INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

F0 slope and mean: cues to speech segmentation in French

Maria del Mar Cordero[1], *Fanny Meunier*[1], *Nicolas Grimault*[2,3], *Stéphane Pota*[4], *Elsa Spinelli*[4]

[1]Université Côte d'Azur, CNRS, BCL, France
[2]University Lyon 1, Lyon, France
[3]CNRS, UMR 5292, INSERM, U1028, Lyon Neuroscience Research Center, Auditory Cognition and Psychoacoustics Team, Lyon, France
[4]Université Grenoble Alpes, CNRS, LPNC, Grenoble, France

maria-del-mar.cordero-rull@etu.univ-cotedazur.fr, fanny.meunier@unice.fr, nicolas.grimault@cnrs.fr, stephane.pota@gmail.com, elsa.spinelli@univ-grenoble-alpes.fr

Abstract

This paper evaluates the use of intonational cues during word segmentation in French. Specifically, we aim to examine how the characteristics of the fundamental frequency (F0) that can be observed at the beginning of words influence their processing. Native speakers of French were presented with phonemically identical sequences, such as /selami/ (*c'est l'amie*/*la mie* "it's the friend/the crumb"). To test which properties of the F0 affect the perceived segmentation, we manipulated the F0 slope and/or the mean value of the first vowel /a/ in consonant-initial items (e.g., *la mie*). To assess differences in off-line vs online processing, we used a two-alternative, forced-choice task in Experiment 1 and a lexical decision task in Experiment 2. A previous study showed that vowel-initial segmentation was enhanced when the F0 mean value increased. However, the present study shows that modifying the F0 slope while keeping the F0 mean value constant also influences speech segmentation in both off-line and online tasks. This suggests that listeners use the F0 slope as a cue at the beginning of content words.

Index Terms: spoken word recognition, lexical segmentation, French, F0, intonational cues

1. Introduction

Natural speech is rich and complex, and one of its striking specificities is that there are no clear marks or pauses for word boundaries, which would be expected to be very challenging for listeners. However, the reality is that listeners recognize words correctly and succeed in understanding continuous speech on a daily basis. Depending on their language, listeners rely on different information to segment a speech stream.

It has been shown that a certain set of allophonic cues correlates with word boundaries. For instance, in languages such as English, French, Dutch and Korean, word-initial consonants (e.g.,/s/ in *I scream*) are longer than word-final consonants (e.g.,/s/ in *ice cream*) [1,2,3,4]. Off-line studies, i.e., studies using tasks performed after the processing of interest, have shown that these durational cues can serve to disambiguate two-word utterances [5]. During online segmentation, i.e., that which is evaluated by tasks that capture the processing as it happens or at least is evaluated early enough not to be contaminated by metacognition, it has been shown that fine-grained acoustic information is also used to modulate the activation of possible candidates [3].

The metric structure of a language has also been identified as another cue that listeners can use when segmenting speech. In English, for example, content words often start with strong syllables. Many experiments have shown the impact of this intonation pattern on the segmentation process. In the early research on this topic, using a word-spotting task in which listeners detected words embedded in nonsense bisyllables (CVCC), the authors [6] showed that detection is slower when the bisyllable has two strong syllables than when it has one strong and one weak syllable. Therefore, *mint* is easier to identify in strong-weak sequences such as *MINTes* than in strong-strong sequences such as *MINTESH*. During online segmentation, the intonation pattern also seems to be used to modulate the activation of competing lexical candidates ([7]; [8] for results in Dutch).

This paper focuses on the role of F0, an intonational cue. Variations in fundamental frequency (F0), duration and intensity have been demonstrated to affect speech perception [9]. For example, higher sounds, as well as sounds with greater intensity, are often experienced as being longer [10]. Nevertheless, F0 alone cannot be attributed to how the duration of sounds is perceived [11].

Of particular interest, F0 has been shown to be involved in the segmentation of continuous speech in French. In French, there is often an F0 rise at the beginning of content words, but an F0 rise also occurs in the final syllables, which are often longer and higher in pitch. Experimental data suggest that an F0 rise in the final syllables contributes to speech segmentation. In a previous study, using an artificial language paradigm, participants had to recognize the words they had previously heard in a learning phase. In this study, Tremblay et al. [12] observed that participants' performance improved when F0 cues were present in the signal compared to their performance when these cues were not present. This suggests that these final prosodic cues are useful during segmentation.

