Lengthened Consonants are Interpreted as Word-Initial

Laurence White[1], Sven Mattys[2], Linda Steffansdottir, Victoria Jones

[1]School of Psychology, Plymouth University, Plymouth, UK
[2]Department of Psychology, University of York, York, UK

laurence.white@plymouth.ac.uk, sven.mattys@york.ac.uk

Abstract

Prosody facilitates listeners' segmentation of the speech stream into a sequence of words and phrases. With regard to speech timing, vowel lengthening is interpreted as a cue to an upcoming boundary, in accordance with the iambic-trochaic law. However, the impact of consonant lengthening on segmentation, in the absence of other boundary cues, has not been tested.

In a series of artificial language learning experiments, we examined how durational variation affects listeners' extraction of novel trisyllables defined by transition probabilities. In line with previous research, syllables containing lengthened vowels were interpreted by listeners as word-final. However, syllables with lengthened onset consonants were interpreted as word-initial. Thus, the structural interpretation of durational variation depends upon localization: longer vowels cue a following boundary; longer consonants cue a preceding boundary.

Index terms: speech timing, speech segmentation

1. Introduction

Variations in suprasegmental dimensions – pitch, duration, loudness – are consistently associated with speech structure at prosodic heads and edges [1]. Firstly, the heads of prosodic domains – stressed syllables and accented words – are more prominent through a combination of higher pitch, greater duration and greater loudness, although the relative contribution of these dimensions is language-specific. Secondly, boundaries between words and between phrases are associated with intonational and durational variation. Boundary-adjacent intonational contours vary between languages [2], but upcoming boundaries may be universally associated with segmental lengthening [1]. Indeed, the slowing of articulation as boundaries approach has been associated with a non-linguistic principle of deceleration at the end of motor sequences [3,4].

It is well established that listeners use suprasegmental variations to segment speech into words and phrases [5,6,7]. Considering specifically speech timing, lengthened vowels are interpreted as word-final in artificial language streams [8]. Similarly, with natural language stimuli, long stressed syllables were more likely to be interpreted as monosyllabic words rather than the start of disyllables (e.g., *ham* vs *hamster* [9]). Additionally, greater magnitude of syllable lengthening is associated by listeners with higher-level phrase or utterance boundaries [7].

Such results are in line with the iambic-trochaic law [10], which proposes that the interpretation of prosodic salience depends on its phonetic realization: in particular, sounds made salient through greater loudness are recognized, other things being equal, as sequence-initial, whilst sounds made salient through lengthening are perceived as final.

Despite English and French differing markedly in the distribution and realization of phrasal prominence, support for the iambic-trochaic law was found with both English and French listeners, and with speech and non-speech sounds [11]. Expanding salience cues to consider also pitch, high-low disyllables were better recalled than low-high; thus, pitch salience, like loudness, is interpreted as sequence-initial [12]. However, considering durational contrast in the same study, recall of disyllabic sequences was better when second syllables had longer vowels compared to when both vowels had similar durations or when first syllables had the longer vowel [12].

For native English-speaking listeners, lengthened syllables promote segmentation of artificial language streams when word-final but not when word-initial [8,13]. In particular, lengthening of vowels in word-initial syllables does not facilitate segmentation for English listeners, despite English having predominantly word-initial stress, associated with segmental lengthening within the syllable, together with pitch excursion, increased loudness and other cues [14]. This raises the question of whether lengthening can only ever serve as a cue to an upcoming boundary rather than a preceding boundary.

Listeners' relative weighting of durational *vs* other cues to prosodic structure (e.g., lexical, segmental, suprasegmental) may be modulated according to language-specific patterns of occurrence [15]. Compared to Dutch, for example, vowel duration in English may be less important for signaling stress than is vowel quality [16, 17]. Additionally, the distribution of lengthening effects within words may serve to disambiguate their different structural interpretations [18, 19]. For example, analysis of a corpus of English speech showed that the distinct patterns of lengthening within words due to lexical stress and word-/phrase-finality appear sufficient to allow listeners to distinguish the two structural interpretations [20]. If so, lengthening could be disambiguated and reliably used as a pre-boundary or stress cue by listeners, even in the absence of additional segmental and suprasegmental cues.

