



HAL
open science

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

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

► To cite this version:

Alejandro Osses, Elsa Spinelli, Fanny Meunier, Etienne Gaudrain, Léo Varnet. Prosodic cues to word boundaries in a segmentation task assessed using reverse correlation. 2023. hal-04121858v1

HAL Id: hal-04121858






<https://hal.science/hal-04121858v1>

Preprint submitted on 8 Jun 2023 (v1), last revised 29 Sep 2023 (v2)

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.

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

Alejandro Osses ¹, Elsa Spinelli ², Fanny Meunier ³, Etienne Gaudrain ⁴, and Léo Varnet ^{1, a)}

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, France*

⁴⁾ *Lyon Neuroscience Research Center, CNRS, INSERM, Université Lyon 1, Lyon, France*

(Dated: 8 June 2023)

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 experimental assumptions. The method was evaluated using pairs of phonemically-identical sentences in French, whose prosody was changed in each trial 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.

I. 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 TRACE¹ consider segmentation as a mere by-product of lexical access, where the selected set of boundaries is the one which 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 Bacri⁵ is an example of such an approach. Separately recorded monosyllables were assembled together into ambiguous disyllabic stimuli like /ba.gaʒ/. 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 to dissociate the perceptual role of segmental duration from the role of other covarying cues. More recently, Shoemaker⁶ tackled this limitation by using stimuli where only the duration dimension was manipulated, while ensuring that all other stimulus 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

^{a)} Author to whom correspondence should be addressed. Electronic mail: leo.varnet@ens.psl.eu.

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.mõ.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 segment⁸ 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 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 reverse-correlation (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 criteria 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 demonstrate 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.

II. 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 materials.

A. 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 fundamental frequency f_0 of 210.8 Hz (LAMI) and 216.5 Hz (LAPEL). In each condition, the two test sentences were aligned temporally and equalized in energy and total duration. The spectrograms of all sentences are shown in Fig. 2A–B.

B. 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. 2. In each trial, one of the two targets was randomly selected and resynthesized with a new prosody, using the WORLD toolbox.¹⁷ A random f_0 vector with $N_{\text{seg}} + 1$ values 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 were shifted by a random Gaussian value (zero mean and SD= 15 ms, bounded to ± 2.2 SD). 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}

C. 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 the 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 1-interval 2-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.

D. Analysis

Reverse correlation 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 ($t_{k,i}$) are 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 :

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

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 $1 - P(r_i = \text{“l’aX”})$.

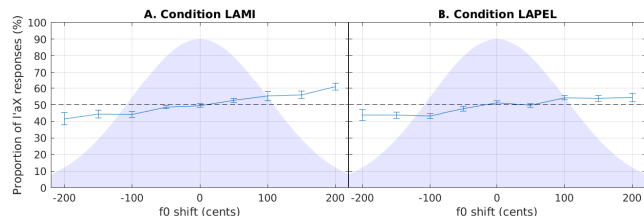


FIG. 1. Proportion of “l’amie” (left panel) and “l’appel” (right panel) responses, displayed as a function of the random f_0 shift at $t = 0.5$ s (left) and $t = 0.4$ s (right), corresponding to the temporal position the vowel /a/. Error bars correspond to ± 1.96 the standard error of the mean (SEM), 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).

III. RESULTS AND DISCUSSION

A. Performance in the task

The prosody resynthesis method (Sec. II B) 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. Fig. 1 shows the percentage of responses “l’aX” as a function of the random f_0 shift imposed on the /a/ segment, irrespective of the fluctuations in the other segments. The monotonically increasing proportions (blue traces in Fig. 1) indicate that shifts closer to -200 or 200 cents systematically biased the participants’ answers towards “la X” or “l’aX,” respectively. These curves suggest that the segmentation judgements 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 $P(r_i = \text{“l’aX”})$ increasing with Δf_0 in both panels of Fig. 1] is consistent with the results obtained by Spinelli *et al.*⁸ (see their Fig. 4). This observation is usually interpreted as an 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}

