Speech Prosody 2022
23-26 May 2022, Lisbon, Portugal

Temporal expectations and the interpretation of timing cues to word boundaries

Laurence White[1], *Sven Mattys*[2], *Sarah Knight*[2], *Tess Saunders*[3], *Laura Macbeath*[1]

[1]School of Education, Communication and Language Sciences, Newcastle University, UK
[2]Department of Psychology, University of York, UK
[3]School of Social Sciences, Nottingham Trent University, UK

laurence.white@newcastle.ac.uk, sven.mattys@york.ac.uk,
sarah.knight3@york.ac.uk

Abstract

In many languages, speech sounds adjacent to prosodic boundaries are lengthened. Moreover, listeners – particularly learners – exploit word-initial consonant lengthening to locate word boundaries, whilst phrase-final lengthening indicates upcoming prosodic breaks and conversational transitions. How lengthening is detected in the temporally-linear speech stream remains unclear, however.

We investigated listeners' possible use of predictive mechanisms to generate hypotheses about upcoming segment durations and thereby interpret deviations from temporal expectations (specifically, lengthening) as linguistically meaningful. In a series of nonword segmentation experiments, listeners heard 90 twelve-syllable nonsense streams, with – on half the trials – trisyllabic nonword targets embedded (e.g., *dumipako<u>libeku</u>binudafolu*). Segment duration was systematically varied. On the 45 target-present trials, targets were early, medial or late in the nonsense stream (but at least two syllables from utterance edges). Listeners had to respond quickly and accurately when detecting targets.

As expected, target-initial consonant lengthening boosted detection, but this was strongly conditioned by target location within utterances. Specifically, differential effects of timing on detection were much stronger for utterance-late targets than utterance-medially (with uniformly poor utterance-early performance). We explore the factors contributing to this (initially unexpected) pattern, in particular, the hypothesis that predictions about segment duration rely on sufficient experience of foregoing speech rate.

Index Terms: speech timing, word segmentation, temporal prediction

1. Introduction

Speech timing provides listeners with information about the prosodic structure of spoken utterances. Thus, in many languages, consonants are longer in word onsets than word-medially [1], [2], with greater lengthening after phrase boundaries [3]. Similarly, vowels and coda consonants are longer in phrase-final syllables [4], [5]. Furthermore, listeners use these localised lengthening effects to detect boundaries, thus guiding segmentation of utterances into words and phrases. Word-final vowel lengthening is a cue to upcoming boundaries in some languages, although appears not to be universally used for segmentation, at least at the word level [6], [7], probably due to other language-specific functions of vowel duration. Onset consonant lengthening may be a widespread cue, cross-linguistically, to preceding lexical boundaries [7], [8].

A key question concerns how listeners are able to judge boundary-adjacent segments as being lengthened, thus allowing them to interpret available timing cues for segmentation. Our working hypothesis is that listeners use foregoing speech rate to generate expectations about segment duration, providing a benchmark against which to judge timing deviations [9]. Thus, it has been shown that *relative* segment duration, assessed against foregoing rate, affects listeners' perception of the phonemic composition of segmental sequences, with faster foregoing rate making reduced segments more likely to be perceived and, conversely, slower foregoing rate leading certain optional segments to be overlooked [10]. Similarly, syllable duration judged relative to foregoing rate affects listeners' perceptions of lexical stress placement [11].

Listeners' tracking of speech rate appears to develops over time: in particular, judgements about the presence or absence of reduced segments are more strongly affected by the average rate of recorded speech as exposure time to that speech increases [12]. Such findings have been linked to work on the entrainment of neural oscillations in the auditory cortex, specifically in the 4-7 Hz theta band, to the amplitude envelope of speech, as indexed by utterance-specific resetting of oscillator phase [13].

Notwithstanding the established importance of foregoing rate for identification of speech sounds and lexical stress, the utterance-level dynamics of speech-rate tracking for lexical segmentation have barely been investigated. One recent finding has, however, provided evidence for a cumulative role of foregoing rate in the interpretation of durational cues to speech structure [14], [15]: in an experiment designed to test a new paradigm for investigating the use of prosodic cues to speech structure (henceforth referred to as the pilot study), we found that listeners used word-initial consonant lengthening for segmentation of words embedded late in utterances, but not in earlier utterance positions.

