspond to clear weights in any of our place-of-articulation ACIs.

Several hypotheses could explain the absence of these cues. First, this might be due to the specific set of recordings used in the experiment, which consisted of one single utterance from a single speaker for each of the six target consonants. However, a crucial feature of the ACI method is that it allows to visualize the cues expected by the participant, even when these cues are not actually present in the stimuli^{32,52}. Therefore, we believe that the choice of a limited number of utterances per target does not restrict the generality of our results. Further arguments supporting this point are provided in the Supplementary Materials. Another factor that may limit the generality of our conclusions is the choice of a single [aCa] context and a particular language. For instance,

since English voiceless stops in this position are typically aspirated, English-speaking listeners might use burst intensity as a voicing cue more systematically than our French-speaking participants. Further studies will be dedicated to extend these results to other contexts and languages.

Second, the method has a limited time-frequency resolution which may restrict the range of observable cues. For instance, a voicing cue on the f_0 CV transition may correspond to a region too small to be detected by the lasso regularization on a Gaussian-pyramid basis. Another situation could be that the cues are too close to each other, or even integrated into a single perceptual property. This has been shown to be the case in voicing perception for the F1 and f_0 transitions with the voicing bar²⁸ and in place perception for the burst and formant transition⁵³.

A third explanation involves the interindividual variability observed in our data. Although we initially hypothesized that the individual ACIs of participants in each phonetic contrast would be generally similar, it turns out that some cues are used by all participants (the F2 onset and the voicing bar for the place-of-articulation contrasts and the voicing contrasts, respectively), while other cues are only found in a subset of participants. Consequently, it is theoretically possible that we have not captured the complete range of cues that are used for these contrasts. This, however, is very unlikely: although the sample size per contrast is small, the phonologically orthogonal experimental design minimizes the risk of missing significant cues because they are expected to appear across several contrasts.

A last potential explanation deserves consideration. Most of the psycholinguistic studies discussed so far focus on phoneme perception in quiet, while the present paper is interested in speech-in-noise perception. Accordingly, the experiment is designed to revealing cues relevant to perception in the presence of a white-noise-like background. Given the very low SNR at which participants performed the task, the ACIs likely represent only the most noise-resistant cues^{24,47,54,55}. Interestingly, from this perspective, the number of cues revealed by the place-of-articulation ACIs appears to be surprisingly large, compared to previous studies on phoneme-in-noise perception. For instance, according to Régner & Allen, recognition of [t] in white noise is based almost exclusively on the audibility of the release burst. Conversely, Alwan et al. conclude that formant transition cues become increasingly important for discrimination between labial and alveolar consonants as SNR degrades. The ACIs measured in the present study indicate that, for place of articulation, the multiplicity of cues involved in the decision is preserved up to very low SNRs.

Cue weighting

Despite a widespread agreement on the existence of specific acoustic cues in consonant perception, there remains ongoing debates about the effective use and the relative importance of these cues in the recognition process. In the context of place perception, for instance, some researchers argue that F2 is the primary cue, supplemented by a secondary burst cue^{7,8}. While acknowledging that both the release burst and formant transitions alone can serve as sufficient cues for place identification¹⁷, they stress that the formant transitions tend to dominate in place-of-articulation perception^{29,30}. On the contrary, other authors contend that the burst is the predominant cue, or even for some phonemes like [t], the only cue used for their recognition²³. According to their findings, this cue is context-independent and essential for accurate perception of all stop consonants^{6,19,23,25}. With respect to the relative importance of F2 and the burst cue in place perception, the ACIs measured in the present study reveal that both cues are present in all contrasts (Fig. 2, top line). This indicates that for natural speech perception in noise, perception of place is not conveyed by a single critical acoustic feature.

