INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

How f0 and phrase position affect Papuan Malay word identification

Constantijn Kaland[1] and Matthew Gordon[2]

[1]Institute of Linguistics, University of Cologne, Germany
[2]Department of Linguistics, University of California, Santa Barbara, USA
ckaland@uni-koeln.de, mgordon@linguistics.ucsb.edu

Abstract

This paper reports a perception experiment on Papuan Malay, an Eastern Indonesian language for which phrase prosody is largely underresearched. While phrase-final f0 movements are the most prominent ones in this language, it remains to be seen to what extent they signal phrase boundaries (demarcating) or whether they contribute to the prosodic prominence of words in that position (highlighting). Crucially, it is unclear whether these functions can actually be teased apart. In an attempt to investigate this issue, a word identification experiment was carried out using manipulated and original f0 word contours in phrase-medial and phrase-final positions. Results indicate that Papuan Malay listeners recognize words faster in phrase-final position, although the shape of the f0 movement did not significantly affect response latencies. The outcomes are discussed in a typological perspective, with particular attention to Trade Malay languages.

Index Terms: prosody, word recognition, f0, intonation, Papuan Malay.

1. Introduction

Intonation research has often distinguished two major types of f0 movements, depending on their linguistic function (e.g. highlighting and demarcating). For example, autosegmental-metrical approaches separate pitch accents from boundary tones (e.g. [1]) and prosody transcription methods have distinguished prominences from boundaries (e.g. [2]). These distinctions are based on the idea that f0 movements either contribute to the highlighting of particular words in a phrase (pitch accents, prominences) or to the marking of phrase edges (boundary tones, boundaries). While this divide has been maintained for many languages (e.g. [3],[4]), it is challenging in practice to tease apart the source of f0 movements as boundary- or prominence-linked, a difficulty that has sparked debates about the analysis of prominence in many languages ([5],[6]). While taxonomy decisions have been largely based on production data, perception experiments potentially also provide insight into the source and function of f0 excursions. More generally, psycholinguistic studies can inform our understanding of the role of f0 and other boundary-associated properties as perceptual landmarks (e.g. [7]).

One group of language varieties beset by analytic indeterminacy surrounding the source of f0 movements is the prosodically underresearched family of (Trade) Malay variants. The only empirical study available for one variety, Ambonese Malay [8], concluded that this language does not make use of pitch accents and does not mark focus prosodically. The commonly observed rise-fall f0 movement in phrase-final position was analysed as a boundary tone with a loose alignment to the segments; i.e. somewhere around the boundary between the final two syllables in the phrase. Although Manado Malay is closely related to Ambonese Malay, it was analysed as a language that marks subject-, object- or verb-focus distinctions using f0 [9]. It should be noted that no acoustic analysis was carried out to confirm this conclusion.

Papuan Malay, another Trade Malay variety, is spoken by over 1 million people in the West-Papua and Papua provinces of Indonesian [10]. It has been more thoroughly researched, but questions still remain about the functions of its phrase prosody. Native listeners who annotated prominences and boundaries in short phrases showed agreement in identifying the position of the latter rather than the former [11], suggesting that boundary demarcation rather than prominence marking is the core prosodic function of f0 (cf. Ambonese Malay). Later work showed that this conclusion required more nuance. The same annotation task carried out with German and Papuan Malay speech materials, both presented to native listeners of each language, showed that Papuan Malay listeners reached higher agreement identifying prominences for German than for their native language [12]. These results were interpreted as an indication that Papuan Malay listeners are sensitive to prominence (i.e. non-boundary) marking, although their language makes little use of it. These results are partially supported by work on word prosody in Papuan Malay showing acoustic, perceptual and lexical support for regular penultimate word stress ([13],[14],[15]). It remains, however, unclear to what extent word stress interacts with phrase prosodic events [5]. F0 movements on the final two syllables in Papuan Malay phrases were largest and often showed a rise on the stressed syllable [16].

