Prosodic cues to word boundaries in a segmentation task assessed using reverse correlation

Alejandro Osses, Elsa Spinelli, Fanny Meunier, Etienne Gaudrain, Léo Varnet

JASA EXPRESS LETTERS ARTICLE pubs.aip.org/asa/jel

Prosodic cues to word boundaries in a segmentation task assessed using reverse correlation

Alejandro Osses,[1,a)] Elsa Spinelli,[2] Fanny Meunier,[3] Etienne Gaudrain,[4] and Léo Varnet[1]

[1]*Laboratoire des Systèmes Perceptifs, Département d'Études Cognitives, École Normale Supérieure, PSL University, CNRS, Paris, France*
[2]*Laboratoire de Psychologie et NeuroCognition, Université Grenoble Alpes, Grenoble, France*
[3]*Université Côte d'Azur, CNRS, BCL, Nice, France*
[4]*Lyon Neuroscience Research Center, CNRS, Inserm, Université Lyon 1, Lyon, France*

ale.a.osses@gmail.com, elsa.spinelli@univ-grenoble-alpes.fr, fanny.meunier@unice.fr, etienne.gaudrain@cnrs.fr, leo.varnet@ens.psl.eu

Abstract: When listening to speech sounds, listeners are able to exploit acoustic features that mark the boundaries between successive words, the so-called segmentation cues. These cues are typically investigated by directly manipulating features that are hypothetically related to segmentation. The current study uses a different approach based on reverse correlation, where the stimulus manipulations are based on minimal assumptions. The method was evaluated using pairs of phonemically identical sentences in French, whose prosody was changed by introducing random f_0 trajectories and segment durations. Our results support a prominent perceptual role of the f_0 rise and vowel duration at the beginning of content words.

[Editor: Douglas D. O'Shaughnessy] https://doi.org/10.1121/10.0021022

Received: 8 June 2023 **Accepted:** 3 September 2023 **Published Online:** 26 September 2023

1. Introduction

A substantial part of the speech perception literature is devoted to the identification of perceptual cues that are conveyed in speech sounds and decoded by the human auditory system. In particular, segmentation cues correspond to the acoustic features that mark the boundaries between successive words in a continuous sound stream. The present study reports an experimental method for assessing these cues, considering only minimal assumptions about their location.

Several sources of information are used by the human brain to find the boundaries between words. In particular, the phonetic and lexical contexts undoubtedly play a role in this process. In the extreme case, lexical competition models such as TRACE[1] consider segmentation as a mere by-product of lexical access, where the selected set of boundaries is the one that forms a sequence of valid words. At the phonetic level, phonotactic rules impose constraints on the position of word boundaries in a sequence of phonemes. For example, /mr/ and /ls/ are not permissible sequences at the onset of words in Dutch.[2] However, the lexical and phonetic levels are not sufficient to account for the ability of humans to segment speech. For instance, listeners are able to discriminate between phonemically identical sentences, such as "c'est l'amie" and "c'est la mie" in French[3] or "known ocean" and "no notion" in English,[4] showing that they can also exploit fine-grained acoustic cues available in the speech sounds.

The acoustic features correlated with the presence of word boundaries are language-specific and multidimensional. In French, they are mainly related to the fundamental frequency (f_0) trajectories and the duration of syllables. Due to the multiple and partly redundant features conveyed in natural speech, researchers have mainly followed a hypothesis-driven approach to identify the segmentation cues, using carefully controlled stimuli where one single cue is varied at a time.

The study by Banel and Bacri[5] is an example of such an approach. Separately recorded monosyllables were assembled together into ambiguous disyllabic stimuli like /ba.ga3/. The authors found that participants were more likely to hear one ("bagage") or two words ("bas gage") depending on whether the second or the first syllable was lengthened, respectively. This result suggests that listeners are sensitive to segmental duration and use this information to modulate the lexical interpretation of spoken French. However, the use of concatenated monosyllables does not allow us to dissociate the perceptual role of segmental duration from the role of other covarying cues. More recently, Shoemaker[6] tackled this limitation by using stimuli where only the duration dimension was manipulated, while ensuring that all other stimulus

a)Author to whom correspondence should be addressed.

features remained unchanged. She confirmed the previous finding that French listeners can exploit segmental duration as a cue for the interpretation of ambiguous phonemically identical sentences.