B. Psychophysical kernels

The derived f_0 and time kernels averaged across all participants ($N = 16$ for LAMI, $N = 18$ for LAPEL) are

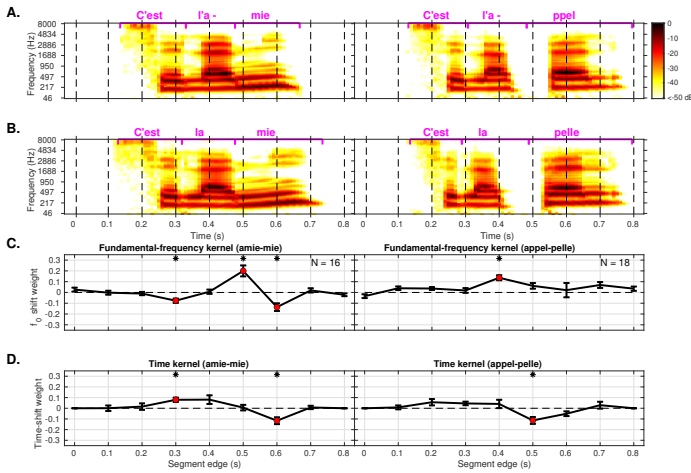


FIG. 2. Targets and mean results for the segmentation task in the LAMI (left panels) and LAPEL conditions (right panels). Top panels (A–B): Spectrograms of test sentences. These spectrograms have a constant frequency (vertical) scaling in the equivalent-rectangular bandwidth (ERB) scale. Bottom panels (C–D): f_0 and time kernels averaged across all participants. Error bars correspond to ± 1.96 SEM, and red markers (and asterisks) indicate the weights that are significantly different from zero (significance threshold $p < 0.005$).

shown in Fig. 2C and 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. 2C).

In the LAMI condition, the f_0 kernel (Fig. 2C, 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 elicits more “c’est la mie” responses. In the time kernel (Fig. 2D, left), the weights associated to 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. 2C, 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. 2D, right), the weight associated to the segment edge at 0.5 s was found to significantly bias the participants’ response towards “la pelle” [$t(17) = -3.47, p < 0.005$].

C. Interpretation of the significant weights

The positive f_0 weight in the two test conditions (Fig. 2C, 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. 1, 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. 2C, 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. 2D, /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. 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. 2D), 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}

IV. CONCLUSION AND OUTLOOK

The revcorr approach allows 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. In order 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 materials. These tests showed that (1) our conclusions are not specific to the selected target sentences or the arbitrary segment edge positions (suppl. Fig. 1), and that (2) the target-specific kernels within each tested sentence pair are very similar (suppl. 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. 2; suppl. Fig. 1), across targets (suppl. Fig. 2), and across segment edge positions (Fig. 2, left; suppl. Fig. 1, right). All in all, our results indicate that the estimated kernels truly reflect a general “segmentation strategy” of the listeners in the task, and that the prosody revcorr approach is suited to investigate psycholinguistic processes. Compared to previous methods which 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 both less time-consuming and can also allow the investigation of potential interactions between cues.

ACKNOWLEDGMENTS

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

¹J. McClelland and J. Elman, “The TRACE model of speech perception,” *Cognitive Psychology* **18**, 1–86 (1986).

²J. McQueen, “Segmentation of continuous speech using phonotactics,” *Journal of Memory and Language* **39**, 21–46 (1998).

³E. Spinelli, P. Welby, and A.-L. Schaegis, “Fine-grained access to targets and competitors in phonemically identical spoken sequences: The case of French elision,” *Language and Cognitive Processes* **22**, 828–859 (2007).

⁴L. Nakatani and K. Dukes, “Locus of segmental cues for word juncture,” *J. Acoust. Soc. Am.* **62**, 714–719 (1977).

⁵M.-H. Banel and N. Bacri, “On metrical patterns and lexical parsing in French,” *Speech Commun.* **15**, 115–126 (1994).

⁶E. Shoemaker, “Durational cues to word recognition in spoken French,” *Applied Psycholinguistics* **35**, 243–273 (2014).