However, it is essential to understand these disagreements within their methodological context. The debate surrounding the use and dominance of acoustic cues in speech perception is deeply influenced by advancements in research techniques and methodologies. Historically, most experimental methods were aimed at identifying the necessary and sufficient acoustic cues for phoneme recognition^{6,36}. A necessary and sufficient cue is defined experimentally as an acoustic property whose presence/absence in the signal dramatically impacts the recognition of the sound: when the cue is removed by filtering¹, truncation¹⁴, or selectively manipulated^{8,25}, phoneme identification is impaired, leading to the sound either being unrecognized or misidentified as a different phoneme²³. As described in the introduction, these methods have been successfully used to reveal the necessary and sufficient cues in a number of phonemes. However, by design, they are not suited for investigating a situation where many cues contribute to the phonetic decision. Even the cue-trading paradigm, widely used for the estimation of the weighting of separate cues, becomes extremely time-consuming when more than two cues are involved⁵⁶ (not to mention that it requires the cues to be known beforehand, making it unsuitable for exploratory research). Similarly, for the 3-dimensional-deep-search method, the presence of multiple, possibly conflicting, cues usually yields complicated patterns of behavior that require difficult interpretation work. As a consequence, these studies have sought for a description of the phonetic decision process based on a minimal number of primary cues, usually one or two, combined with secondary cues, either optional or contextual⁵⁰. Although this conceptual framework has proved useful, it has also led to deadlocked controversies, such as the F2 vs. burst debate.

In contrast, recent theoretical considerations have argued that, depending on the research questions, it may make sense to talk about primary and contextual cues in a parallel way, as they both constitute sources of information that individuals weigh differently⁵⁰, or even that there is no objective basis for a distinction between these two types of cues⁵⁷. The innovative feature of the ACI method for acoustic cues investigation is to allow a data-driven exploration of the information used by a listener when discriminating natural speech sounds in noise. Critically, it does not rely on the notion of necessary and sufficient cue. Rather, the method returns a spectrotemporal map of the regions where the presence of noise influenced perception in a systematic way. Any information contributing to the decision can theoretically be identified in this way, even though this information is not critically needed for the identification of the phoneme. Furthermore, this approach involves a minimum number of assumptions on the nature and number of cues being sought. In the present study, a majority of the

ACIs (40 out of 49) reveal the use of more than a single cue, with some listeners relying on at least 4 cues for a given contrast (e.g. S13 in condition APTA), confirming that the phoneme recognition process is not based on the detection of a single acoustic feature but rather involves the analysis of complex spectrotemporal patterns.

As noted above, one striking aspect of our results is the presence of a significant interindividual variability in the pattern of cues extracted by each listener in a given task. This variability is not entirely random; rather, it exhibits a structured pattern that provides insights into the underlying processes. In particular, the F2 CV transition is used by all 28 listeners engaged in a place categorization task while the burst cue is found only in a smaller subset of 19 individuals (Table 1). Therefore, although the ACI method does not formally distinguish between primary and secondary cues, our results indicate that F2 may be said to be the dominant cue over the burst cue for place perception in white noise because it is more prevalent within the tested participants. Furthermore, the burst cue can hardly be considered a “necessary” cue for this contrast, contrary to the claim of some authors²³, because some listeners appear to not use this cue at all. Similarly, despite the presence of interindividual differences, the voicing bar cue appears in all individual voicing ACIs, except the one that did not reach significance. This was not the case for the burst cue or the CV F1 transition.

As a conclusion, we think that the use of ACIs offers an appealing methodological avenue for psycholinguistics: in showing in a fine-grained manner, and with only few theoretical assumptions, the information that a particular listener uses in a given task, the ACI approach may be able to reconcile the historical research that has focused on the identification of necessary and sufficient cues and the growing body of work investigating individual differences in the perception of speech sounds.

Data availability

This study—including hypotheses, stimuli and procedures – was preregistered on OSF (<https://osf.io/fqejt>). All datasets are available within a Zenodo repository: <https://doi.org/10.5281/zenodo.11507060>. The experiment and all analyses can be replicated using the fastACI MATLAB Toolbox, available on Github: <https://github.com/aosses-tue/fastACI>. A more detailed procedure is provided in the Supplementary Materials.