In parallel, other researchers have studied the role of the f_0 trajectory in the segmentation of an input stream into words. Based on the study of pitch patterns from recorded utterances, sophisticated theories have been developed to explain the complex structure of f_0 fluctuations in French.[7] The role of these hypothetical cues during speech comprehension was then confirmed using perceptual experiments carried with manipulated stimuli. In particular, Spinelli *et al.*[8] tested the use of a word-initial ("early") f_0 rise as a segmentation cue by native French listeners. Based on ambiguous phoneme sequences arising from elision, such as /se.la.mi/ ("c'est la mie" or "c'est l'amie"), they generated new stimuli where the mean f_0 on the /a/ segment was artificially modified. This variable was shown to influence the perceived segmentation of the sentence, with a higher f_0 leading to more "c'est l'amie" answers. This result supported the view of the early f_0 rise as a cue to a content word beginning. A similar finding had been obtained by Welby using nonsensical sentences,[9] where the manipulation process was applied to the whole f_0 contour of the stimuli. Consistent with the notion of early f_0 rise, Welby demonstrated that listeners interpreted nonsense sequences like /me.la.mɔ̃.din/ as a single nonword ("mélamondine") when the f_0 rise occurred in the first syllable and as a determiner followed by a nonword ("mes lamondines") when the rise occurred in the second syllable. However, that study also indicated that the early f_0 rise was not a necessary cue and that a simple inflection or "early elbow" in the f_0 contour was also interpreted as a marker for word beginning.

The above-mentioned studies employed a variety of psycholinguistic methods to investigate the role of different cues in the segmentation process, including cross-splicing,[3–5] duration manipulation,[6] and manipulation of the f_0 information in one phonetic segment[8] or in more complex f_0 contours.[9] An argument in favor of these manipulation-based experiments is that the speech features can be carefully controlled. However, the major disadvantage is that they require prior knowledge of the specific cues that can play a role in the segmentation process. Consequently, they cannot probe the large space of all possible cues available to the listeners, undermining the possibility to investigate potential interactions between cues.

In the present study, we introduce a new method for the investigation of segmentation cues in perception, based on the reverse-correlation (revcorr) paradigm. Contrary to the previously described experiments, the proposed method offers a comprehensive approach to explore the cues to word boundaries, where the test stimuli are processed with minimal assumptions. In this sense, the method is efficient, because it allows the investigation of several acoustic dimensions at the same time.

The revcorr approach consists in introducing random fluctuations into a stimulus and measuring how specific patterns of fluctuations affect recognition. In psycholinguistics, revcorr has been used to reveal the acoustic cues underlying phoneme identification[10,11] or the prosodic cues to paralinguistic information, such as the intention and emotional state of the speaker.[12,13] In the latter case, participants are asked to make sense of utterances resynthesized with a random prosody. For example, Ponsot *et al.*[12] used a voice-processing algorithm to systematically randomize the prosody of existing speech recordings,[14] asking their participants to judge the newly generated sounds along a predefined criterion of trustworthiness. The authors then associated the specific random prosody with the corresponding behavioral responses by computing the difference between the mean pitch contour of voices classified as trustworthy and non-trustworthy. This analysis returned a first-order kernel, interpreted as a "mental template," that is able to capture which aspects of the random prosody critically affected the participants' responses.

The objective of the experiment reported here was to evaluate the efficiency of a prosody revcorr protocol in the study of segmentation cues. Using a segmentation task based on the work by Spinelli *et al.*,[3] we measured the perceptual kernels of two pairs of phonemically identical sentences for a group of participants.

2. Materials and methods

The segmentation task was implemented as a listening experiment using the fastACI toolbox.[15,16] In each condition, the test sounds were two phonemically identical sentences whose details are given next. Complementary details for replicating the experiment or reproducing any of the analyses presented in this study are given in the supplementary material.