As mentioned earlier, an F0 rise is present at the beginning of content words in French. This can be seen through the comparison of homophonic utterances, i.e., phonemically identical utterances. The segmentation process of these utterances is different; however, correct segmentation is crucial for successful understanding. For example, /lami/ matches *l'amie*, 'the friend', but also *la mie*, 'the crumb'. These two ambiguous utterances differ in their F0 rise, which is present at the beginning of [la] in *l'amie* but later in *la mie*, at [mi]. Therefore, in this case, the F0 is lower for /a/ in *la mie* than for /a/ in *l'amie*, where it is the first phoneme of the word. Thus, this cue is crucial

http://dx.doi.org/10.21437/Interspeech.2020-2509

for proper segmentation and successful understanding. Again, the F0 rise has been found to be used during online segmentation of French. It has been shown that listeners can discriminate between *la mie* and *l'amie* in an ABX task, as well as identify which utterance they heard in a 2AFC task [13]. Increasing the value of F0 in /a/ in *la mie* increased both the responses to vowel-initial choices (e.g., *amie*) and the activation of the lexical representation of vowel-initial targets [14].

The acoustic cues used to discriminate between homophonic sequences in French, such as *l'amie – la mie*, have proven to be robust even with the introduction of intraindividual variations in the sequences tested. A recent study examined the electrophysiological correlates to these acoustic cues and found that homophonic sequences elicited a mismatch negativity (MMN) using a modified oddball paradigm [15]. The oddball paradigm consists of a series of frequent sounds (standards) in which a rare stimulus (deviant) sometimes appears. If the deviant is processed as such by the brain, its appearance prompts an MMN, which appears between 150 and 250 ms after the onset of the stimulus (deviant) and is located in the frontal to central brain regions. The standards in the authors' study were different [la]s extracted from different productions of *l'amie*, while the deviant [la] was extracted from *la mie* (and reversed). This implies that the listener's brain gathers cues that are sufficiently and regularly associated with each of the intended segmentations.

All the studies we have carried out on this topic have considered and characterized the F0 based on its mean value. It could be argued that this is a rather static measure for a characteristic that varies so much and in such a continuous manner. In the experiments we present in this paper, we explore the role of another aspect of the F0: its temporal dynamics that are revealed by the value of its slope.

Recently, the importance of taking into account the value of the F0 slope was demonstrated by a study in Korean [16]. In the Korean language, the canonical intonation pattern is L(HL)H tones (L=Low, H=High), with the word-final H tone indicating a word limit, particularly if followed by an L tone. In the study, the authors used an artificial language in which they manipulated both the timing of the peak in final H tones and the value of initial L tones. The results showed that when segmenting continuous speech, Korean listeners mainly used the difference in pitch value in word boundaries, that is, between final H and initial L tones. When the value of the initial L tone was manipulated and raised, segmentation performances dropped. Interestingly, in this situation, it seems that listeners used the slope between the final H tone and the initial L tone to segment the speech stream.

As an extension of our previous work, our purpose here was to examine the role of different F0 characteristics during segmentation processing. We did this by resynthesizing the F0 of the stimuli used in our work [13,14]. We asked participants to perform two-alternative forced-choice tasks in ambiguous sequences, such as *C'est la mie* 'It's the crumb' vs *C'est l'amie* 'It's the friend', both /selami/. We used five experimental conditions for each sequence: (1- Co-nat) natural production of the utterance with the content word starting with the consonant (e.g., *la mie*, consonant-initial, no resynthesis), and (2- Vow-nat) natural production of the utterance with the content word starting with the vowel (e.g., *l'amie*, vowel-initial, no resynthesis). To create the resynthesized stimuli, we took the utterances of the content word starting with a consonant, such as *C'est la mie*, and resynthesized the vowel (/a/) to obtain the F0 characteristics of /a/ of the vowel-initial content word (e.g., *l'amie*). For the resynthesized conditions, (3- Co-slope) we matched the slope (e.g., *la mie*, with the F0 slope of the vowel /a/ resynthesized to equal that of the first vowel /a/ of its homophone, *l'amie*); (4- Co-shift) we matched the mean value by applying a shift (e.g., *la mie*, with the F0 of /a/ resynthesized to equal that of the vowel-initial homophone); or (5 Co-slope+shift) we matched both the slope and the mean value (e.g., *la mie*, with the F0 slope and the F0 mean value of /a/ resynthesized to be equal to that of the vowel-initial homophone).