While the studies on Trade Malay languages all indicate that the f0 movements in phrase-final position are the most prominent, their function is unclear. While Ambonese Malay only marks boundaries and Manado Malay allows highlighting, the work on Papuan Malay does not suggest a strong distinction between a highlighting and demarcating function. Word recognition studies provide one way to further investigate this issue. It is known, for example, that pitch accents speed up the recognition of words ([17],[18]). Research has also shown that f0 facilitates word recognition when its pattern matches the location of the stressed syllable [19]. For Papuan Malay, such a facilitation effect was indeed found, however, the effect was more consistent when the word was presented in a phrase than when presented in isolation [14], indicating that (prosodic) context plays a crucial role. This is confirmed by cross-linguistic research on the facilitation effect of rhythmical expectations and f0 contour in phrases, regardless of the target word's acoustic realisation (e.g. [20],[21]). Furthermore, phrase-final position appears to facilitate word recognition to a larger extent than phrase-

http://dx.doi.org/10.21437/Interspeech.2021-6

medial position [22]. The latter effect has been explained as the result of context; listeners have heard more semantic context and thus become better in predicting upcoming words towards the end of a phrase. In the case of Papuan Malay, it is thus plausible that highlighting is restricted to phrase-final positions, as in Manado Malay. This would support the idea that Papuan Malay makes use of pitch accents. However, if prominent f0 movements always occur in phrase-final positions, listeners also have a reliable cue to detect phrase boundaries in addition to any other boundary cues that may be present. Note that f0 could be used for both highlighting and demarcating, perhaps even synergistically, which challenges the traditional distinction between these two functions of prosody. The question therefore remains how the specific shape of the f0 movement and phrase position affect word recognition in Papuan Malay. Answering this question will shed light on whether phrase-final f0 movements should be interpreted as pitch accents due to their shape (and plausibly alignment with the stressed syllable), as boundary tones due to their position, or as an amalgamate where both f0 and phrase position affect word recognition.

Thus, if phrase-final f0 movements are indeed pitch accents with a particular (rise-fall) shape, word recognition is predicted to be slower when this movement is absent. And if phrase-final position constitutes a privileged location in the prosodic structure, words should be recognized faster phrase-finally than phrase-medially. Both factors are tested in a word identification task outlined in the next section.

2. Method

A reaction time (RT) experiment was set up to investigate native listeners' word identification latencies in Papuan Malay phrases. The target words appeared in either phrase-medial or phrase-final position and either had an original f0 contour or a manipulated (flat) f0 contour.

2.1. Recordings

For the current experiment recordings from [23] were used, consisting of Papuan Malay words embedded in a matrix clause, read by a male native speaker. The matrix clause was constructed in such a way that the target word appeared either in phrase-medial (1a) or in phrase-final position (1b).

(1a)	ko	pu	kata	__	itu,	sa	blum	taw
	2SG	POSS	word	__	D.DIST	1SG	not.yet	know

'that word __ of yours, I don't yet know (it)'

(1b)	sa	blum	taw	ko	pu	kata	itu,	kata	__
	1SG	not.yet	know	2SG	POSS	word	D.DIST	word	__

'I don't yet know that word of yours, the word __'

From the recordings a subset was selected for use in the current experiment. Because Papuan Malay makes use of a considerable number of loanwords, only native Papuan Malay roots were selected. Furthermore, several recordings that were unclear due to the low intensity of the speaker's voice were not used in the current study. The selected set consisted of 20 recordings with phrase-medial target words and 20 recordings with phrase-final target words.

2.2. Design

In the experiment, participants listened to the matrix sentences. Their task was to indicate as fast as possible which target word they heard in the sentence. One stimulus consisted of the entire matrix sentence including the target word in either phrase-medial or in phrase-final position. For each stimulus, participants could choose between two response words, of which only one matched the target. The incorrect response word (distractor) was chosen such that it partially matched the target word. That is, the most frequently occurring syllable structure and stress pattern in Papuan Malay is ˈCV.CV [10]. The second syllable of the distractor was identical to the second syllable in the target. This was done to guarantee that the crucial cue to identify the target word was the first (stressed) syllable. Specifically, the distractor was chosen such that the difference with respect to the target was the vowel in the first syllable. This was done to make sure that the most sonorous part of the stressed syllable would always contribute to the identification of the target word. For some distractor words the consonant in the first syllable was also different from the target (due to the limited number of suitable words). For example, when the stimulus was "sa blum taw ko pu kata itu, kata *laki*" (I don't yet know that word of yours, the word *husband*"), the distractor was *hoki* (*plant stem*).