2.1 Target sentences

Four ambiguous sentences from Spinelli *et al.*[3] were used as targets in the present experiment. The sounds are recordings of the sentences "c'est l'amie/la mie" ("this is the friend/the crumb," condition LAMI) or "c'est l'appel/la pelle" ("this is the call/the shovel," condition LAPEL), all uttered by a female speaker with an average f_0 of 210.8 Hz (LAMI) and 216.5 (LAPEL). In each condition, the two test sentences were aligned temporally and set to have the same total duration and root-mean-square level. The temporal alignment was done manually to roughly match the phonetic segments of each sentence pair. The spectrograms of all sentences are shown in Figs. 1(A) and 1(B).

2.2 Stimuli preparation: Prosody resynthesis

The target sentences were divided into N_{seg} segments of 100 ms, irrespective of their phonetic content, as indicated by the vertical dashed lines in Fig. 1. In each trial, one of the two targets was randomly selected and resynthesized with a new

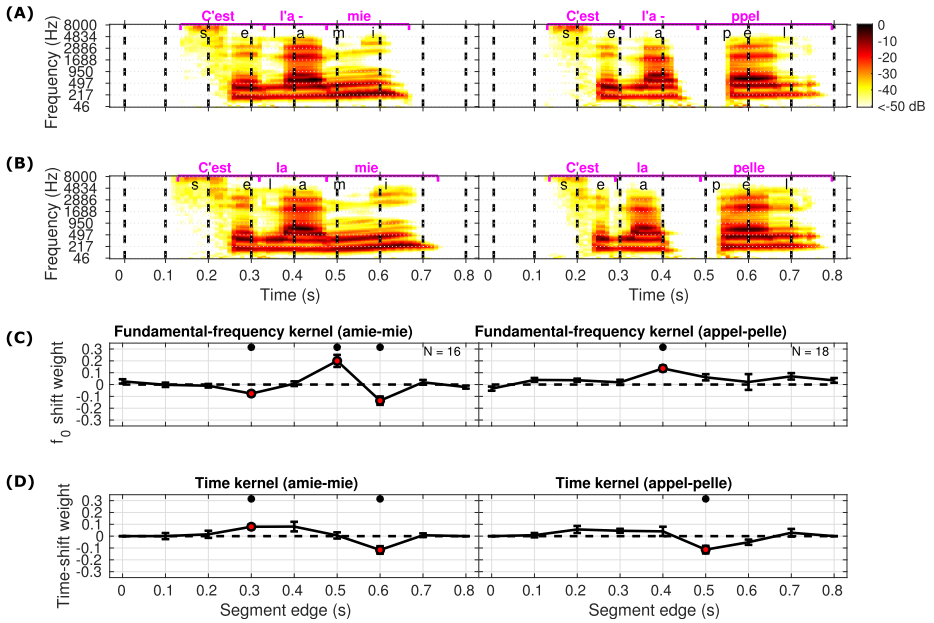

Fig. 1. Targets and mean results for the segmentation task in the LAMI (left panels) and LAPEL conditions (right panels). Top panels [(A) and (B)]: Spectrograms of test sentences. These spectrograms have a constant frequency (vertical) scaling in the equivalent-rectangular bandwidth (ERB) scale. Bottom panels [(C) and (D)]: f_0 and time kernels averaged across all participants. Error bars correspond to ± 1.96 standard error of the mean (SEM), and red markers (and asterisks) indicate the weights that are significantly different from zero (significance threshold $p < 0.005$).

prosody, using the WORLD toolbox.[17] A random f_0 vector with N_{edge} values (with $N_{edge} = N_{seg} + 1$) was drawn from a Gaussian distribution with a mean f_0 of 210.8 Hz (LAMI) or 216.5 Hz (LAPEL) and a standard deviation (SD) of 100 cents (one musical semitone), with the constraint that new f_0 values should not deviate by more than 2.2 times the SD (i.e., 220 cents) from the corresponding mean f_0. For each 100-ms long segment, the original f_0 trajectory of the sentence was replaced by a monotonic f_0 transition between the two random f_0 values of the corresponding segment edges. Simultaneously, the edge positions (excluding the first and last edges) were shifted by a random Gaussian value (zero mean and SD = 15 ms, bounded to ± 2.2 SD), leading to stretched segment durations with an SD of about 20 ms. These shifts were used by WORLD to resynthesize sentences with randomly compressed or stretched segments, which were eventually used in the segmentation experiment. This process was repeated 400 times for each sentence, resulting in speech sounds (800 per condition) with unreliable duration cues and entirely neutralized f_0 cues, while keeping the natural intensity and formants of the original sentences. This choice of resynthesis parameters is in line with previous prosody revcorr studies.[12,13]

2.3 Participants and experimental protocol

Data were collected for two independent groups of 16 and 18 participants for the LAMI and LAPEL conditions, respectively. The participants were all students of the bachelor program in psychology from Grenoble Alpes University. All students had self-reported normal hearing and were rewarded with course credits for their participation in the experiment.