Given the features of French intonation and previous results, we hypothesized that manipulating the characteristics of F0 would modify speech segmentation. We expected that increasing the value of the F0 slope of /a/ in *la mie* would lead to a greater segmentation of vowel-initial items. Our experiment allowed us to test whether this effect is due to an increase in the slope, an increase in the mean frequency, or an increase in both. Comparing the different experimental conditions allowed us to clarify which characteristics between the slope and the mean value are more important for the segmentation process and whether both are useful.

2. Experiment 1

2.1. Methods

2.1.1. Participants

Fifty native speakers of French (29 females) with an age range of 18 to 26 years (M=21 years, SD=2) took part in this study. No hearing impairments were reported.

2.1.2. Materials

The stimuli used were 29 pairs of phonemically identical sequences consisting of a noun preceded by a definite article (e.g., *la mie/l'amie,* 'the crumb/the friend') framed in a neutral context (*c'est,* 'it is'). The sequences were grouped into minimal pairs that differed in the beginning of the noun of each member of the pair (e.g., *amie* vs *mie*). The vowel /a/ of the consonant-initial items (e.g., *la mie*) had a consistently lower F0 (M=183 Hz) and F0 slope (-0.06 Hz/ms) than that of vowel-initial items (e.g., *l'amie*, M=199 Hz; F0 slope= +0.12 Hz/ms).

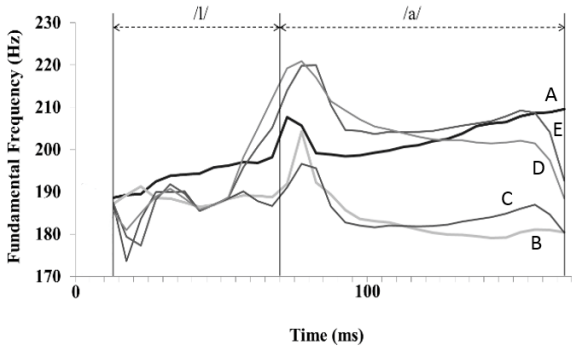

Figure 1: *Illustration of the F0 resynthesis of the first two phonemes (/la/) for the pair "la mie/l'amie". All conditions are represented: Vow-nat (A), Co-nat (B), Co-slope (C), Co-shift (D), and Co-slope+shift (E).*

Three additional tokens were created for each pair by resynthesizing the F0 of the consonant-initial item. Figure 1

shows the mean value and slope for both natural and resynthesized conditions. For the resynthesis process, the STRAIGHT software was used [17]. We computed the average F0 value (Hz) and the F0 slope (Hz/ms) of /a/ from vowel-initial words to modify the F0 value of the first vowel in each of the 3 new tokens. These parameters were applied to the F0 contour of Co-nat in a time window corresponding to the /a/ with 20-ms cosine ramps. Then, we used the F0 curve for the vowel /a/ in Co-nat to generate the three additional conditions: it was multiplied by a scaling factor to reach the same F0 mean value as that in the /a/ in Vow-nat (Co-shift); it was rotated to reach the slope value of the /a/ in Vow-nat while keeping its F0 value (Co-slope); and it was both shifted and rotated (Co-slope+shift).

2.1.3. Procedure

Stimuli were presented aurally over Sennheiser HD 212 Pro headphones at a comfortable listening level. Each participant listened to only one member of each ambiguous pair from each condition and was asked to identify the noun through a forced choice between the two possible nouns (e.g., *mie* vs *amie*).

2.2. Results

The percentage of correct answers in natural productions was 74.4% (79.9% for vowel-initial items and 68.8% for consonant-initial), similar to what was found in previous studies [13,14].

To analyze our data, we used one-way repeated measure ANOVAs with the five condition levels (Co-nat, Co-slope, Co-shift, Co-slope+shift, and Vow-nat). F-values are reported for the analyses, with the participants (*F1*) and items (*F2*) as random factors. The percentages of vowel choices are shown in Figure 2.

