9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Neural network modeling of prosodic prominence in Besemah (Malayic, Indonesia)

Bradley McDonnell, Rory Turnbull

University of Hawai'i at Mānoa, USA

mcdonn@hawaii.edu, rory.turnbull@hawaii.edu

Abstract

A number of recent studies have proposed that various languages in western Indonesia do not show evidence of word-level stress, and they only exhibit evidence for sentence-level prominence [1, 2]. This study examines the acoustic realization of prosodic prominence within different domains in Besemah, a little-described Malayic language of southwest Sumatra, Indonesia.

The present study reports the results of a production experiment in which six female native speakers of Besemah completed an information gap task where target words were uttered in different frames that varied along two dimensions: information status and position within the sentence. Based on the results of a neural network analysis that used acoustic features to predict syllable position in the word, information status, and sentence position, this study shows that information status cannot be predicted above chance, but that both position of the syllable in the word and the position within in sentence can be predicted with above chance levels of accuracy. These patterns are consistent with the hypothesis that Besemah has predictable word-level stress, sentence-level prosodic boundary marking, and does not use prosodic means to mark focus.

Index Terms: Austronesian, word stress, pitch accent, information status, neural networks

1. Introduction

There has been much disagreement over the status of word-level stress in the languages of western Indonesia, particularly in regards to well-known varieties of Malay, such as Standard Indonesian [3]. Word-level stress has been claimed to fall predictably on the penultimate syllable unless it contains a schwa in which case it falls on the ultima [4]. However, since the late 1990s, an increasing number of studies have questioned this position for Standard Indonesian [5], and more recent studies have pointed out the complications associated with studies of Standard Indonesian due to significant influence from substrate languages like Javanese [1].

At the same time, there has been an increasing number of studies of prosodic prominence in other languages of western Indonesia that do not have the same complications associated with substrate influences. These studies also question the presence of word-level stress and propose that prosodic prominence is a property of the sentence and not the word. For example, based on a production experiment, Betawi Malay has been shown to lack any evidence for word-level stress, and pitch accents do not predictably anchor to a given syllable within a word [6]. Furthermore, again based on production experiments, Ambon Malay has been shown to lack any evidence for word-level stress, prosodically-marked focus, or pitch accents [2]. It remains an open question how widespread these features of prosodic prominence are in the languages of western Indonesia.

1.1. Word stress in Besemah

Besemah /bəsəmah/ is an under-described Malayic language spoken in the highlands of southwest Sumatra by approximately 400,000 people [7]. It is considered a 'vernacular Malay' meaning it is primarily spoken as a first language and was regularly inherited from Proto-Malayic. This status distinguishes it from Standard Indonesian, which is primarily spoken as a second language, and Ambon Malay, which is a 'Pidgin-derived Malay' that was not regularly inherited from Proto-Malayic [8].

Like many of the languages of western Indonesia, Besemah has received little attention since the Dutch colonial period [9]. The only study of Besemah prosody tentatively concluded that word-level stress in Besemah falls on the final syllable of the word and is cued by increased intensity [10]. However, there are several complications to this analysis, and word-level stress appears to be affected by the presence of final L boundary tones.

The present study is designed to tease apart different factors relating to prosodic prominence in Besemah. It does so by investigating the realization of common acoustic correlates of stress (f_0, duration, intensity, and spectral tilt) within target words that vary in their sentence position (sentence-medial vs. sentence-final position) and information status ('in focus' vs. 'out of focus').

2. Methods

2.1. Design

The experiment was designed to elicit target words that vary along two dimensions, sentence position and information status. This part of the design is similar to the study of Betawi Malay [6].

Target words were collected using an information gap task. This task involved two talkers: a confederate who asked questions and a naïve participant who answered them. They sat facing each other, each able to see only their own laptop screen. Both the confederate and participant could see a question on the top of the screen that the confederate was to ask, but only the participant's screen displayed the answer. The confederate also had a sheet of paper with all possible answers. The task was for the confederate to ask the question and the participant to provide an answer in a complete sentence modeled upon the question. The confederate then circled the answer on the paper. One female native speaker of Besemah acted as the confederate for all participants.

2.2. Materials

Question-answer pairs were constructed which varied in information status (target word 'in focus' or 'out of focus') and sentence position (target word sentence-medial or sentence-final). The type and structure of the question depended upon the combination of its sentence position and information status. Ques-

10.21437/SpeechProsody.2018-154

Table 1: *Example question-answer pairs for 'in focus' condition.*

Position		Question/Answer
final	Q:	*Sebelah kanan kate ape?*
		'On the right side is what word?'
	A:	*Sebelah kanan kate **susu**.*
		'On the right side is **milk**.'
medial	Q:	*Kate ape sebelah kanan?*
		'What word is on the right side?'
	A:	*Kate **susu** sebelah kanan.*
		'The word **milk** is on the right side.'

tion type differed based upon whether the target word was 'in focus' or 'out of focus'.[1] The sentence position simply altered the shape of the question so that the target word would occur in sentence-medial or sentence-final position. Table 1 presents example question-answer pairs for the 'in focus' condition. In this table the target word *susu* 'milk' is used as an example and displayed in boldface.