The listening experiment was implemented using a one-interval two-alternative forced choice paradigm. In each trial, the participant was instructed to indicate whether the presented stimulus was "l'aX" ("l'amie" or "l'appel," option 1) or "la X" ("la mie" or "la pelle," option 2). Feedback was provided after each trial. Each experiment consisted of two sessions of 400 trials, with a short training preceding the first experimental session. The stimuli were presented at a comfortable level, with a random roving—drawn from a uniform distribution between ± 2.5 dB—to partly discourage the use of absolute loudness cues during the task.

2.4 Analysis

Revcorr kernels were computed using a generalized linear model (GLM) approach.[10,11,18] The GLM relates the exact random prosody in each trial to the corresponding response of the participant by attributing a decision weight β_k to each element in the f_0 vector and a decision weight γ_k to each element in the time-shift vector. These sets of weights, β_k and γ_k, define the f_0 and time kernels, respectively. The kernels represent the influence of the f_0 and timing information at each segment edge k on the decision of the participant, i.e., how a particular shift in f_0 or timing can systematically bias the participant into responding "l'aX" or "la X." Within the GLM, the information about f_0 ($f_{k,i}$) and timing ($t_{k,i}$) is linearly

combined with f_0 and time weights at each edge, β_k and γ_k. The weighted sum is subsequently transformed using a logistic function Φ, resulting in a probability of response for trial i,

$$P(r_i = \text{“l'aX”}) = \Phi\left(\sum_k \beta_k \cdot f_{k,i} + \sum_k \gamma_k \cdot t_{k,i} + c\right), \tag{1}$$

where c is a term that reflects the participant's overall preference (or bias) for one of the two responses. The kernels defined by β_k and γ_k reveal which temporal segments are effectively used by the participant to respond “l'aX” with a probability $P(r_i = \text{“l'aX”})$ or to respond “la X,” with a probability $P(r_i = \text{“la X”})$, obtained as $1 - P(r_i = \text{“l'aX”})$.

3. Results and discussion

3.1 Psychophysical kernels

The derived f_0 and time kernels averaged across all participants ($N = 16$ for LAMI, $N = 18$ for LAPEL) are shown in Figs. 1(C) and 1(D) (left for LAMI, right for LAPEL), respectively. The kernels provide a visualization of the listening strategy of the participants, i.e., they emphasize the prosody aspects they relied on during the task. For instance, the above-mentioned effect of the f_0 shift on /a/ is captured by the GLM, which attributes a significant positive weight to the f_0 information on the 0.5-s segment edge in LAMI and on the 0.4-s segment edge in LAPEL [Fig. 1(C)].

In the LAMI condition, the f_0 kernel [Fig. 1(C), left] reveals three critical regions, significantly different from zero at the group level with a positive weight at 0.5 s [$t(15) = 3.84, p < 0.005$] surrounded by negative weights at 0.3 s [$t(15) = -3.80, p < 0.005$] and 0.6 s [$t(15) = -3.79, p < 0.005$]. This indicates that when presented with a high f_0 at 0.5 s and/or low f_0 at 0.3 and 0.6 s, the participants were more likely to perceive the stimulus as “c'est l'amie.” On the contrary, the opposite patterns of f_0 elicit more “c'est la mie” responses. In the time kernel [Fig. 1(D), left], the weights associated with the segment edges at 0.3 and 0.6 s were found to be significant [segment at 0.3 s: $t(15) = 3.80, p < 0.005$; segment at 0.6 s: $t(15) = -3.79, p < 0.005$].