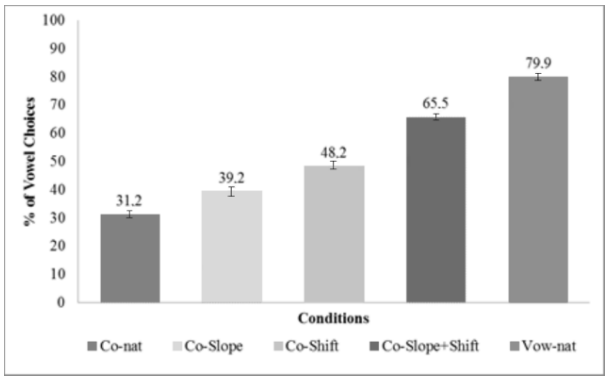

Figure 2: *Mean percentages of vowel responses in the five conditions in Experiment 1. Error bars show standard errors of the means.*

There was a main effect of condition [*F1*(4,196)=67.2, *p*=.0004; *F2*(4,112)=40.6, *p*=.0012]. Importantly, there was a significant difference between the Co-nat and Co-slope conditions [*F1*(1,49)=4.3, *p*=.0434; *F2*(1,28)=4.2, *p*=.0499], and more vowel choices were made in the resynthesized slope condition (39.2%) than in the natural consonant condition (31.2%). Moreover, the Co-slope condition (39.2%) differed significantly from the Co-shift condition (48.2%) [*F1*(1,49)=5.5, *p*=.0231; *F2*(1,28)=8.6, *p*=.0066]. There was also a difference between the Co-shift (48.2%) and Co-slope+shift (65.5%) conditions, with the latter condition receiving more vowel responses. This difference was significant for both participants and items [*F1*(1,49)=16.8, *p*=.0002, *F2*(1,28)=16.5, *p*=.0004]. A significant difference was also found between the Vow-nat and Co-slope+shift conditions [*F1*(1,49)=16.2, *p*=.0002; *F2*(1,28)=12.2, *p*=.0016].

2.3. Discussion

Our results provide further evidence that listeners can discriminate between phonemically identical sequences and can identify much better than chance which word they heard. This finding is in line with studies indicating that listeners use the acoustic marking of word boundaries to correctly segment speech [5,18,3].

Our data show that all three resynthesized conditions lead to an increased segmentation of vowel-initial content words (e.g., *l'amie*). Crucially, when the F0 slope of /a/ in *la mie* was replaced by the F0 slope from *l'amie* but the F0 mean value was kept constant, the responses to initial-vowel items increased from 31.2% to 39.2%. This finding suggests that the subtle slope change (+0.18 Hz/ms) is used during segmentation processes. In line with our previous results [14], when the F0 mean value was increased, there was a greater increase in vowel-initial responses (up to 48.2%).

The condition encompassing two simultaneously converging cues (Co-slope+shift) had higher levels of vowel segmentation (65.5%) than conditions with only one cue. However, the degree of segmentation of vowel-initial natural productions was even higher (79.9%). This finding supports the idea that there are other cues involved in the process of segmentation and disambiguation for the correct understanding of speech, including durational and formant cues.

Our results indicate the relevant role of F0 in the segmentation of ambiguous sequences. To confirm the role of the F0 slope in online segmentation, we conducted a second experiment using a cross-modal identity-priming paradigm with a lexical decision task. Participants listened to short homophonic sequences in French, such as [lami], and then had to decide whether the target, displayed on the computer screen as a string of letters, such as *amie*, was a French word. If vowel-initial nouns are sufficiently activated by the slope cue introduced in the Co-slope condition, the facilitating effect yielded by the cue should be similar to the identity priming observed in the Vow-nat condition compared to an unrelated baseline.

3. Experiment 2

3.1. Methods

3.1.1. Participants

Thirty-two French native speakers participated. No hearing impairments were reported, and none had participated in Experiment 1.

3.1.2. Materials

Twenty-seven vowel-initial nouns from Experiment 1 (e.g., *l'amie*) were selected as experimental visual targets. Each target was associated with three auditory primes. For the primes, we used the article + noun sequences from Experiment 1, eliminating the contextual part (*C'est…*) with Adobe Audition software. Two of the primes were related to the target, and a third was unrelated; this last prime served as a baseline. For related primes, we extracted the sequence from the Vow-nat condition as one prime and the sequence from the Co-slope as the other related

prime. Half of the unrelated primes were elided sequences of article + vowel-initial nouns (e.g., *l'assaut,* "the assault") and the other half were article + consonant-initial nouns (e.g., *la route*, "the road").