Of course, a full account of the distribution of durational effects associated with prosodic structure must also include consonantal lengthening. In particular, lengthening of consonants in word-initial position is consistently observed in several studied languages. For English, syllable onset consonants are substantially longer when uttered in word-initial position than word-

10.21437/SpeechProsody.2014-82

medially [21], an effect subsequently observed in French, Korean, and Taiwanese [22]. Whilst multiple consonants within the onset may be lengthened, the durational effect does not extend to the vowel nucleus of that syllable [18, 21]. As with lengthening of domain-final vowels, the magnitude of the consonant-lengthening effect increases at higher prosodic boundaries [23, 24], although consonants may be as short in absolute utterance-initial position as when word-medial, perhaps in part because the termination of silence serves as an unambiguous cue to prosodic structure [18].

Several studies have investigated the impact of consonant length, in conjunction with other durational and segmental cues, on word -level segmentation [5, 25, 26, 27]. Typically, these studies have manipulated a within-task contrast between consonants in initial and medial/final position. For example, in segmentally ambiguous phrases like *two lips* vs *tulips* (near-homophonous in American English), a longer consonant in word-initial position (/l/) encouraged cross-modal priming of *lips* [25]. Also for English, word-initial consonant lengthening, together with word-final vowel lengthening and other naturally occurring cues to prosodic boundaries, affected the interpretation of ambiguous sequences such as *paper* vs *pay per* in adults and infants as young as ten months [5, 6].

In Dutch, listeners' interpretation of segmentally ambiguous sequences like *diep in* vs *die pin* was affected by duration of the pivotal consonant, which tended to be interpreted as word-initial when relatively long [26]. Consonant duration affected Italian listeners' judgements regarding both lexical segmentation and identification of geminates *vs* singletons [27], whilst French listeners interpreted longer consonants as more likely to be word-initial than in liaison context (e.g., *dernier oignon* vs *dernier rognon*, [28]). There is also evidence that phrase-initial lengthening and articulatory strengthening affect listeners' interpretation of the structure of ambiguous phrases [29].

All of the foregoing studies suggest that lengthened consonants tend to be interpreted as word-initial by listeners. However, all used natural speech – sometimes resynthesized to manipulate segment durations – with multiple potential cues to word boundaries. In particular, segmental cues, such as boundary-related allophonic variations, and other suprasegmental cues, including lengthening of word-final vowels, were also available to listeners [26]. Furthermore, participants' awareness of the implicit contrast between the two interpretations of near-homophonous sequences (e.g., *two lips* vs *tulips*) might modulate their use of segmentation cues relative to when there is only one lexical solution available.

To address these confounds, we used an artificial language learning paradigm to focus on lengthening of consonants and lengthening of vowels, both separately and together. Listeners have consistently been shown to be able to learn and subsequently recall novel words from a nonsense speech stream when the syllable-to-syllable transition probabilities within words are higher than those between words [8]. Using such a paradigm, with artificial speech streams created through diphone synthesis, we obviate the need to use near-homophonous sequences from natural languages, and eliminate the presence of other potential cues to word boundaries. This allows us to focus precisely on the question: does longer duration increase the tendency for consonants to be interpreted as word-initial?

2. Experiment 1

2.1. Method

Three durational manipulations of two artificial languages were used to determine the impact of consonantal lengthening on segmentation, and thereby on subsequent recall, of statistically-defined words. We predicted that words should be better recalled when word-initial consonants are lengthened during language exposure, relative to when all consonants had the same duration or when word-medial consonants were lengthened.

2.1.1. Participants

We tested 120 native British English speakers, with no reported speech or hearing problems. They were randomly allocated to the three duration conditions (40 in each condition). Within each duration condition, 20 participants were allocated to Stream 1 and 20 to Stream 2. All participants received a small honorarium or course credit for their participation.

2.1.2. Materials

We prepared two different artificial languages, similar to those used in earlier studies [30], each comprising four trisyllabic words (C1V1-C2V2-C3V3).