2.3. F0 manipulation

From each of the selected recordings (Section 2.1) two versions were created in which f0 was manipulated using Time-Domain Pitch-Synchronous Overlap-and-Add (TD-PSOLA; [24]) as implemented in Praat [25]. It has been shown that the naturalness of speech due to TD-PSOLA resynthesis is somewhat decreased compared to unmanipulated speech [26]. Therefore, in one version, the f0 contour was only stylized using a frequency resolution of 2 semitones in Praat [25]. The stylised f0 contour closely followed the one in the original recording. In this way, the recording would undergo TD-PSOLA resynthesis without a change in the trajectory of the f0 contour. This was done to decrease the naturalness of the stimuli to a level comparable to the other version, which underwent TD-PSOLA resynthesis for the purpose of f0 manipulation. The rationale behind this procedure is that potential side-effects on participants' response latencies due to naturalness were balanced.

In the other version, the stylised f0 contour and the contour of the target word were flattened such that no pitch excursions occurred within the target word. This was done differently for phrase-medial target words than for phrase-final target words. For phrase-medial words, the original f0 level at the start of the first syllable was maintained throughout the entire word. The rise that originally occurred on the first syllable of the target word was then shifted onto the first syllable of the next word in the matrix sentence (*i* in *itu*), see Figure 1 bottom left. In this way, the f0 in the remaining part of the matrix sentence would follow its original contour. As for phrase-final words, the f0 was kept at the level at the start of the first syllable with a declination towards the end of the word (equivalent to the phrase end). The declination was determined by taking the original f0 end point in the phrase. This was done to minimize the impression that the contour was manipulated.