Received: 27 June 2024; Accepted: 23 October 2024

Published online: 08 November 2024

References

- Fletcher, H. The nature of speech and its interpretation. *J. Franklin Inst.* **193**, 729–747. [https://doi.org/10.1016/S0016-0032\(22\)90319-9](https://doi.org/10.1016/S0016-0032(22)90319-9) (1922).
- French, N. R. & Steinberg, J. C. Factors governing the intelligibility of speech sounds. *J. Acoust. Soc. Am.* **19**, 90–119. <https://doi.org/10.1121/1.1916407> (1947).
- Allen, J. B. *Articulation and Intelligibility* (Morgan & Claypool Publishers, 2006).
- Miller, G. A. & Nicely, P. E. An analysis of perceptual confusions among some english consonants. *J. Acoust. Soc. Am.* **27**, 338–352. <https://doi.org/10.1121/1.1907526> (1955).
- Allen, J. B. Consonant recognition and the articulation index. *J. Acoust. Soc. Am.* **117**, 2212–2223 (2005).
- Singh, R. & Allen, J. B. The influence of stop consonants' perceptual features on the Articulation Index model. *J. Acoust. Soc. Am.* **131**, 3051–3068. <https://doi.org/10.1121/1.3682054> (2012).
- Lieberman, A. M., Delattre, P. C. & Cooper, F. S. The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *Am. J. Psychol.* **65**, 497–516 (1952).
- Lieberman, A. M., Delattre, P. C., Cooper, F. S. & Gerstman, L. J. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychol. Monogr. Gen. Appl.* **68**, 1–13. <https://doi.org/10.1037/h0093673> (1954).
- Delattre, P. From acoustic cues to distinctive features. *Phonetica* **18**, 198–230. <https://doi.org/10.1159/000258610> (1968).
- Repp, B. H. Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychol. Bull.* **92**, 81–110. <https://doi.org/10.1037/0033-2909.92.1.81> (1982).
- Varnet, L., Langlet, C., Lorenzi, C., Lazard, D. S. & Micheyl, C. High-frequency sensorineural hearing loss alters cue-weighting strategies for discriminating stop consonants in noise. *Trends Hear.* [SPACE] <https://doi.org/10.1177/2331216519886707> (2019).
- Clayards, M. Differences in cue weights for speech perception are correlated for individuals within and across contrasts. *J. Acoust. Soc. Am.* **144**, EL172–EL177. <https://doi.org/10.1121/1.5052025> (2018).
- Shultz, A. A., Francis, A. L. & Llanos, F. Differential cue weighting in perception and production of consonant voicing. *J. Acoust. Soc. Am.* **132**, 95–101. <https://doi.org/10.1121/1.4736711> (2012).
- Schatz, C. D. The role of context in the perception of stops. *Language* **30**, 47–56. <https://doi.org/10.2307/410219> (1954).
- Halle, M., Hughes, G. W. & Radley, J.-P.A. Acoustic properties of stop consonants. *J. Acoust. Soc. Am.* **29**, 107–116. <https://doi.org/10.1121/1.1908634> (1957).
- Ohman, S. E. Perception of segments of VCCV utterances. *J. Acoust. Soc. Am.* **40**, 979–988. <https://doi.org/10.1121/1.1910222> (1966).
- Dorman, M. F., Studdert-Kennedy, M. & Raphael, L. J. Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Percept. Psychophys.* **22**, 109–122. <https://doi.org/10.3758/BF03198744> (1977).
- Malécot, A. The role of releases in the identification of released final stops: A series of tape-cutting experiments. *Language* **34**, 370–380. <https://doi.org/10.2307/410929> (1958).
- Cole, R. A. & Scott, B. The phantom in the phoneme: Invariant cues for stop consonants. *Percept. Psychophys.* **15**, 101–107. <https://doi.org/10.3758/BF03205836> (1974).
- Winitz, H., Scheib, M. E. & Reeds, J. A. Identification of stops and vowels for the burst portion of (p, t, k) isolated from conversational speech. *J. Acoust. Soc. Am.* **51**, 1309–1317. <https://doi.org/10.1121/1.1912976> (1972).
- Warren, P. & Marslen-Wilson, W. Continuous uptake of acoustic cues in spoken word recognition. *Percept. Psychophys.* **41**, 262–275. <https://doi.org/10.3758/BF03208224> (1987).
- Fujimura, O., Macchi, M. J. & Streeter, L. A. Perception of stop consonants with conflicting transitional cues: a cross-linguistic study. *Lang. Speech* **21**, 337–346. <https://doi.org/10.1177/002383097802100408> (1978).
- Li, F., Menon, A. & Allen, J. B. A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. *J. Acoust. Soc. Am.* **127**, 2599–2610. <https://doi.org/10.1121/1.3295689> (2010).
- Régner, M. S. & Allen, J. B. A method to identify noise-robust perceptual features: application for consonant /t/. *J. Acoust. Soc. Am.* **123**, 2801–2814. <https://doi.org/10.1121/1.2897915> (2008).
- Li, F. & Allen, J. B. Manipulation of consonants in natural speech. *IEEE Trans. Acoust. Speech Signal Process.* **19**, 496–504. <https://doi.org/10.1109/TASL.2010.2050731> (2011).