In the LAPEL condition, the f_0 kernel [Fig. 1(C), right] was found to only have one critical region with a positive weight at 0.4 s [$t(17) = 5.73, p < 0.0001$]. This indicates that when presented with a high f_0 at 0.4 s, the participants were more likely to perceive the stimulus as “c'est l'appel.” In the time kernel [Fig. 1(D), right], the weight associated with the segment edge at 0.5 s was found to significantly bias the participants' response toward “la pelle” [$t(17) = -3.47, p < 0.005$].

An indication of the kernels' goodness of fit can be obtained from Eq. (1) and the average kernels from Fig. 1, resulting in response predictions significantly above chance for both conditions with respect to the participants' data [55.7% (SD = 4.8%) for LAMI; 54.5% (SD = 2.2%) for LAPEL] (further details can be found in the supplementary material).

3.2 Interpretation of the significant weights

The positive f_0 weight in the two test conditions [Fig. 1(C), edges at 0.5 and 0.4 s for LAMI and LAPEL, respectively] coincides with the segment containing the vowel /a/. As also shown in Fig. 2, this positive weight indicates that a higher f_0 in this segment is more likely related to a “l'aX” response, supporting the notion of early f_0 rise cue for segmentation.[3,9] The significant weight at time 0.6 s of the LAMI f_0 kernel [Fig. 1(C), not visible for LAPEL] can be similarly interpreted in terms of the early f_0 rise. This segment roughly corresponds to the /i/ vowel in the targets. The negative f_0 weight indicates that a downward f_0 shift elicits more “l'amie” responses or, in other words, that an upward f_0 shift elicits more “la mie” responses. Consequently, an /i/ vowel with a high f_0 supports the segmentation at the syllable that contains it, resulting in the consonant-initial word “mie.”

The significant weight at time 0.3 s [Fig. 1(D), /e/ vowel in the LAMI targets, not visible for LAPEL] is harder to interpret. This weight may correspond to a form of reference point for the assessment of the f_0 height on the subsequent /a/ vowel.

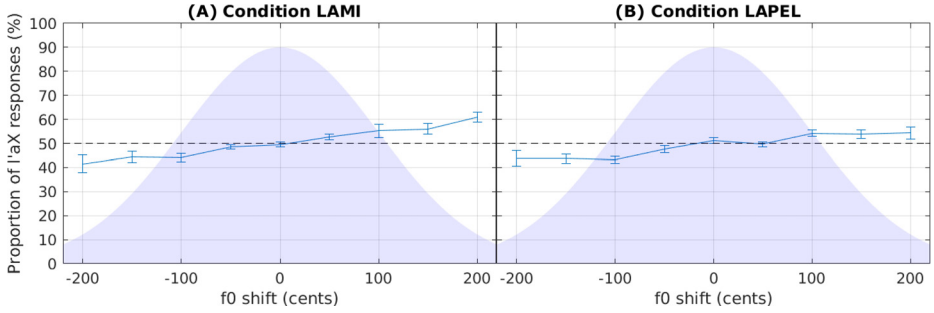

Fig. 2. Proportion of “l'amie” (panel (A)) and “l'appel” (panel (B)) responses, displayed as a function of the random f_0 shift at $t = 0.5$ s (A) and $t = 0.4$ s (B), corresponding to the temporal position the vowel /a/. Error bars correspond to ±1.96 SEM, comparable to 95% confidence intervals under the assumption of normality. The dotted line indicates the absence of bias, where “l'aX” and “la X” were equally chosen by the participants. The shaded area corresponds to the Gaussian distribution from which the f_0 shifts were drawn (no unit).

The explanation could be related to the "early elbow,"[9] where the presence of an inflection in the f_0 contour is equivalent, from an intonational perspective, to an early f_0 rise cue. In our case, a heightened f_0 on /e/ creates an artificial elbow in the f_0 contour, which in turn influences the segmentation process.