Stream 1 words: *pabiku*, *golatu*, *tinudo*, *daropi*

Stream 2 words: *tudaro*, *bikuti*, *golatu*, *nudopa*

Six-minute streams containing these words in random sequence were generated using the *en1* male British English voice in the diphone synthesizer MBROLA [31]. Fundamental frequency was a constant 120Hz. To eliminate the strong segmentation cue of hearing silence at the beginning and end of the stream, the streams were faded in and faded out with five-second ramps.

As each word could be followed by any of the other three words, but not by itself, the transition probability between words was always 1/3. As each syllable only occurred once within the language, the transition probability between within-word syllables was always 1.

Total trisyllabic word duration was kept constant at 720ms, whilst the duration of individual segments was manipulated to generate three "lengthening" conditions.

Flat: All segments – vowels and consonants – were 120ms.

C1: The onset consonant of the first syllable of each word (*pabiku* etc.) was 170ms *vs* 110ms for all other segments.

C2: The onset consonant of the second syllable of each word (*pabiku* etc.) was 170ms *vs* 110ms for all other segments.

In the test phase, following exposure to the six-minute stream, isolated words and foils were played to participants. Test-phase foils were constituted of the syllables of the language, either part-words derived from the end of one word and start of another (e.g., Stream 1: *bikuti* from *pabiku tinudo*) or non-words, syllable strings that never occurred in the language (e.g., *tipala*). The words in Stream 1 were part-words in Stream 2 and *vice*

versa. Words and foils for the test phase were synthesized with flat durational profiles (all segments 120ms) in all three conditions.

2.1.3. Procedure

Participants were told they would hear an artificial language through headphones for six minutes, and that their task was to listen and try to discover the words in the language. After the exposure phase, they were given test phase instructions. In the test phase, they heard 24 pairs of trisyllabic strings, where one string was a word in the language stream just heard and the other string was a part-word or non-word. The two trisyllabic strings were separated by 500ms. For each pair, participants had to press the left shift key on a computer keyboard if the artificial language word was the first string of the pair, and the right shift key if it was the second string.

2.1.4. Statistical analysis

All analyses were carried out on the raw response data – "correct" or "incorrect" – using mixed-effects logistic regression models, including the random factors of subjects and trials (*lmer* package in *R*, [32]). Models were compared using log-likelihood χ^2 tests.

2.2. Results and discussion

Mean correct responses by lengthening condition are shown in Figure 1. In the Flat condition, mean correct was 67%, $z = 5.07$, $p < .0001$, replicating previous findings that listeners can recognize words defined by transition probabilities in novel streams of syllables [8]. Above chance performance was also found in the other lengthening conditions: C1 – 73%, $z = 6.12$, $p < .0001$; C2 – 63%, $z = 4.29$, $p < .0001$.

A logistic regression including fixed factors of Lengthening (*Flat vs C1 vs C2*) and Stream (1 *vs* 2) found a main effect of Lengthening, $\chi^2(2) = 11.59$, $p < .005$. There was also a main effect of Stream, $\chi^2(1) = 4.51$, $p < .05$, with words from Stream 1 recalled better than those from Stream 2. There was no interaction between Lengthening and Stream, $\chi^2(5) = 0.11$, $p = .95$. Thus, the advantage of Stream 1 over Stream 2 was consistent across the three lengthening conditions, and so further pairwise analyses were collapsed across streams.

Lengthening of the consonant in the first syllable (C1) improved performance compared to lengthening of the consonant in the second syllable (C2), $\chi^2(1) = 9.93$, $p < .005$, and compared to the Flat condition, $\chi^2(1) = 4.20$, $p < .05$. There was no difference between *C2* vs *Flat*, $\chi^2(1) = 2.15$, $p = .14$. These results indicate that segmentation of the artificial language was promoted by localized lengthening of the word-initial consonant. Thus, consonantal lengthening appeared to cue listeners to the presence of an immediately preceding boundary.