26. Wright, R. A review of perceptual cues and cue robustness. In Hayes, B., Steriade, D. & Kirchner, R. (eds.) *Phonetically Based Phonology*, 34–57. <https://doi.org/10.1017/CBO9780511486401.002> (Cambridge University Press, Cambridge, 2004).
27. Lisker, L. “Voicing” in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language Speech* **29**, 3–11. <https://doi.org/10.1177/002383098602900102> (1986).
28. Kingston, J., Diehl, R. L., Kirk, C. J. & Castleman, W. A. On the internal perceptual structure of distinctive features: The [voice] contrast. *J. Phon.* **36**, 28–54. <https://doi.org/10.1016/j.wocn.2007.02.001> (2008).
29. Walley, A. C. & Carrell, T. D. Onset spectra and formant transitions in the adult’s and child’s perception of place of articulation in stop consonants. *J. Acoust. Soc. Am.* **73**, 1011–1022. <https://doi.org/10.1121/1.389149> (1983).
30. Kewley-Port, D., Pisoni, D. B. & Studdert-Kennedy, M. Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *J. Acoust. Soc. Am.* **73**, 1779–1793. <https://doi.org/10.1121/1.389402> (1983).
31. Osses, A. & Varnet, L. A microscopic investigation of the effect of random envelope fluctuations on phoneme-in-noise perception. *J. Acoust. Soc. Am.* **155**, 1469–1485. <https://doi.org/10.1101/2022.12.27.522040> (2024).
32. Varnet, L., Knoblauch, K., Meunier, F. & Hoen, M. Using auditory classification images for the identification of fine acoustic cues used in speech perception. *Front. Hum. Neurosci.* **7**, 865 (2013).
33. Osses, A. & Varnet, L. fastACI toolbox: the MATLAB toolbox for investigating auditory perception using reverse correlation (v1.0). doi:10.5281/zenodo.5500139 (2021).
34. Wesker, T. et al. Oldenburg logatome speech corpus (OLLO) for speech recognition experiments with humans and machines. In *Proceedings of Interspeech*, 1273–1276 (2005).
35. Varnet, L. & Lorenzi, C. Probing temporal modulation detection in white noise using intrinsic envelope fluctuations: A reverse-correlation study. *J. Acoust. Soc. Am.* **151**, 1353–1366. <https://doi.org/10.1121/10.0009629> (2022).
36. Li, F., Trevino, A., Menon, A. & Allen, J. B. A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise. *J. Acoust. Soc. Am.* **132**, 2663–2675. <https://doi.org/10.1121/1.4747008> (2012).
37. Varnet, L., Meunier, F., Trollé, G. & Hoen, M. Direct viewing of dyslexics’ compensatory strategies in speech in noise using auditory classification images. *PLoS ONE* **11**, e0153781. <https://doi.org/10.1371/journal.pone.0153781> (2016).
38. Hazan, V. & Rosen, S. Individual variability in the perception of cues to place contrasts in initial stops. *Percept. Psychophys.* **49**, 187–200. <https://doi.org/10.3758/BF03205038> (1991).
39. Yu, A. C. L. & Zöllou, G. Individual differences in language processing: Phonology. *Ann. Rev. Linguist.* **5**, 131–150. <https://doi.org/10.1146/annurev-linguistics-011516-033815> (2019).
40. Kong, E. J. & Edwards, J. Individual differences in categorical perception of speech: Cue weighting and executive function. *J. Phon.* **59**, 40–57. <https://doi.org/10.1016/j.wocn.2016.08.006> (2016).
41. Schertz, J., Cho, T., Lotto, A. & Warner, N. Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *J. Phon.* **52**, 183–204. <https://doi.org/10.1016/j.wocn.2015.07.003> (2015).
42. Osses, A. & Varnet, L. Consonant-in-noise discrimination using an auditory model with different speech-based decision devices. In *DAGA, 47th Annual Conference on Acoustics DAGA 2021 Wien*, 298–301 (Vienne, Austria, 2021).