1.1. Pilot study

The full design – replicated in the present study – is explained below, but the key features of the pilot study [14], [15] are that trisyllabic nonword targets were embedded early, medially or late in 12-syllable nonsense utterances ("test phrases"). After each synthesised test phrase, listeners heard a trisyllabic probe and had to judge whether it had been embedded in the foregoing test phrase. A range of timing conditions were used, relative to a baseline condition (Flat) in which all test-phrase segments had the same duration. In line with the timing studies reviewed above, target-initial consonant lengthening (C1) and target-final vowel lengthening (V3) were expected to boost listeners' detection of embedded words, whilst lengthening of target-medial vowels (V1) and

10.21437/SpeechProsody.2022-66

consonants (C2) should have inhibited detection by implying a boundary within the target trisyllable rather than at its edge.

The placement of targets at three different positions within the test phrase was included to forestall listeners' strategic attendance to a certain section of the phrase as the experiment proceeded, but there was an unexpected effect of position on target detection accuracy. As shown in Figure 1, detection of early targets was at chance level (50%). It was above chance in medial and final targets; however, in neither position were there the expected differential effects of timing cues. The exception was the C1 condition in late targets, where detection accuracy was reliably better than all other late-target timing conditions.

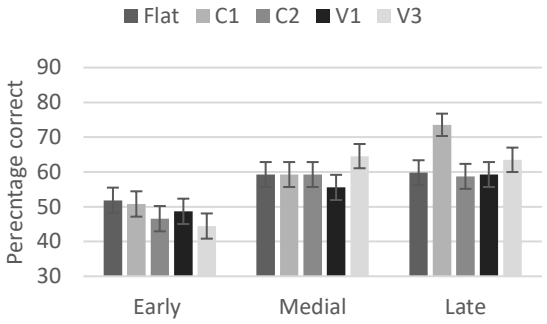

Figure 1: *Pilot study target detection accuracy by timing condition (Flat, C1, C2, V1, V3) and test-phrase target position (Early, Medial, Late).*

We interpreted the chance-level detection of early targets as a straightforward memory effect: in a stream of nonsense syllables, there is little opportunity for higher-level processing of input and thus the memory trace of early syllables decays in the phonological loop [16] as later syllables are heard.

The differential effect of timing according to target position in the utterance was unexpected. Based on earlier studies using other paradigms, our hypothesis was that word-initial consonant lengthening (C1) would consistently boost listeners' detection of targets, but the expected benefit was only found in utterance-late targets. Our post hoc interpretation – in line with the "foregoing speech rate" effects reviewed above – was that listeners require some minimum amount of utterance exposure through which to calibrate speech rate and thus to set a temporal baseline against which lengthening effects can be detected.

Therefore, the present study is in part a replication of the pilot study. Our interpretation of the differential effect of timing cues according to position was post hoc and thus we need to validate it in a replication. Because of this, the systematic variation of target position within utterances is retained in the new design, as are the Flat and C1 timing conditions.

A further unexpected outcome of the pilot study was that other timing conditions had no differential effects, even in the utterance-late position. Target-final vowel lengthening (V3) had been expected to promote target detection, but we provisionally conclude that listeners may benefit more, in this task design, from cues to target onset than to target offset. Thus, in the present study, we lengthened the vowel immediately preceding the target (V0), thus cueing the pre-target boundary, and included another condition where both the pre-target vowel and the target-initial consonant were lengthened (V0C1).

Target-medial vowel (V1) and consonant (C2) lengthening were expected to inhibit target detection in the pilot by cueing an internal boundary. It is unclear why this was not observed, so the present study also included a double timing cue to a target-medial boundary, with both elements of a vowel-consonant diphone (V1C2) lengthened. Thus, the present study is both a replication and an extension of the pilot.

2. Method

2.1. Participants

Participants (N=45) were all native English speakers, at least 18 years old, with no reported speech, hearing or language difficulties. Multilingual speakers were included if English was their first language. A small number of additional participants were excluded due to problems during the testing session, such as noise or failure to understand the procedure. Moreover, the original intention had been to collect data from 63 participants, but laboratory studies had to be halted due to unexpected external circumstances, so we analysed the available data with some caution (see discussion, below, of the statistical analyses).

2.2. Design

The experiment had the same essential design as the pilot study [14], [15], the only difference being in the timing conditions detailed below. Thus, nine trisyllabic nonword targets were constructed from nine consonants and five vowels (tense monophthongs and diphthongs), creating 27 different CV syllables. No phonemes were repeated in any trisyllabic target.