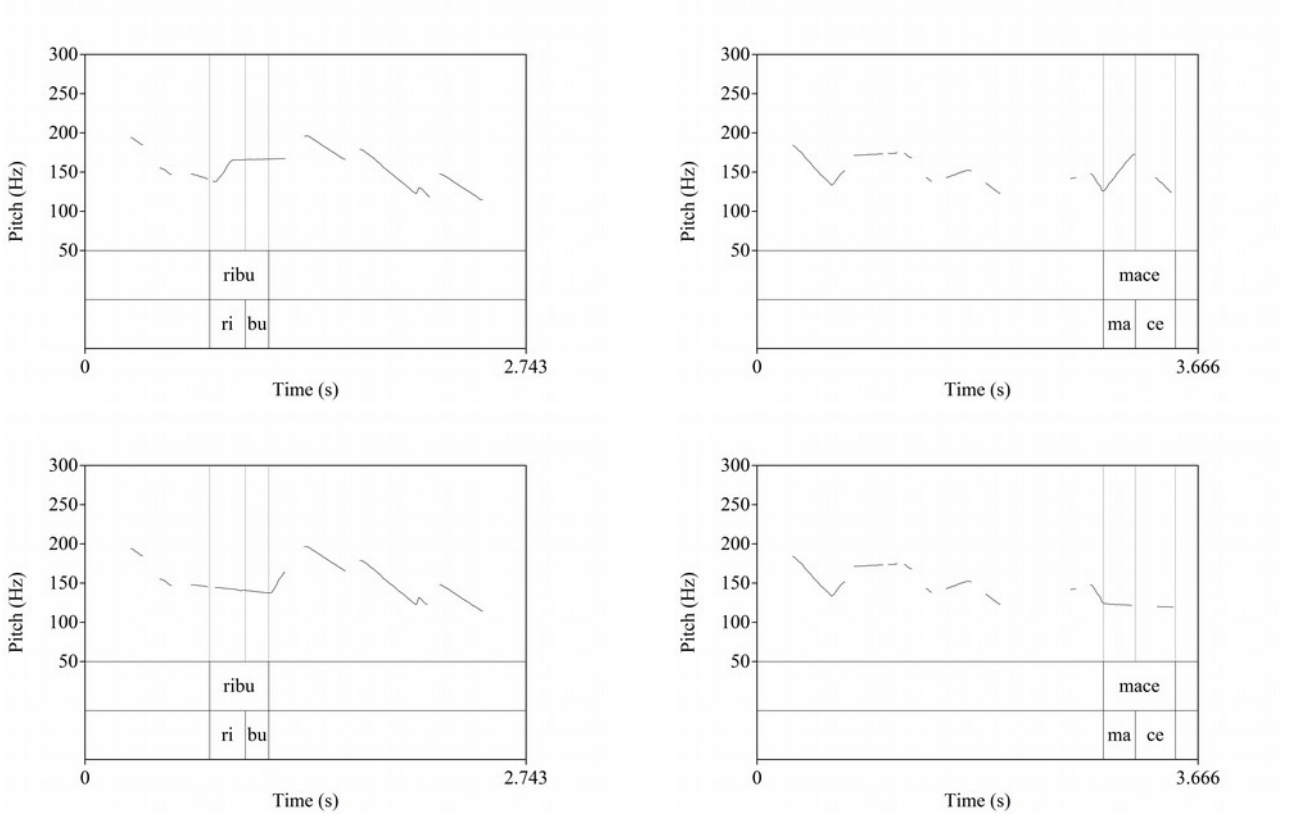

Figure 1: *Stylized f0 contours on target words following the original (top) and manipulated (bottom) contour in phrase-medial (left) and phrase-final (right) positions. Target words are segmented on the word level (top tier) and syllable level (bottom tier).*

2.4. Participants

In total, 20 native speakers of Papuan Malay without hearing problems completed the experiment for course credit (13 females, 7 males; mean age: 21.2; age range 18-41).

2.5. Procedure

The word identification task was designed using OpenSesame [27]. The experiment consisted of a Python [28] script and 80 wave files (stimuli). For each stimulus, the script generated a screen. The screen displayed the question "Kata mana yang Anda dengar?" (Which word did you hear) and two buttons. On each of the buttons was shown the corresponding key that should be pressed to choose one of the response words (either 1 for the word on the left, or 0 for the word on the right). The response words were written underneath the respective buttons. Target and distractor were randomly assigned as left or right word on the screen, differently for each participant. The stimulus screen was displayed for five seconds in order to let participants familiarize themselves with the two response options and to prepare them to hear the stimulus. During the last three seconds the participant heard three successive tones of 1 kHz that cued the upcoming stimulus. The first two were 250 ms in length and the last one lasted for the entire final second before the stimulus was played. The stimulus screen was displayed until participants had pressed "1" or "0". After making their choice participants needed to press the space bar to initiate the next stimulus. The space bar was chosen so that participants could keep their hands on the keyboard during the entire experiment. The stimulus familiarisation time was fixed (five seconds) to make sure all participants underwent the same procedure. Note that the time between successive stimuli was participant-initiated to allow participants to set the pace of the experiment, which has been shown to lead to lower rates of missed responses and to improve participants' compliance [29]. This aspect is crucial for the current study's participants, who had little to no familiarity with RT experiments. RTs were measured between the offset of the target word in the stimulus sentence and the moment the participant had pressed "1" or "0". Commonly, stimulus onset latencies are reported in word recognition tasks, although stimulus offset latencies better account for differences in stimulus duration [30], which were present in the stimuli of the current experiment. Phrase positions were balanced across the parts of the experiment before and after the break. That is, one half of the participants was presented with phrase-medial targets in the first part and phrase-final targets in the second part. The other half was presented with phrase-final targets in the first part and phrase-medial targets in the second part. The presentation order of the stimuli was random and different for each participant. This was done to balance potential effects of handedness (faster with preferred hand), as well as other side-effects potentially associated with a fixed order (e.g. learning effects).