⁷S.-A. Jun and C. Fougeron, “A phonological model of French intonation,” in *Intonation*, edited by A. Botinis (Springer Netherlands, 2000) pp. 209–242.

⁸E. Spinelli, N. Grimault, F. Meunier, and P. Welby, “An intonational cue to word segmentation in phonemically identical sequences,” *Attention, Perception & Psychophysics* **72**, 775–787 (2010).

⁹P. Welby, “The role of early fundamental frequency rises and elbows in French word segmentation,” *Speech Commun.* **49**, 28–48 (2007).

¹⁰L. Varnet, K. Knoblauch, F. Meunier, and M. Hoen, “Using auditory classification images for the identification of fine acoustic cues used in speech perception.” *Frontiers in Human Neuroscience* **7**, 865 (2013).

¹¹A. Osses and L. Varnet, “A microscopic investigation of the effect of random envelope fluctuations on phoneme-in-noise perception,” (2022).

¹²E. Ponsot, J. Burred, P. Belin, and J.-J. Aucouturier, “Cracking the social code of speech prosody using reverse correlation,” *Proceedings of the National Academy of Sciences* **115**, 3972–3977 (2018).

¹³L. 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,” *Nature Communications* **12**, 861 (2021).

¹⁴J. 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).

¹⁵A. Osses and L. Varnet, “fastACI toolbox: the MATLAB toolbox for investigating auditory perception using reverse correlation (v1.2),” (2022).

¹⁶A. 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,” (2023).

¹⁷M. 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).

¹⁸Note 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 than the more usual difference-of-means procedure, and the interpretation of the resulting kernels remains the same.

APPENDIX A: SUPPLEMENTARY MATERIALS

The information in this section is supplementary to our main study. For this reason, the explanations in these supplementary materials may not be self-explanatory. In such cases, the reader is referred to the main text or the raw- and post-processed data of our study. The raw- and post-processed data can be retrieved from Zenodo.¹⁶ All analyses and figures can be recreated using the fastACI toolbox,¹⁵ from now on referred to as “the toolbox.”

1. Segmentation task using additional sentence pairs

Two extra pairs of sentences were tested using the same experimental protocol as presented in the main text of this study. The new pairs were “c’est l’accroche”/“c’est la croche” (“this is the catch phrase/the eighth note,” condition LACROCH), “c’est l’alarme”/“c’est la larme” (“this is the alarm/the tear,” condition LALARM). The sentences had a mean f_0 of 207.6 Hz (LACROCH) or 207.9 Hz (LALARM).

A total of $N = 5$ participants were enrolled in each new condition and, as in the main experiments, they were bachelor students from the Grenoble Alpes University. The resulting f_0 and time kernels are shown in suppl. Fig. 1C–D, for the LACROCH (left) and LALARM (middle) conditions, respectively. Due to the low number of participants in these conditions, significance was not tested. However, both f_0 kernels have a prominent positive cue at the time segment that marks the end of the /a/ vowel, consistent with the f_0 kernels from Fig. 2. In terms of the obtained time kernels, in both conditions we found a positive weight at around 0.4 s, followed by negative weights in the subsequent segment, an effect that was stronger in the LALARM condition. This effect was also found in the main conditions LAMIE and LAPEL. In the LACROCH condition (suppl. Fig. 1D, left), an additional negative weight at 0.2 s was found. That segment edge is located in the middle of the syllable /se/. This cue was not found in other conditions, but when inspecting the spectrograms, it is possible to note that in the sentence “c’est la croche” (suppl. Fig. 1B, left), the syllable /se/ is somewhat longer and starts earlier than in the sentence “c’est l’accroche.” Hence, this cue might reflect that participants benefit for this slight misalignment in the contrasting sentences.

2. Segmentation task in the LAMI condition for different segment edges

The condition LAMI was re-evaluated using segment edges that were arbitrarily shifted by 20 ms. The purpose of this analysis was to show that the derived kernels are not critically affected by the choice of segment edges. As indicated by the vertical dashed lines in suppl. Fig. 1A–B (right), the original LAMI sentences had different knee points for the f_0 randomization and time compression/-expansion.