For the time kernels [Fig. 1(D)], the significant weights are located at times 0.3 and 0.6 s for LAMI and at time 0.5 s for LAPEL, indicating that participants relied on the relative durations of the first and second syllables. This is in line with the durational differences observed in the production of "l'aX" and "la X" sentences, where the first syllable (/la/) was found to be shorter in "la X" than in "l'aX," while the reverse was true for the second syllable.[3] In our experiment, a shift of the 0.6 s (or 0.5 s) segment edge affects the perceived duration of the /mi/ (or /pel/) syllable and, hence, the segmentation process. Also, in line with Spinelli *et al.*'s observations, the positive weight at time 0.3 s (only significant for LAMI) indicates that longer /la/ segments (and, thus, shorter /e/ segments) are associated with more "l'amie" responses. More generally, the measured time kernels support the conclusions that listeners use segmental durations near word boundaries as a cue for segmentation.[5,6]

3.3 Performance in the task

The prosody resynthesis method (Sec. 2.2) was designed to neutralize the prosody cues of the original target sentences, reducing the performance in the segmentation task from an expected score of 75% (Spinelli *et al.*,[8] ambiguous condition) to an average percentage of correct responses close to chance [57.3% (SD = 6.7%) for LAMI; 49.3% (SD = 4.3%) for LAPEL], despite the presence of intensity and formant cues that remained available in the stimuli. Nevertheless, this performance does not imply that the participants were responding at random. Given the strong f_0 weighting shown in Fig. 1 for /a/ segments at 0.5 s (LAMI) and 0.4 s (LAPEL), we depict the percentage of "l'aX" responses as a function of the random f_0 shift in those segments, irrespective of the fluctuations in the other segments. The monotonically increasing proportions (blue traces in Fig. 2) indicate that shifts closer to −200 or 200 cents systematically biased the participants' answers toward "la X" or "l'aX," respectively. These curves suggest that the segmentation judgments were driven by the random prosody imposed on the speech target, and not by the target itself. The fact that a null f_0 shift led to an overall score of 50% indicates that there was no bias in favor of one response or the other.

The finding that raising the f_0 in the /a/ vowel leads to more vowel-initial segmentation [with $P(r_i =$ "l'aX") increasing with Δf_0 in both panels of Fig. 2] is consistent with the results obtained by Spinelli *et al.*[8] (see their Fig. 4). This observation is usually interpreted as evidence for the role of the early f_0 rise in French, where the presence of a heightened f_0 on the first syllable of a content word is used as a perceptual cue by the listener to segment this word.[8,9]

4. Conclusion and outlook

The revcorr approach allows us to test different prosody dimensions (f_0 and duration cues) at the same time. Although the stimulus preparation was based on a resynthesis algorithm, the set of parameters used in the algorithm did not require *a priori* assumptions about the location of the cues under investigation, contrary to more traditional hypothesis-driven approaches.

Our results show that first-order kernels can be derived for a segmentation experiment using a prosodic revcorr approach. The estimated weights can be interpreted by relating the arbitrary time positions of the segment edges to the phonetic content of the target speech sounds. To show that our observations did not critically depend on the specific set of target sentences or the adopted resynthesis parameters, we tested three additional control conditions in a reduced group of participants. All details of these extra conditions are presented in the supplementary material. These tests showed that (1) our conclusions are not specific to the selected target sentences or the arbitrary segment edge positions (supplemental Fig. 1), and (2) the target-specific kernels within each tested sentence pair are very similar (supplemental Fig. 2), indicating that potential subtle acoustic differences between contrasting sentences did not strongly influence our main results.

It is important to note that, although the kernels are presented as contours, the prosody revcorr method assesses the effect of f_0 and timing at each segment edge independent of the other segment edges. However, the auditory system does not process information from different time points independently. For this reason, we should refrain from interpreting kernels as "prosodic prototypes" in general. Nevertheless, as a means of confirmation, we resynthesized the targets by employing the f_0 and time kernels as prosodic contours. As expected, this resynthesis indeed led to highly confusing stimuli.

In general, the prosodic revcorr method presented in this study showed high replicability across sentences (Fig. 1; supplemental Fig. 1), across targets (supplemental Fig. 2), and across segment edge positions (Fig. 1, left; supplemental Fig. 1, right). Despite the modest sample size in this study, our results suggest that the estimated kernels truly reflect a general "segmentation strategy" of the listeners in the task. Consequently, the prosody revcorr approach seems to be well suited for the investigation of psycholinguistic processes. Compared to previous methods that rely on the manipulation of a single cue, the prosody revcorr method requires minimal assumptions about the potential cues that participants could use. Probing the large space of all possible cues available to the listeners is less time-consuming and can also allow the investigation of potential interactions between cues.