Lengthening of a vowel in a similar artificial language stream has been shown to act as a cue to a following boundary [8]. This suggests a functional difference in listeners' interpretation of vowel and consonant lengthening. An alternative hypothesis is that longer syllables – whether through greater vowel or consonant duration – tend to be perceived as word-edges, which could be either initial or final. This view is not encouraged by findings that vowel lengthening in word-initial syllables failed to facilitate segmentation relative to no lengthening [8]. However, in that experiment, initial syllable vowels were only lengthened in half of the six artificial words. In order to confidently assert our interpretation – that consonant lengthening, in contrast with vowel lengthening, is a cue to a preceding boundary – we attempted a more direct replication of [8] with our artificial language materials, testing the effect on segmentation of lengthening the first syllable vowel in every trisyllabic word.

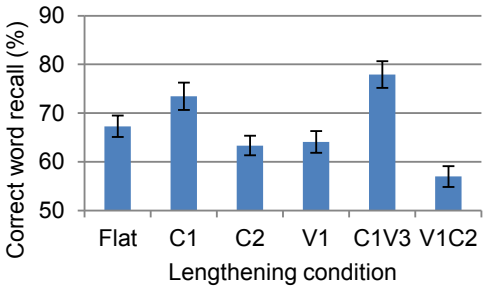

Figure 1: *Mean correct responses and standard errors: Exp1 – Flat, C1, C2; Exp 2 – V1; Exp 3 – C1V3, V1C2.*

3. Experiment 2

3.1. Method

The participants and experimental procedure were as for Experiment 1. However, in Experiment 2, participants heard the artificial language streams with the first vowel of each word lengthened: thus, the underlined vowel in *pabiku* etc. was 170ms vs 110ms for all other segments. This *V1* condition was implemented for both artificial language streams, with 20 participants hearing each one. As before, after exposure to the streams, participants heard 24 pairs of trisyllabic strings – words and foils – with all segments having 120ms duration, and had to choose which string within a pair belonged to the artificial language.

3.2. Results and discussion

Overall mean correct word recognition in the *V1* condition was 64%, reliably above chance, $z = 4.66$, $p < .0001$. To test the hypothesis regarding the localization of durational segmentation cues, the important comparisons are with the *Flat* and *C1* conditions in Experiment 1. Figure 1 shows the mean correct responses for the three critical conditions.

Collapsing, as before, across the two artificial language streams, there was no difference in recognition between the *Flat* and *V1* conditions, $\chi^2(1) = 1.15$, $p = .28$, replicating previous findings that lengthening of the vowel in a word-initial syllable does not serve as a cue to a preceding boundary for English listeners, despite the prevalence of word-initial stress in English [8, 13].

Performance on the *C1* condition was reliably better than the *V1* condition, $\chi^2(1) = 7.57$, $p < .01$. This supports the hypothesis that localization of lengthening is important for segmentation: a lengthened consonant cues a preceding boundary; a lengthened vowel cues a following boundary. In Experiment 3, to explore the power of such cues further, we tested the efficacy of vowel and consonant lengthening in combination. In particular, we examined whether a lengthened vowel

immediately followed by a lengthened consonant was a strong cue to an intervening boundary. We contrasted two cases, one where the lengthened consonant-vowel sequences coincided with the location of boundaries indicated by syllable transition probabilities and the other where the durational and statistical cues conflicted.

4. Experiment 3

4.1. Method

The participants and experimental procedure were as for Experiment 1, but here the duration of the artificial language streams was manipulated in two new conditions. In condition *C1V3*, the first consonant and the final vowel of each word (e.g., *pabiku*) were 160ms, *vs* 100ms for all other segments. In condition *V1C2*, the vowel of the first syllable and the consonant of the second syllable (e.g., *pabiku*) were 160ms, *vs* 100ms for all other segments. This was effectively a composite of the V1 and C2 conditions. Note that the lengthened segments were 160ms and the others 100ms, in contrast with 170ms and 110ms in the other experiments: this was to preserve total word duration at 720ms in all conditions across the three experiments.

4.2. Results and discussion

As shown in Figure 1, performance was reliably above chance in the *C1V3* condition, 78%, $z = 8.15$, $p < .0001$, and the *V1C2* condition, 57%, $z = 2.29$, $p < .05$. However, performance was significantly better in the *C1V3* condition, where the lengthened vowel and consonant straddled a statistically-defined word boundary, $\chi^2(1) = 31.69$, $p < .001$.