Targets were arbitrarily split into three groups (Table 1) for the purpose of creating the carrier utterances in which targets were embedded. For each target group, a set of nine-syllable carrier utterances was created by random sampling of the other two groups' target syllables. Each 12-syllable sequence of carrier utterance and embedded target – together comprising the *test phrase* – thus contained no repeated syllables.

Table 1: *Nonword targets.*

Group 1	Group 2	Group 3
makolu	*fusame*	*bupedi*
ponubi	*kefiso*	*nedupa*
silona	*libeku*	*damifo*

There were three possible positions for the embedded target within the carrier utterance, indexed here by the ordinal position of the target's first syllable within the full test phrase sequence: *early* – syllable 3; *medial* – syllable 5; *late* – syllable 8. Examples are shown below, with targets underlined. The same carrier sequence is shown here purely for clarity of illustration (each carrier utterance actually comprised a new random sample of non-target-group syllables).

- **Early**: *dumi<u>libeku</u>pakobinudafolu*
- **Medial**: *dumipako<u>libeku</u>binudafolu*
- **Late**: *dumipakobinuda<u>libeku</u>folu*

There were five timing conditions. In the **Flat** condition, all test phrase segments had the same duration of 120ms. In the other four conditions, specific segments in the embedded target, or immediately preceding it, were longer (170ms) than the rest of the segments in the test phrase (120ms):

- **C1:** target first-syllable onset consonant.
- **V0:** carrier-utterance vowel immediately preceding target.

- **V0C1:** carrier-utterance vowel immediately preceding target, plus target first-syllable onset consonant.
- **V1C2:** target first-syllable vowel rhyme and second-syllable onset consonant.

In the experiment, each carrier-plus-target test phrase was followed by a *target probe*. In *target-present* trials, this probe was the same trisyllabic sequence as the embedded target. In the *target-absent* trials, each test-phrase-embedded target was changed by only its final syllable, which was substituted for another CV syllable that did not otherwise appear in that test phrase sequence. Other than this target-final syllable, target-present and target-absent trials were matched pairwise, including on utterance position and timing conditions, and with the same target probe.

Test phrases and target probes were diphone synthesised with MBROLA [17], voice *en1* (British English male). Test phrase fundamental frequency was 120 Hz throughout. To avoid an exact acoustic match with test-phrase-embedded targets, target probes were synthesised at 105 Hz throughout. The duration of all target-probe segments was 120ms.

Three sets of experimental trials were constructed, with the utterance position of each group of three nonword targets being fixed within a trial set. Thus, in one trial set, the three Group 1 targets (*makolu*, *ponubi*, *silona*) were always utterance-early, Group 2 targets (*fusame*, *kefiso*, *libeku*) utterance-medial and Group 3 targets (*bupedi*, *nedupa*, *damifo*) utterance-late. Group position was systematically rotated between the three sets. All target words were presented in all five timing conditions within each set. Each target-present trial was paired with an equivalent target-absent trial.

There were 90 trials per participant: 3 targets x 3 utterance positions x 5 timing conditions x target-present *vs* target-absent. For each trial set, three different pseudo-random trial orders were prepared, with no more than three target-present or target-absent trials in a row and with the same target probe never used on successive trials. Thus there were nine 90-trial scripts in total: 3 trial sets x 3 pseudo-random orders. Five participants were assigned to each of the nine scripts.

2.3. Procedure

Participants wore headphones adjusted to a comfortable volume. On each trial, participants first heard the 12-syllable test phrase, followed 1000ms after its offset by the trisyllabic target probe. Immediately after probe offset, the words "absent" and "present" were displayed. Participants were instructed to press the right shift key if they had heard the probe in the test phrase and the left shift key otherwise. Both speed and accuracy were encouraged.

Prior to the experimental trials, there were three practice trials, one selected from each experimental set. None of the practice trials were repeated within the first five experimental trials. After the practice trials, participants had an opportunity to check the procedure with the experimenter before beginning the experimental trials, which were presented in two blocks, with an opportunity for a break halfway through.

2.4. Statistical analysis

Analyses, based on the 45 target-present trials, employed the *lme4* package in R [4]. Given the relatively small sample size, the regression model protocol was designed to result in maximally *parsimonious* final models: thus, terms that were not found to contribute significantly to model fit were removed.