Additionally, 27 nonwords were created as visual targets for the lexical decision task. Similarly, each nonword target was associated with 3 auditory primes. To prevent participants from associating the phonological resemblance with an affirmative response, phonologically related pairs were linked to nonword targets. Two of the related primes consisted of an article + noun ambiguous sequence (e.g., *l'adroite*, 'the skillful one'; *la droite*, 'the right'). A third unrelated prime served as a baseline: half of the primes were elided sequences of article + vowel-initial nouns (e.g., *l'araignée*, 'the spider'), and the other half were article + consonant-initial nouns (e.g., *la tisane*, 'the herbal tea'). One hundred additional targets (half real words, half nonwords) were also presented in unrelated conditions to reduce the proportion of related pairs to 25%.

3.1.3. Procedure

The experiment was performed with E-Prime software. Each participant was presented with all 3 priming conditions (Vow-nat, Co-slope and Baseline) but saw each target only once. Fifty milliseconds after the offset of the auditory prime, the target was displayed in lowercase letters in the center of a computer screen. Participants had to indicate whether the visual target was a real word or a nonword by pressing one of the two response buttons as quickly and accurately as possible. The computer clock was activated when the target was displayed on the screen and stopped when the participant responded. Response latencies and errors were collected.

3.2. Results

We calculated the reaction times (RTs) starting from the beginning of the visual target presentation up to the button-press time point. For statistical analysis, only data from words were taken into account. RTs ±2SDs per condition/participant were removed (4.5%) as well as 4.7% of responses errors. We performed one-way repeated measure ANOVAs with 3 levels (Vow-nat, Co-slope and Baseline).and with the subjects (*F1*) and items (*F2*) as random factors.

The results are presented in Table 1. RT analyses highlighted a significant effect of the priming condition ($F1(2,62)=4.3$, $p=.018$; $F2(2,26)=10.056$, $p=.0002$). The Vow-nat condition had a significantly greater facilitating effect than the Baseline condition ($F1(1,31)=6.074$, $p=.019$; $F2(1,26)=14.3$, $p=.0008$). This effect was also found for the Co-slope condition relative to the Baseline condition ($F1(1,31)=4.95$, $p=.034$; $F2(1,26)=14.5$, $p=.0008$). No significant difference was shown between the Co-slope and Vow-nat conditions (both Fs>1).

Table 1: *Reaction times (RTs, in milliseconds), standard deviations (SDs) and error percentages for the three priming conditions in Experiment 2.*

	Priming conditions		
	Vow-nat	Co-slope	Baseline
RT	570	581	626
SD	146	117	107
% error	2.8%	2.1%	9.4%

3.3. Discussion

These results show that listeners correctly activated the targets (e.g., *amie*) under the intended conditions (e.g., vowel-initial: *l'amie*). Crucially, we found evidence that this effect also arises when only the F0 slope of the first vowel (e.g.,/a/ in *la mie*) is changed to match that of the first /a/ of the vowel-initial homophone (e.g., *l'amie*). This finding suggests that the increase in responses to vowels observed in Experiment 1 for the Co-slope condition is indeed linked to online segmentation and the boost of activation of vowel-initial lexical competitors.

4. General discussion and conclusion

This study explored the cues that can be used for speech segmentation in French. We focused on two F0 characteristics, the mean and the slope, which belong to the initial phoneme of content words. We evaluated their role in segmentation in French. It has been previously established that the F0 is used during processing. Here, we extended this result by verifying the engagement of each characteristic and evaluating the relative weight of each characteristic.

We used sequences consisting of article+noun that were homophonic and could be segmented in two different ways, such as *l'amie* vs *la mie*, both /lami/. We performed three different types of manipulations of the F0 contour of the first vowel /a/ on consonant-initial members of the sequences (e.g., *l<u>a</u> mie*). The manipulations concerned the F0 slope, the F0 shift of the mean value, and both aspects simultaneously (slope+shift).

Remarkably, our results in the off-line task revealed that all these acoustic manipulations enhanced the segmentation of vowel-initial members. Indeed, in the first experiment we showed that when we applied the F0 characteristics of vowel-initial words (such as *l'amie*) to a consonant sequence such as *la mie*, the number of times vowel-initial words were recognized increased. More specifically, we observed that while the combination of the slope and the mean was the most effective change, it was still "incomplete" and did not yield the value observed when the natural vowel-initial content word was presented. This implies that other cues are also involved in the segmentation process. Through the modification of either the slope or the mean of the sequences of consonant-initial content words, it seems that the shift in the mean is more effective than the rotation of the slope, with a difference of 9%. However, slope rotation alone still yields an 8% difference in response changes. These results highlight that the manipulation of the F0 slope influences participants' perception during online segmentation, causing them to activate more vowel-initial competitors.