Comparison with the earlier experiments showed that performance on *C1V3* was no better than on *C1*, $\chi^2(1) = 1.10$, $p = 0.29$. However, performance on *C1V3* was better than on all other conditions ($p < .001$ in all cases). This may be due to intrinsic performance limitations on this type of language learning task, given the memory component combined with the repeated exposure to words and foils during the 24 two-alternative forced-choice test trials.

Performance in the *V1C2* condition was worse than in all other conditions ($p < .05$ for all comparisons). In this case, the word boundary implied between the lengthened vowel and the subsequent lengthened consonant was incongruent with that defined by transition probabilities. This accords with previous findings that statistically-defined trisyllables that straddled intonationally-defined boundaries in artificial language streams were not well recognized [33].

5. Conclusion

The three experiments demonstrate that segmental lengthening can serve as a cue to both preceding and following prosodic boundaries, depending on its distribution. As shown in Figure 1, performance was best in the two conditions (*C1* and *C1V3*) where the onset consonant of the first syllable in each word was lengthened. The worst performance was in condition *V1C2*, where a lengthened vowel was followed by a lengthened consonant within the same word.

Thus, even in the absence of other segmental and prosodic cues, listeners interpret lengthened consonants to indicate the start of a new word. This suggests that a modification is required to the iambic-trochaic law for spoken language to reflect the perceptual importance of the *locus* of prosodic lengthening effects [18, 19]: lengthened vowels cue a following boundary; lengthened consonants cue a preceding boundary.

The relative importance of timing compared to other boundary cues is not examined here. Natural speech typically provides multiple congruent sources of information about segmentation: when higher-level cues, such as lexicality and syntactic structure, offer an unambiguous guide to structure, acoustic-phonetic cues appear to be minimally exploited by listeners [15].

Our proposal for a functional division between vowels and consonants is congruent with claims that they carry distinct informational loads in speech processing, even for neonates [34, 35]. Language experience is probably required, however, before the development of differential sensitivity to localized durational effects in vowels and consonants. A similar preference for initial pitch-salience to that established in adults has been shown with 7-month-old infants, but no distinction between initial and final length-salience was found at the same age [12], indicating that more linguistic exposure is required before vowel lengthening is associated with a following boundary. The same probably applies for the interpretation of consonant lengthening as a cue to a preceding boundary. It remains to be seen whether this functional distinction holds in languages other than English.

6. Acknowledgements

This work was supported by a British Academy grant to the first author. We thank Elizabeth Gabe-Thomas, Laura König and Jean Roper for help with running experiments.

7. References

[1] Beckman, M. E. (1992). Evidence for speech rhythms across languages. In Y. Tohkura, E. Vatikiotis-Bateson & Y. Sagisaka (eds.), *Speech Perception, Production and Linguistic Structure* (pp. 457-463). Oxford: IOS Press.

[2] Ladd, D.R. (1996). *Intonational Phonology*. Cambridge: Cambridge University Press.

[3] Fowler, C.A. (1990). Lengthenings and the nature of prosodic constituency: Comments on Beckman and Edwards's paper. In J. Kingston & M.E. Beckman (Eds.), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*. (pp. 201-207). Cambridge: Cambridge University Press.

[4] Tyler, M. D., & Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *Journal of the Acoustical Society of America*, *126*, 367-376.

[5] Christophe, A., Peperkamp, S., Pallier, C., Block, E., & Mehler, J. (2004). Phonological phrase boundaries constrain lexical access I. Adult data. *Journal of Memory and Language*, *51*, 523-547.

[6] Gout, A., Christophe, A., & Morgan, J. L. (2004). Phonological phrase boundaries constrain lexical access II. Infant data. *Journal of Memory and Language*, *51*, 548-567.

[7] Price, P.J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America*, *90*, 2956-2970.