Correct-incorrect response data were analysed with generalised linear mixed-effects models, with a binomial distribution and logit link. The initial model included fixed factors of Timing, Utterance Position, and their interaction, and random slopes, by Subjects and by Targets, for the fixed factors. However, this random-effects structure produced a singular fit, indicating overfitting. Random effects were thus systematically tested for contribution to model fit and removed where model comparisons (likelihood ratio tests) indicated redundancy. Order of random effect removal was determined by Akaike information criterion (AIC) values: in each case, the reduced model producing the greatest drop in AIC was used as the new baseline for further nested comparisons. The final random-effects structure included Subject and Target intercepts, and random by-Subject and by-Target slopes for Utterance Position. Fixed effects were then tested with model comparisons using likelihood ratio tests. No fixed effects were removed.

Response latency analyses were based on correct-response target-present trials. Latencies shorter than 250ms or slower than a participant's mean plus 2.5 standard deviations were removed as outliers. The effects of Timing, Utterance Position and their interaction on latencies were analysed using linear mixed-effects models. Treatment of random effects and model comparisons were otherwise as for the correct-incorrect response data; here the final, implemented model had a by-Subject intercept as the only random term and Utterance Position as the only fixed effect (see below).

For accuracy and latency, post hoc pairwise comparisons (Tukey-adjusted) were calculated with the *emmeans* package.

3. Results

3.1.1. Correct target-detection response rates

Mean correct response rates are shown in Figure 2. The response profile was somewhat similar to that of the pilot (Fig. 1 [14], [15]): utterance-early target detection was at/below chance; utterance-medial target detection was above chance, but varied little by timing condition; utterance-late target detection was above chance and varied by timing condition.

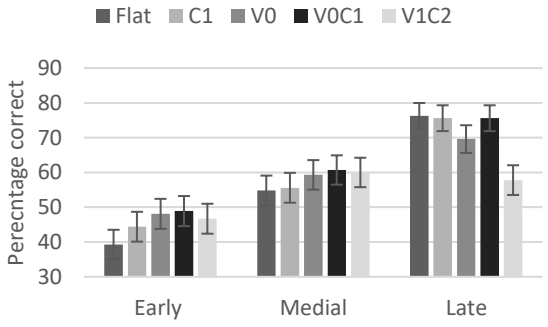

Figure 2: *Target detection accuracy rates by timing condition (Flat, C1, V0, V0C1, V1C2) and test-phrase target position (Early, Medial, Late).*

Model comparisons showed an effect of Utterance Position on correct response rates, $X^2(2) = 31.66$, $p < .001$, an effect of Timing, $X^2(4) = 17.68$, $p = .001$, and an Utterance Position x Timing interaction, $X^2(8) = 17.87$, $p = .022$.

Accuracy was below chance for utterance-early targets, 45%, $t(44) = -2.031$, $p = .048$, but above chance for utterance-

medial targets, 58%, $t(44) = 2.909$, $p = .006$, and utterance-late targets, 71%, $t(44) = 9.384$, $p < .001$. We also note that overall correct response rates were reliably greater for utterance-late than utterance-medial targets, $t(44) = 4.025$, $p < .001$.

The finding that performance was at chance when targets were early in the test phrase accords with that of the pilot study [14], [15]. In both cases, this pattern is probably indicative of short-term memory effects, specifically the decay of auditory traces within the phonological loop, with rehearsal blocked by subsequent input [16].

Given the Timing x Utterance Position interaction, and in line with our primary experimental question, we tested for differential timing effects separately for the three utterance positions. For both utterance-early and utterance-medial targets, there were no pairwise differences between any of the timing conditions ($p > .10$). For utterance-late targets, however, performance on the V1C2 condition was reliably poorer than on the three other timing conditions:

- *vs* C1, $z = 3.293$, $p = .009$.
- *vs* Flat, $z = 3.434$, $p = .005$.
- *vs* V0C1, $z = 3.293$, $p = .009$.

There were no other pairwise differences between timing conditions for utterance-late targets ($p > .10$).

Thus, as in the pilot study [14], [15], differential effects of timing on accuracy were found only when targets were late in the test phrases. In the present data, however, the key effect was that the performance was poorest in the condition (V1C2) in which timing cues were expected to inhibit target detection, rather than (as in the pilot) a performance boost due to the word-initial cues expected to promote target detection (e.g., C1).