The RTs from Experiment 2 show that homophonic utterances in which the slope is modified prime both the vowel target and repetition. This suggests that the activation of vowel-initial content words is enhanced by the presence of the slope cue and that such activation is sufficient to produce a facilitating effect that is not different from identity priming.

Our study extends existing evidence of the role of intonational cues in word segmentation in French. In particular, we provide direct evidence of the participation of the F0 slope in segmentation processes.

5. Acknowledgements

This research was supported by a grant from Région Rhône-Alpes (ARC 2).

6. References

[1] Ito, K., & Strange, W. Perception of allophonic cues to English word boundaries by Japanese second language learners of English. *The Journal of the Acoustical Society of America. 125*, 2348–2360, 2009.

[2] Spinelli, E., McQueen, J., & Cutler, A. Processing Resyllabified Words in French. *Journal of Memory and Language, 48*, 233–254, 2003.

[3] Shatzman, K. B., & McQueen, J. M. Segment duration as a cue to word boundaries in spoken-word recognition. *Perception & Psychophysics, 68*(1), 1–16, 2006.

[4] Cho, T., & Keating, P. Articulatory and acoustic studies of domain-initial strengthening in Korean. *Journal of Phonetics, 29*, 155–190, 2001.

[5] Quené, H. Integration of acoustic-phonetic cues in word segmentation. In M. E. H. Schouten (Ed.). *The auditory processing of speech: From sounds to words*. Berlin: Mouton de Gruyter, pp. 349–356, 1992.

[6] Cutler, A., & Norris, D. The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 14*(1), 113–121, 1988.

[7] van Donselaar, W., Koster, M., & Cutler, A. Exploring the role of lexical stress in recognition. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 58*(A), 251–273, 2005.

[8] Reinisch, E., Jesse, A., & McQueen, J. M. (2010). Early use of phonetic information in spoken word recognition: Lexical stress drives eye-movements immediately. *Quarterly Journal of Experimental Psychology, 63*(4), 772−783, 2010.

[9] Xu, Y. (2011). Speech prosody: A methodological review. Journal of Speech Sciences 1(1), 85–115.

[10] Moore, BC. *An introduction to the psychology of hearing*. 6th Edition. Bingley: Emerald Group, 2012.

[11] Dawson C, Aalto D, Simko J, Vainio M. The influence of fundamental frequency on perceived duration in spectrally comparable sounds. *PeerJ 5*:e3734, 2017.

[12] Tremblay, A., Cho, T., Kim, S., & Shin, S. Gradient Effects of Tonal Scaling in the Segmentation of Korean Speech: An Artificial-Language Segmentation Study. *Proceedings of the International Conference on Speech Prosody*. Vol. 2018, pp. 65−69, 2018.

[13] Spinelli, E., Welby, P., & Schaegis, A. L. Fine-grained access to targets and competitors in phonemically identical spoken sequences: The case of French elision. *Language & Cognitive Processes, 22*, 828−859, 2007.

[14] Spinelli, E., Grimault, N., Meunier, F., and Welby, P. An intonational cue to word segmentation in phonemically identical sequences. *Attention Perception & Psychophysics 72,* 775–787, 2010.

[15] Do Carmo-Blanco, N., Hoen, M., Pota, S., Spinelli, E., Meunier, F. Processing of non-contrastive subphonemic features in French homophonous utterances: An MMN study. *Journal of Neurolinguistics*, *52*, 100849, 2019.

[16] Tremblay, A., Cho, T., Kim, S., and Shin, S. Phonetic and phonological effects of tonal information in the segmentation of Korean speech: An artificial-language segmentation study. *Applied Psycholinguistics, 40*, 1–20, 2019.

[17] Kawahara, H., Masuda-Katsuse, I., & de Cheveigné, A. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech communication, 27*(3), 187–207, 1999.

[18] Gow, D. W., Jr., & Gordon, P. C. Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance, 21*(2), 344–359, 1995.

[19] Weber, A., and Scharenborg, O. Models of spoken-word recognition. *Wiley Interdisciplinary Reviews: Cognitive Science, 3*, 387–401, 2012.

[20] McClelland, J. L., & Elman, J. L. The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86, 1986.

[21] Norris, D. G. Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*, 189–234, 1994.