SUPPLEMENTARY MATERIAL

See supplementary material at https://doi.org/10.1121/10.0021022 for the evaluation of the prosody revcorr paradigm in control conditions using additional pairs of sentences.

Acknowledgments

This study was supported by the French National Research agency through the grants fastACI (Grant No. ANR-20-CE28-0004; A.O., L.V.), FrontCog (Grant No. ANR-17-EURE-0017; A.O., L.V.), CeLyA (Grant No. ANR-10-LABX-0060; E.G.), and IDEXLYON (Grant No. 16-IDEX-0005; E.G.).

AUTHOR DECLARATIONS

Conflict of interest

The authors declare they have no conflicts of interest.

DATA AVAILABILITY

The raw data of this study are available on Zenodo (Ref. 16) and the fastACI toolbox, required to replicate all figures and analyses, can be retrieved from Ref. 15.

References

[1]J. McClelland and J. Elman, "The TRACE model of speech perception," Cogn. Psychol. **18**, 1–86 (1986).
[2]J. McQueen, "Segmentation of continuous speech using phonotactics," J. Mem. Lang. **39**, 21–46 (1998).
[3]E. Spinelli, P. Welby, and A.-L. Schaegis, "Fine-grained access to targets and competitors in phonemically identical spoken sequences: The case of French elision," Lang. Cogn. Proc. **22**, 828–859 (2007).
[4]L. Nakatani and K. Dukes, "Locus of segmental cues for word juncture," J. Acoust. Soc. Am. **62**, 714–719 (1977).
[5]M.-H. Banel and N. Bacri, "On metrical patterns and lexical parsing in French," Speech Commun. **15**, 115–126 (1994).
[6]E. Shoemaker, "Durational cues to word recognition in spoken French," Appl. Psycholinguist. **35**, 243–273 (2014).
[7]S.-A. Jun and C. Fougeron, "A phonological model of French intonation," in *Intonation*, edited by A. Botinis (Springer, Amsterdam, 2000), pp. 209–242.
[8]E. Spinelli, N. Grimault, F. Meunier, and P. Welby, "An intonational cue to word segmentation in phonemically identical sequences," Atten. Percept. Psychophys. **72**, 775–787 (2010).
[9]P. Welby, "The role of early fundamental frequency rises and elbows in French word segmentation," Speech Commun. **49**, 28–48 (2007).
[10]L. Varnet, K. Knoblauch, F. Meunier, and M. Hoen, "Using auditory classification images for the identification of fine acoustic cues used in speech perception," Front. Hum. Neurosci. **7**, 865 (2013).
[11]A. Osses and L. Varnet, "A microscopic investigation of the effect of random envelope fluctuations on phoneme-in-noise perception," bioRxiv 2022.12.27 (2022).
[12]E. Ponsot, J. Burred, P. Belin, and J.-J. Aucouturier, "Cracking the social code of speech prosody using reverse correlation," Proc. Natl. Acad. Sci. U.S.A. **115**, 3972–3977 (2018).
[13]L. Goupil, E. Ponsot, D. Richardson, G. Reyes, and J.-J. Aucouturier, "Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature," Nat. Commun. **12**, 861 (2021).
[14]J. Burred, E. Ponsot, L. Goupil, M. Liuni, and J.-J. Aucouturier, "CLEESE: An open-source audio-transformation toolbox for data-driven experiments in speech and music cognition," PLoS One **14**, e0205943 (2019).
[15]A. Osses and L. Varnet, "fastACI toolbox: The MATLAB toolbox for investigating auditory perception using reverse correlation (v1.3)," https://zenodo.org/record/7888588 (Last viewed 9 September 2023).
[16]A. Osses, E. Spinelli, F. Meunier, E. Gaudrain, and L. Varnet, "Raw and post-processed data for the study of prosodic cues to word boundaries in a segmentation task using reverse correlation," https://zenodo.org/record/7865424 (Last viewed 9 September 2023).
[17]M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," IEICE Trans. Inf. Syst. **E99.D**, 1877–1884 (2016).
[18]Note that, in the present experiment where the random predictors are drawn from independent distributions and have an overall small effect on perception, the GLM approach gives similar results as the more usual difference-of-means procedure, and the interpretation of the resulting kernels remains the same.