Before the start of the experiment participants received verbal instructions about the course of the tasks. They were instructed to press the corresponding button on their keyboard as quickly as possible when they heard one of the words displayed on the screen. Then, they took a seat behind a computer and completed two subsequent parts of the experiment. First, they received written instructions on the

screen about their task. To familiarize themselves with the task, participants completed a practice round consisting of five stimuli. At the end of the practice round participants were asked whether they felt they needed to practice more or whether they were ready to start the actual task. When more practice was needed, participants were presented additional stimuli. After each additional practice stimulus, participants could end the practice round. Second, when participants ended the practice session, they were asked to start the actual word identification task. After completing 50% of the actual identification task, participants were instructed to take a short break. The experiment lasted approximately 25 minutes. Results were collected on a computer. RTs shorter than 200 ms after target onset (N = 3) and RTs longer than two seconds after target offset were discarded (N = 1) as these were plausibly the result of erroneous responses.

2.6. Statistical analysis

Statistical analyses were carried out using R [31] and the lme4 package [32]. Linear mixed model analyses (LMM) fit by maximum likelihood (using Satterthwaite approximations to degrees of freedom to calculate p-values) were carried out on the offset RTs with f0 manipulation (2 levels: original, manipulated) and phrase position (2 levels: medial, final) as predictors. Participant and final word in the phrase were included as random intercepts with f0 manipulation as by-participant random slope (the maximal converging model).

3. Results

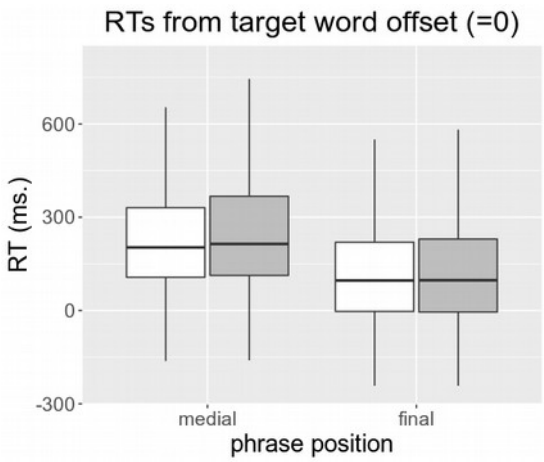

Figure 2: *Boxplots of the RTs to correct target identification in the respective phrase positions and f0 manipulations (white = original, grey = manipulated).*

Table 1: *Mean RTs (standard deviations) to correct target word identification*

phrase position	f0 manipulation	RT (ms)
medial	original	280.79 (*281.72*)
	manipulated	298.97 (*285.62*)
final	original	129.92 (*221.22*)
	manipulated	146.24 (*234.41*)

Figure 2 and Table 1 show that participants were faster identifying the target word when its f0 was original and when its position was phrase-final. Results of the LMM showed that f0 manipulation did not have a significant effect (β = 15.57, SE = 16.49, t = 0.95, *n.s.*), whereas phrase position did have a significant effect (β = 151.55, SE = 25.55, t = 5.93, p < 0.001). The interaction did not show a significant effect.

4. Conclusions

The results of the RT experiment show that only the phrase position, not the f0 shape, has a statistically significant effect on the word identification latencies of Papuan Malay listeners. This outcome indicates that the phrase-final position constitutes a privileged location in the prosodic structure. It can be ruled out that the effect of phrase position in this study was the result of semantic context. The matrix phrase did not provide clues on the basis of which participants could have predicted target words with more accuracy in phrase-final position than in phrase-medial position. For this reason, it can be predicted that in natural Papuan Malay phrases, which do provide a more useful semantic context for predicting upcoming words, the effect of phrase position could be even stronger. Although the manipulated (flat) f0 contour somewhat slowed down participants' identification times, the shape of the f0 movements is not as crucial as the phrase position. Together, these results favor the interpretation that Papuan Malay does not make use of specific pitch accents. Rather, the main prosodic phenomenon appears to be boundary-related, although, given the fact that tokens with flat terminal f0 contours also facilitated recognition in phrase-final position, it is less clear which particular boundary properties are most important from a processing standpoint. It is plausible that prosodic phenomena such as final lengthening contribute to the acoustic clarity (or prominence) of phrase-final words. In this way, Papuan Malay may be a language that exploits phrase-final positions for both highlighting and demarcating purposes, without strong demands on f0 shape. Concerning demarcation, more perceptual research needs to be done in order to investigate listeners' ability to detect phrase boundaries when the acoustic properties of phrase-final syllables are manipulated. Furthermore, it is plausible that word recognition is expedited by rhythmic expectations ([20], [21]), such as regular penultimate word stress in Papuan Malay [13]. Determining the role of experiential rhythmic priming in speech processing, including f0 as well as other acoustic cues, could shed crucial light on the potential interplay of word stress and phrase prosody. Research has only begun to scratch the surface in this respect [16] and could fruitfully be extended to other underresearched languages [7]. This small study already demonstrated the importance of broadening typological knowledge of the role of prosody in speech processing and its results challenge the usefulness of traditionally maintained categories of prosodic functions.