The lack of differential timing effects for utterance-early targets is explained by the task being too difficult: indeed, accuracy was below chance when targets were followed by a further seven syllables. Listeners were able to detect targets in utterance-medial position, however, so the lack of differential timing effects there requires another explanation. Below we consider possible mechanisms that could explain the selective power of timing cues for targets late in the test phrase.

3.1.2. Correct target detection response latencies

Correct target-present response latencies are shown in Table 2. A full fixed-effects regression model – Timing, Utterance Position and their interaction – showed no significant effects. A backwards stepwise procedure, based on likelihood ratio tests and AIC values, was thus used to remove non-significant fixed effects (see above). This generated a model with Utterance Position as the only predictor, $X^2(2) = 22.35$, $p < .001$.

Post hoc tests indicate that responses were faster to targets in utterance-late position, M=930ms, than in utterance-early position, M=1099ms, $t = -4.028$, $p < .001$, and utterance-medial position, M=1082ms, $t = -3.968$, $p < .001$, but there was no difference between utterance-early and utterance-medial targets, $t = 0.359$, $p > .10$.

This pattern of response latencies, like that for correct scores, indicates that the task was easier when targets were late in the test phrase than otherwise. There is no evidence of a trade-off between speed and accuracy, as both were better for utterance-late targets.

The shorter response latencies for utterance-late targets are likely due to recency. Listeners will tend to have more response confidence to the probe where less time intervenes after hearing the prior target. The lack of a Timing effect or a Timing x Utterance Position interaction, in contrast to the accuracy rate analyses, might be due to the relative insensitivity of this offline task to subtle cue effects. Below we discuss a modified task with an online response to target detection.

Table 2: *Response latencies (ms) for target detection*

	Utterance Position		
Timing	*Early*	*Medial*	*Late*
Flat	1156	1036	964
C1	1192	1044	943
V0	1085	1155	837
V0C1	1095	1097	909
V1C2	967	1077	999

4. Discussion

This study's primary purpose was to validate the unexpected finding [14], [15] that the impact of timing cues to word boundaries was mediated by the position of targets within utterances. A similar differential effect was also found here: while both utterance-medial and utterance-late targets had above-chance detection accuracy, differences between timing conditions only emerged for utterance-late targets.

This finding reinforces the earlier post hoc interpretation of the differential power of timing cues in utterance-late targets. It accords with studies of the role of foregoing speech rate in mediating judgements of the presence of reduced syllables [18] or lexical stress [11]. The proposed mechanism is that the listener uses the foregoing utterance rate to gradually build up expectations about segment duration. In the current case, the salience of longer-than-expected utterance-late segments would license their interpretation as localised cues to word boundaries; however, utterance-medial timing cues would not be preceded by sufficient phonetic material to effectively calibrate rate.

Differential cue effectiveness contrasted with the pilot data, where initial consonant lengthening (C1) boosted accuracy over all other conditions. Here, both conditions which included C1 lengthening had high accuracy, but performance was no better than for the Flat baseline. The differential effect found here was lower accuracy where target-internal segments were lengthened (V1C2): this should promote perception of a *target-internal* boundary and, indeed, appeared to inhibit target detection.

This could be interpreted as a recency effect, as these cues occur slightly later than the positive boundary cues (C1, V0C1). C1 phrases had identical timing profiles in the pilot and boosted accuracy there, however, while later V1 and C2 cues were not detrimental relative to baseline. Moreover, the latest-occurring V3 cue had no impact relative to baseline in the pilot. It may be that the high performance in the flat baseline here is anomalous, but if so, explanations for that anomaly are not apparent.

To separate memory effects from the hypothesised priming-by-foregoing-rate mechanism, we have run a further study [14], with a similar design, except that probes were presented *prior* to test phrases. Listeners responded immediately after detecting targets and thus had no variation in *target* memory demands by utterance position. Further evidence of differential timing effects in utterance-late targets would strongly reinforce the working hypothesis that temporal expectations are mediated by foregoing speech rate, given sufficient exposure.

5. References

[1] D. K. Oller, 'The effect of position in utterance on speech segment duration in English', *The Journal of the Acoustical Society of America*, vol. 54, no. 5, pp. 1235–1247, 1973.