A total of $N = 3$ participants were recruited for this new experimental condition. The resulting “LAMI_shifted” kernels (black traces) are reasonably well overlapped with the kernels obtained for the original task. For ease of comparison, the LAMI kernels from Fig. 2 were replotted and are indicated by the gray dashed line and circle markers. The only LAMI cue that vanished was the negative cue in the f_0 kernel (Fig. 1C, right) at 0.6 s, which is not visible at time 0.62 s in the new black traces. One possible reason is that such cue may be related to the /m/ phoneme, whose time-frequency content was almost entirely contained in the LAMI segment between 0.5 and 0.6 s. In contrast, in the new kernel, the /m/ onset is contained by the preceding segment (between 0.42 and 0.52 s), as indicated by the pink markers in the spectrograms. However, no strong conclusions can be drawn from the new (low powered) analysis due to the low number of participants. In terms of the time kernel, the magnitude of the weights are consistently lower than in the case of the f_0 kernels and are also overlapped across conditions.

3. Control analysis: Target-specific kernels

This last control analysis investigates the potential effect of subtle timing differences between tested sentence pairs, that could have affected the main results presented in Fig. 2, as suggested by the time cue at 0.2 s in LACROCH (suppl. Sec. A1). A straightforward way to investigate this question is to compute separate kernels for each test sentence.

For this analysis, we re-processed the data for all participants for the LAMI, LAPEL, LACROCH, and LALARM conditions (excluding LAMI_shifted). The only difference with respect to the analysis described in Sec. IID (and Equation 1) is that only trials leading to either “l’aX” or “la X” were used. The obtained kernels are shown as separate blue and red traces in suppl. Fig. 2, respectively. The obtained results support the idea that our findings do not strongly depend on one of the two targets. Still, there are some occasional differences in the weights, for example in LAPEL at $t = 0.4$ s. This is possibly due to the fact that in the two /lapel/ targets, the 0.4-s segment edge falls just before or just after the first syllable offset, making it less efficient to manipulate the duration of the /a/ segment).

4. Reproducing the experiment using the fastACI toolbox

The segmentation revcorr experiment can be replicated using the toolbox.¹⁵ The presented segmentation task is included as of version 1.3 of the toolbox under the name ‘segmentation.’ The segmentation experiment is thus coded into four MATLAB scripts: `segmentation_init.m`, `segmentation_cfg.m`, `segmentation_set.m`, and `segmentation_user.m`. The experiment can be run for an arbitrary participant ‘STest’ in the condition LAMI, using the following command (Listing 1):