[8] Saffran, J.R., Newport, E.L., & Aslin, R.N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.
[9] Salverda, A.P., Dahan, D., & McQueen, J.M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, *90*, 51-89.
[10] Hayes, B. (1995). *Metrical Stress Theory: Principles and Case Studies*. Chicago: University of Chicago Press.
[11] Hay, J.S., & Diehl, R.L. (2007). Perception of rhythmic grouping: Testing the iambic/trochaic law. *Perception and Psychophysics*, *69*, 113-122.
[12] Bion, R. A., Benavides-Varela, S., & Nespor, M. (2011). Acoustic markers of prominence influence infants' and adults' segmentation of speech sequences. *Language and Speech*, 54, 123-140.
[13] Toro, J. M., Sebastian-Galles, N., & Mattys, S. L. (2009). The role of perceptual salience during the segmentation of connected speech. *European Journal of Cognitive Psychology*, *21*, 786-800.
[14] Cutler, A., & Carter, D.M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, *2*, 133-142.
[15] Mattys, S.L., White, L., & Melhorn, J.F. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General*, *134*, 477-500.
[16] Cutler, A., Wales, R., Cooper, N., & Janssen, J. (2007). Dutch listeners' use of suprasegmental cues to English stress. In *Proceedings of the XVIth International Congress of Phonetic Sciences* (pp. 1913-1916).
[17] Cutler, A., & Pasveer, D. (2006). Explaining cross-linguistic differences in effects of lexical stress on spoken-word recognition. *Proceedings of Speech Prosody 2006*, *Dresden* (pp. 237-240).
[18] White, L. (2002). *English Speech Timing: A Domain and Locus Approach*. University of Edinburgh PhD dissertation.
[19] White, L. (under revision). Communicative function and prosodic form in speech timing: Structure is signalled by localised lengthening effects
[20] Monaghan, P., White, L., & Merkx, M.M. (2013). Disambiguating durational cues for speech segmentation. *Journal of the Acoustical Society of America*, *134*, EL45-EL51.
[21] Oller, D.K. (1973). The effect of position in utterance on speech segment duration in English. *Journal of the Acoustical Society of America*, *54*, 1235-1247.
[22] Keating, P. A., Cho, T., Fougeron, C., & Hsu, C. (2003). Domain-initial strengthening in four languages. In J. Local, R. Ogden, & R. Temple (Eds.). *Papers in Laboratory Phonology 6* (pp. 145-163). Cambridge: Cambridge University Press.
[23] Fougeron, C. & Keating, P.A. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, *101*, 3728-3740.
[24] Byrd, D., Lee, S., Riggs, D., & Adams, J. (2005). Interacting effects of syllable and phrase position on consonant articulation. *Journal of the Acoustical Society of America*, *118*, 3860-3873.
[25] Gow, D.W. & Gordon, P.C. (1995). Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 344-359.
[26] Quené, H. (1992). Durational cues for word segmentation in Dutch. *Journal of Phonetics*, *20*, 331-350.
[27] Tagliapietra, L., & McQueen, J. M. (2010). What and where in speech recognition: Geminates and singletons in spoken Italian. *Journal of Memory and Language*, *63*, 306-323.
[28] Spinelli, E., McQueen, J. M., & Cutler, A. (2003). Processing resyllabified words in French. *Journal of Memory and Language*, *48*, 233-254.
[29] Cho, T., McQueen, J., & Cox, E. (2007). Prosodically driven detail in speech processing: the case of domain-initial strengthening in English. *Journal of Phonetics*, *35*, 210-243.
[30] Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month old infants. *Science*, 274, 1926-1928.
[31] Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & Van Der Vreken, O. (1996). The MBROLA project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. In *Proceedings of the International Conference on Spoken Language Processing, Philadelphia* (pp. 1393-1396).
[32] Baayen, R.H., Davidson, D.J., & Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390-412.
[33] Shukla, M., Nespor, M. & Mehler, J. (2007) Interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology*, *54*, 1-32.
[34] Benavides-Varela, S., Hochmann, J.R., Macagno, F., Nespor, M., & Mehler, J. (2012). Newborn's brain activity signals the origin of word memories. *Proceedings of the National Academy of Sciences*, *109*, 17908-17913.
[35] Bonatti, L. L., Pena, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations. The role of consonants and vowels in continuous speech processing. *Psychological Science*, *16*, 451-459.