For stimuli that seek to collect target words that are 'out of focus', the confederate asked a different set of questions. Examples are shown in Table 2. Unlike the 'in focus' condition, the same target word was mentioned four times in immediately preceding turns. This mentioning of the target word served to keep the word as discourse-given and thus 'out of focus'. Further, each time the word appeared for the first time in a block it was introduced by asking the question *Titu kate ape?* 'What word is this?'. The target word in this question is 'in focus' in this utterance and not considered in the analysis. However, this allowed the confederate to ask questions about the target word without it being 'in focus'. Thus, the subsequent repetitions of the target word ask where the target word appears on the screen, the top, bottom, left or right side. In these four questions, the target word is 'out of focus' while the word describing its position is 'in focus'. Table 2 presents example question-answer pairs for the 'out of focus' condition.

Table 2: *Example question-answer pairs for 'out of focus' condition.*

Position		Question/Answer
final	Q:	*Sebelah mane kate **susu**?*
		'Which side is the word **milk** on?'
	A:	*Sebelah bawah kate **susu**.*
		'On the left side is the word **milk**.'
medial	Q:	*Kate **susu** sebelah mane?*
		'The word **milk** is on which side?'
	A:	*Kate **susu** sebelah kidau.*
		'The word **milk** is on the left side.'

The twelve target words are shown in Table 3. Syllables in Besemah are maximally CVC with some restrictions on the coda consonants and words are most commonly bisyllabic. The words feature all four phonotactically legal combinations of light (CV) and heavy (CVC) syllables. Each of the four vowels /i, u, a, ə/ was present and matched in both syllables to allow for easier intra-word comparisons. Phonotactically, the high vowels /i, u/ show no restrictions, but /a/ does not occur word-finally and /ə/ does not occur in final closed syllables.

[1] 'Focus', as a term of art in linguistics, is fraught with definitional difficulties. In our 'in focus' condition, the target word is the answer to the current question under discussion [11]. In the 'out of focus' condition, on the other hand, the target word is given, not new [12]. This definition may not match exactly with the definitions used by other researchers, and we invite the reader to mentally substitute their favorite terms that fit their proclivities if they are so inclined.

Table 3: *Target words used in constructing question-answer pairs.*

Vowel	Besemah	Gloss
/i/	/titi/	'to cross over'
	/pipis/	'to pulverize'
	/tʃiŋki/	'must have'
	/sintiŋ/	'crooked'
/u/	/susu/	'milk'
	/tutus/	'to pound'
	/tuŋku/	'hearth'
	/tuntun/	'to watch'
/a/	—	
	/tatap/	'to touch'
	—	
	/pantas/	'to be fitting'
/ə/	/tʃətə/	'to be exact'
	—	
	/təmpə/	'to forge (metal)'
	—	

2.3. Procedure

There were four experimental blocks that combined the possible combinations of information status and sentence position:

- Block A: sentence-medial, 'in focus'
- Block B: sentence-final, 'in focus'
- Block C: sentence-medial, 'out of focus'
- Block D: sentence-final, 'in focus'

All recordings were made on a Marantz PMD-670 solid state recorder with a sampling frequency of 48kHz and stored as wav files. The participant was recorded on the left channel with a Shure WH-30 headset microphone while the confederate was recorded on the right channel with a Shure SM10A headset microphone.

Because Besemah does not have a standard orthography, participants were first asked to read example sentences with each of the 12 target words in context. This ensures that the speakers knew how to read each of the target words and understood their meanings. Participants were then introduced to the information gap task and asked to practice each of the four blocks with eight filler words for each condition. Participants then began the task, which took 20 to 30 minutes.

Each target word was repeated four times for each of the four blocks (12 words $\times$ 4 repetitions $\times$ 4 blocks = 192 tokens). The presentation of these words was randomized and the presentation of each block was also randomized.