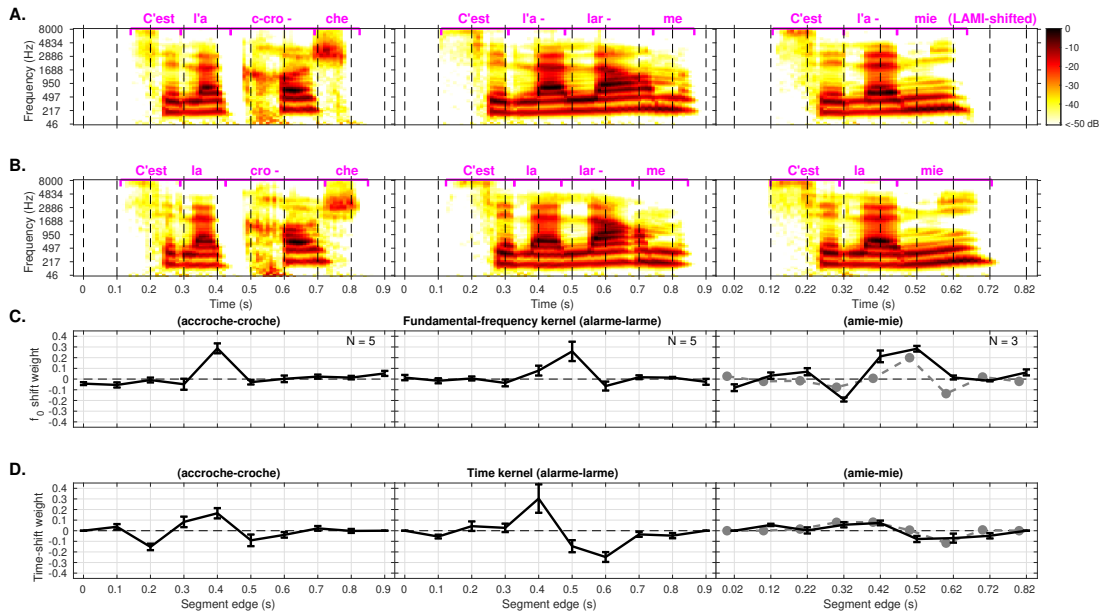


FIG. 1. Spectrograms of the target sounds (top panels, rows A–B) and assessed kernels averaged across participants (bottom panels, rows C–D). Same caption as in Fig. 2. For comparability purposes, in LAMI_shifted (right column), the gray curves indicate the comparable kernels for the LAMI sentences from Fig. 2. Error bars represent ± 1.96 SEM. The gray and black curves are well overlapped (more details are given in the text).

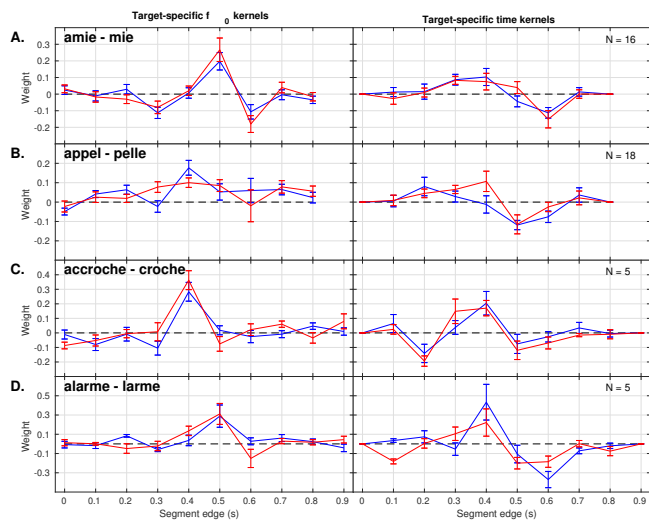


FIG. 2. Mean f_0 (left panels) and time kernels (right panels) in each condition, computed for each independent sentence (blue: “laX”, red: “la X”). Error bars correspond to ± 1.96 SEM.

Listing 1. MATLAB commands required to run the segmentation task using the condition LAMI

```

1 expname='segmentation'; % experiment name
2 Subj_ID='STest'; % Participant ID
3 Condition='LAMI'; % Main condition 1
4 fastACI_experiment(expname,Subj_ID,Condition);

```

To run the experiment using the main condition LAPEL, line 3 from Listing 1 needs to be changed to:

```
3 Condition='LAPEL'; % Main condition 2
```

To run the experiment using the extra condition LACROCH, line 3 from Listing 1 needs to be changed to:

```
3 Condition='LACROCH'; % Extra condition
```

To run the experiment using the extra condition LALARM, line 3 from Listing 1 needs to be changed to:

```
3 Condition='LALARM'; % Extra condition
```

To run the experiment using the extra condition LAMI_shifted, line 3 from Listing 1 needs to be changed to:

```
3 Condition='LAMI_shifted'; % Extra condition
```

However, for a exact reproduction of our analysis, the user is required to store locally all the experimental data of the current study.¹⁶ Then each of the study figures can be obtained by (independently) running the following lines:

```

5 publ_osses2023a_JASA_EL_figs('fig1');
6 publ_osses2023a_JASA_EL_figs('fig2');
7 publ_osses2023a_JASA_EL_figs('fig1_suppl');
8 publ_osses2023a_JASA_EL_figs('fig2_suppl');

```