43. Varnet, L., Knoblauch, K., Serniclaes, W., Meunier, F. & Hoen, M. A psychophysical imaging method evidencing auditory cue extraction during speech perception: A group analysis of auditory classification images. *PLoS ONE* **10**, e0118009 (2015).
44. Varnet, L., Wang, T., Peter, C., Meunier, F. & Hoen, M. How musical expertise shapes speech perception: Evidence from auditory classification images. *Sci. Rep.* **5**, 14489. <https://doi.org/10.1038/srep14489> (2015).
45. Delattre, P. C., Liberman, A. M. & Cooper, F. S. Acoustic loci and transitional cues for consonants. *J. Acoust. Soc. Am.* **27**, 769–773. <https://doi.org/10.1121/1.1908024> (1955).
46. Ohde, R. N. & Stevens, K. N. Effect of burst amplitude on the perception of stop consonant place of articulation. *J. Acoust. Soc. Am.* **74**, 706–714 (1983).
47. Alwan, A., Jiang, J. & Chen, W. Perception of place of articulation for plosives and fricatives in noise. *Speech Commun.* **53**, 195–209. <https://doi.org/10.1016/j.specom.2010.09.001> (2011).
48. Storme, B. Contrast enhancement as motivation for closed syllable laxing and open syllable tensing. *Phonology* **36**, 303–340. <https://doi.org/10.1017/S0952675719000149> (2019).
49. Ohman, S. E. Coarticulation in VCV utterances: Spectrographic measurements. *J. Acoust. Soc. Am.* **39**, 151–168. <https://doi.org/10.1121/1.1909864> (1966).
50. Schertz, J. & Clare, E. J. Phonetic cue weighting in perception and production. *WIREs Cogn. Sci.* **11**, e1521. <https://doi.org/10.1002/wcs.1521> (2020).
51. Repp, B. H. Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. *Lang. Speech* **22**, 173–189. <https://doi.org/10.1177/002383097902200207> (1979).
52. Gosselin, F. & Schyns, P. G. RAP: A new framework for visual categorization. *Trends Cogn. Sci.* **6**, 70–77. [https://doi.org/10.1016/1364-6613\(00\)01838-6](https://doi.org/10.1016/1364-6613(00)01838-6) (2002).
53. Stevens, K. N. & Blumstein, S. E. Invariant cues for place of articulation in stop consonants. *J. Acoust. Soc. Am.* **64**, 1358–1368. <https://doi.org/10.1121/1.382102> (1978).
54. Serniclaes, W. & Arrouas, Y. Perception des traits phonétiques dans le bruit. *Verbum* 131–144 (1995).
55. Phatak, S. A., Lovitt, A. & Allen, J. B. Consonant confusions in white noise. *J. Acoust. Soc. Am.* **124**, 1220–1233. <https://doi.org/10.1121/1.2913251> (2008).
56. Bailey, P. J. & Summerfield, Q. Information in speech: Observations on the perception of [s]-stop clusters. *J. Exp. Psychol. Hum. Percept. Perform.* **6**, 536–563. <https://doi.org/10.1037/0096-1523.6.3.536> (1980).
57. Port, R. How are words stored in memory? Beyond phones and phonemes. *New Ideas Psychol.* **25**, 143–170. <https://doi.org/10.1016/j.newideapsych.2007.02.001> (2007).

Acknowledgements

This study was supported by the French National Research agency through the ANR grants “fastACI” (Grant No. ANR-20-CE28-0004), “DIA-SYN-PHON” (Grant No. ANR-21-CE28-0008) and “FrontCog” (Grant No. ANR-17-EURE-0017).

Author contributions

G.C., M.G. and L.V. conceived the experiment. G.C., L.V., C.C. and P.F. collected the data. L.V. analysed the results and ran the simulations. L.V. and G.C. wrote the manuscript. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-77634-w>.

Correspondence and requests for materials should be addressed to L.V.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024