5. Acknowledgements

Research for this paper was funded by the German Research Foundation (DFG,) – Project-ID 281511265 – SFB 1252. The authors thank the staff of the Center for Endangered Languages Documentation (CELD, Manokwari, West-Papua) for facilitating the experiment and for participant recruitment, and three anonymous reviewers for valuable comments.

6. References

[1] J. Pierrehumbert and J. Hirschberg, "The Meaning of Intonational Contours in the Interpretation of Discourse," 1990, doi: 10.7916/d8kd24fp.

[2] J. Cole and S. Shattuck-Hufnagel, "New Methods for Prosodic Transcription: Capturing Variability as a Source of Information," Laboratory Phonology, vol. 7, no. 1, pp. 1–29, Jun. 2016, doi: 10.5334/labphon.29.

[3] S.-A. Jun, Ed., Prosodic typology: the phonology of intonation and phrasing. Oxford; New York: Oxford University Press, 2005.

[4] S.-A. Jun, Ed., Prosodic typology II: the phonology of intonation and phrasing. Oxford: Oxford University Press, 2014.

[5] M. Gordon, "Disentangling stress and pitch-accent: a typology of prominence at different prosodic levels," in Word Stress, H. van der Hulst, Ed. Cambridge, UK: Cambridge University Press, 2014, pp. 83–118.

[6] M. Gordon, "Word stress and intonational prominence in highly synthetic languages," in Prominence in Languages with Complex Morphology, H. van der Hulst and K. Bogomolets, Ed. Oxford, UK: Oxford University Press, to appear.

[7] A. Cutler, Native listening: language experience and the recognition of spoken words. Cambridge, UK: MIT Press, 2012.

[8] R. Maskikit-Essed and C. Gussenhoven, "No stress, no pitch accent, no prosodic focus: the case of Ambonese Malay," Phonology, vol. 33, no. 2, pp. 353–389, Aug. 2016, doi: 10.1017/S0952675716000154.

[9] R. B. Stoel, "The intonation of Manado Malay," in Prosody in Indonesian Languages, vol. 9, van Heuven, Vincent J. and van Zanten, E, Eds. Utrecht: LOT, Netherlands Graduate School of Linguistics, 2007, pp. 117–150.

[10] A. Kluge, A grammar of Papuan Malay. Berlin, Germany: Language Science Press, 2017.

[11] S. Riesberg, J. Kalbertodt, S. Baumann, and N. P. Himmelmann, "On The Perception Of Prosodic Prominences And Boundaries In Papuan Malay," in Perspectives on information structure in Austronesian languages, S. Riesberg, A. Shiohara, and A. Utsumi, Eds. Berlin, Germany: Language Science Press, 2018, pp. 389–414.

[12] S. Riesberg, J. Kalbertodt, S. Baumann, and N. P. Himmelmann, "Using Rapid Prosody Transcription to probe little-known prosodic systems: The case of Papuan Malay," Laboratory Phonology: Journal of the Association for Laboratory Phonology, vol. 11, no. 1, p. 8, Jul. 2020, doi: 10.5334/labphon.192.

[13] C. C. L. Kaland, "Acoustic correlates of word stress in Papuan Malay," Journal of Phonetics, vol. 74, pp. 55–74, May 2019, doi: 10.1016/j.wocn.2019.02.003.

[14] C. C. L. Kaland, "Offline and online processing of acoustic cues to word stress in Papuan Malay," The Journal of the Acoustical Society of America, vol. 147, no. 2, pp. 731–747, Feb. 2020, doi: 10.1121/10.0000578.