[2] P. Keating, T. Cho, C. Fougeron, and C.-S. Hsu, 'Domain-initial articulatory strengthening in four languages', in *Phonetic Interpretation: Papers in Laboratory Phonology VI*, J. Local, R. Ogden, and R. Temple, Eds. Cambridge: Cambridge University Press, 2004, pp. 143–161.

[3] C. Fougeron and P. A. Keating, 'Articulatory strengthening at edges of prosodic domains', *The journal of the acoustical society of America*, vol. 101, no. 6, pp. 3728–3740, 1997.

[4] A. E. Turk and S. Shattuck-Hufnagel, 'Multiple targets of phrase-final lengthening in American English words', *Journal of Phonetics*, vol. 35, no. 4, pp. 445–472, 2007.

[5] D. H. Klatt, 'Linguistic uses of segmental duration in English: Acoustic and perceptual evidence', *The Journal of the Acoustical Society of America*, vol. 59, no. 5, pp. 1208–1221, 1976.

[6] M. Ordin, L. Polyanskaya, I. Laka, and M. Nespor, 'Cross-linguistic differences in the use of durational cues for the segmentation of a novel language', *Mem Cogn*, vol. 45, pp. 863–876, Mar. 2017, doi: 10.3758/s13421-017-0700-9.

[7] L. White, S. Benavides-Varela, and K. Mády, 'Are initial-consonant lengthening and final-vowel lengthening both universal word segmentation cues?', *Journal of Phonetics*, vol. 81, p. 100982, Jul. 2020, doi: 10.1016/j.wocn.2020.100982.

[8] L. White, S. L. Mattys, L. Stefansdottir, and V. Jones, 'Beating the bounds: Localized timing cues to word segmentation', *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 1214–1220, 2015, doi: 10.1121/1.4927409.

[9] L. C. Dilley and M. A. Pitt, 'Altering context speech rate can cause words to appear or disappear', *Psychological Science*, vol. 21, no. 11, pp. 1664–1670, Nov. 2010, doi: 10.1177/0956797610384743.

[10] T. H. Morrill, M. Baese-Berk, C. Heffner, and L. Dilley, 'Interactions between distal speech rate, linguistic knowledge, and speech environment', *Psychonomic Bulletin & Review*, vol. 22, no. 5, pp. 1451–1457, Oct. 2015, doi: 10.3758/s13423-015-0820-9.

[11] E. Reinisch, A. Jesse, and J. M. McQueen, 'Speaking rate affects the perception of duration as a suprasegmental lexical-stress cue', *Language and Speech*, vol. 54, no. 2, pp. 147–165, Jun. 2011, doi: 10.1177/0023830910397489.

[12] M. M. Baese-Berk, C. C. Heffner, L. C. Dilley, M. A. Pitt, T. H. Morrill, and J. D. McAuley, 'Long-term temporal tracking of speech rate affects spoken-word recognition', *Psychol Sci*, vol. 25, no. 8, pp. 1546–1553, Aug. 2014, doi: 10.1177/0956797614533705.

[13] H. Luo and D. Poeppel, 'Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex', *Neuron*, vol. 54, no. 6, pp. 1001–1010, Jun. 2007, doi: 10.1016/j.neuron.2007.06.004.

[14] L. White, S. Mattys, S. Knight, T. Saunders, and L. Macbeath, 'Nonword segmentation suggests tracking of speech rate to interpret timing cues to word boundaries', in prep.

[15] L. White, S. Benavides-Varela, K. Mády, and S. Mattys, 'The primary importance of onsets: Timing and prediction in speech segmentation', presented at the 3rd Phonetics and Phonology in Europe conference, Lecce, 2019.

[16] A. D. Baddeley, N. Thomson, and M. Buchanan, 'Word length and the structure of short-term memory', *Journal of Verbal Learning and Verbal Behavior*, vol. 14, no. 6, pp. 575–589, 1975.

[17] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. Van der Vrecken, 'The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes', in *Proceedings of the Fourth International Conference on Spoken Language Processing*, Philadelphia, 1996, vol. 3, pp. 1393–1396.

[18] T. H. Morrill, L. C. Dilley, J. D. McAuley, and M. A. Pitt, 'Distal rhythm influences whether or not listeners hear a word in continuous speech: Support for a perceptual grouping hypothesis', *Cognition*, vol. 131, no. 1, pp. 69–74, Apr. 2014, doi: 10.1016/j.cognition.2013.12.006.