[15] C. C. L. Kaland and V. J. van Heuven, "Papuan Malay word stress reduces lexical alternatives," in Proc. 10th International Conference on Speech Prosody 2020, 2020, pp. 454–458, doi: 10.21437/SpeechProsody.2020-93.

[16] C. C. L. Kaland and S. Baumann, "Demarcating and highlighting in Papuan Malay phrase prosody," The Journal of the Acoustical Society of America, vol. 147, no. 4, pp. 2974–2988, Apr. 2020, doi: 10.1121/10.0001008.

[17] A. Cutler and D. J. Foss, "On the Role of Sentence Stress in Sentence Processing," Lang Speech, vol. 20, no. 1, pp. 1–10, Jan. 1977, doi: 10.1177/002383097702000101.

[18] B. Braun, A. Dainora, and M. Ernestus, "An unfamiliar intonation contour slows down online speech comprehension," Language and Cognitive Processes, vol. 26, no. 3, pp. 350–375, Apr. 2011, doi: 10.1080/01690965.2010.492641.

[19] C. K. Friedrich, S. A. Kotz, A. D. Friederici, and K. Alter, "Pitch modulates lexical identification in spoken word recognition: ERP and behavioral evidence," *Cognitive Brain Research*, vol. 20, no. 2, pp. 300–308, Jul. 2004, doi: 10.1016/j.cogbrainres.2004.03.007.

[20] A. Cutler, "Phoneme-monitoring reaction time as a function of preceding intonation contour," Perception & Psychophysics, vol. 20, no. 1, pp. 55–60, Jan. 1976, doi: 10.3758/BF03198706.

[21] J. L. Shields, A. McHugh, and J. G. Martin, "Reaction time to phoneme targets as a function of rhythmic cues in continuous speech.," Journal of Experimental Psychology, vol. 102, no. 2, pp. 250–255, 1974, doi: 10.1037/h0035855.

[22] D. J. Foss, "Decision processes during sentence comprehension: Effects of lexical item difficulty and position upon decision times," Journal of Verbal Learning and Verbal Behavior, vol. 8, no. 4, pp. 457–462, Aug. 1969, doi: 10.1016/S0022-5371(69)80089-7.

[23] A. Kluge, B. A. W. Rumaropen, and L. Aweta, "Papuan Malay data - Word list," SIL International, Dallas, TX, 2014. https://www.sil.org/resources/archives/59649.

[24] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Communication, vol. 9, no. 5, pp. 453–467, Dec. 1990, doi: 10.1016/0167-6393(90)90021-Z.

[25] Boersma, Paul and Weenink, David, Praat: doing Phonetics by Computer. 2019.

[26] H. Mixdorff and D. Mehnert, "Exploring the naturalness of several German high-quality-text-to-speech systems," 1999.

[27] S. Mathôt, D. Schreij, and J. Theeuwes, "OpenSesame: An open-source, graphical experiment builder for the social sciences," Behav Res, vol. 44, no. 2, pp. 314–324, Jun. 2012, doi: 10.3758/s13428-011-0168-7.

[28] G. Van Rossum and J. De Boer, "Interactively testing remote servers using the Python programming language," CWI Quarterly, vol. 4, no. 4, pp. 283–304, Dec. 1991.

[29] H. Krinzinger et al., "Sensitivity, Reproducibility, and Reliability of Self-Paced Versus Fixed Stimulus Presentation in an fMRI Study on Exact, Non-Symbolic Arithmetic in Typically Developing Children Aged Between 6 and 12 Years," Developmental Neuropsychology, vol. 36, no. 6, pp. 721–740, Aug. 2011, doi: 10.1080/87565641.2010.549882.

[30] J. Lipinski and P. Gupta, "Does neighborhood density influence repetition latency for nonwords? Separating the effects of density and duration," Journal of Memory and Language, vol. 52, no. 2, pp. 171–192, Feb. 2005, doi: 10.1016/j.jml.2004.10.004.

[31] R Core Team, R: The R Project for Statistical Computing. 2019.

[32] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," Journal of Statistical Software, vol. 67, no. 1, pp. 1–48, Oct. 2015, doi: 10.18637/jss.v067.i01.