2.3.1. Participants

The experiment was conducted by the first author over several days in March 2015 in the Besemah village of Karang Tanding. The experiment was conducted entirely in Besemah with no interlanguage. The six female participants in this study are native

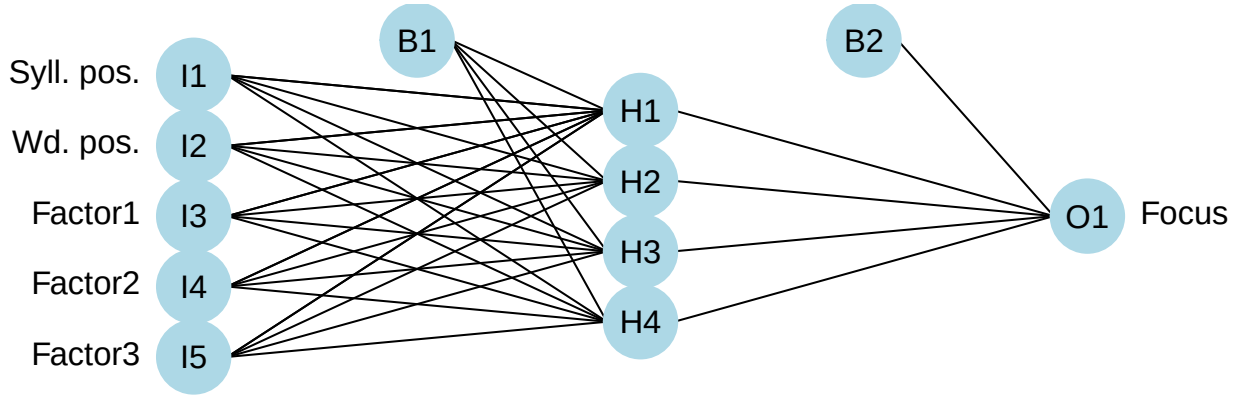

Figure 1: *Example neural network with four hidden nodes. As explained in the text, models were fit to predict three distinct output variables, and number of hidden nodes varied from 1 to 23.*

speakers of Besemah and reside in or near the village of Karang Tanding in South Sumatra province. They range in age from 19 to 30 years old.

2.4. Acoustic analysis

For each vowel of each target word, measures of f_0, duration, intensity, and spectral tilt were taken. The f_0 was measured at 7 equally-spaced timepoints throughout the vowel, that is, at 12.5%, 25%, 37.5%, 50%, 62.5%, 75%, and 87.5%. Intensity was estimated as dB averaged over the entire vowel. Two measures of spectral tilt were estimated: the difference in amplitude between the first and second harmonics ($H_1 - H_2$); and the difference in mean amplitude between the 0–4 kHz and the 4–8 kHz bins of the vowel spectrum ('spectral balance').

For analysis, the f_0 measures at the seven timepoints were fit to a cubic polynomial. Each word's four coefficients (intercept, linear term, quadratic term, cubic term) were the f_0 data used in the modeling reported here.

To avoid collinearities among the acoustic variables, and to simplify the subsequent analyses, the eight acoustic variables (duration, $H_1 - H_2$, spectral balance, intensity, and the four f_0 polynomial coefficients) were subjected to factor analysis to reduce the dimensionality of the data. The variables were z-scored prior to being entered into the model. The first three factors had eigenvalues greater than 1 and thus were selected. The factor loadings for the three factors are shown in Table 4.

As can be seen, the first factor is loaded most heavily on f_0 trajectory curvature, with a smaller contribution of duration. The second factor is largely $H_1 - H_2$, with duration, intensity, and f_0 intercept as secondary features. The third factor is not dominated by a single variable but is composed of intensity, spectral balance, and the higher-order f_0 terms.

2.5. Statistical analysis

Simple feed-forward neural networks with a single hidden layer were trained to predict the binary factors of sentence position, syllable position in word, and information status. The input to each neural network was the three factor scores output from the factor analysis, sentence position, syllable position, and information status. Figure 1 depicts an example neural network. Models were trained with 1,000 iterations of Monte Carlo cross-validation using a 90:10 training:test split.

Table 4: *Factor loadings greater than* $|0.1|$ *in acoustic factor analysis*

	Factor1	**Factor2**	**Factor3**
Duration	0.409	-0.185	
$H_1 - H_2$		0.993	
Intensity		-0.153	0.344
Spectral balance			-0.260
f_0 intercept	-0.377	0.406	
f_0 linear term	0.938		0.337
f_0 quadratic term	-0.985		0.145
f_0 cubic term	0.936		-0.324

This procedure was repeated with varying numbers of hidden nodes per model from 1 through 23. Thus, a total of 69,000 (1,000 iterations $\times$ 23 hidden node sizes $\times$ 3 dependent variables) networks were created and evaluated. The R package `nnet` [13] was used for all modeling.

3. Results

Figure 2 depicts mean neural network classification accuracy as a function of number of hidden nodes. The horizontal lines depict the limits of chance performance ($p < .05$) for the test data set. There is a clear distinction between predictive ability for the three dependent variables. Syllable position was predicted with just under 80% accuracy, sentence position with just over 70% accuracy, and focus accuracy wavered around 55% at chance level. For sentence position and focus, there was no clear effect of the number of hidden units; for syllable position, accuracy rose steeply from 1 to 2 hidden units, remained steady up to 5 hidden units, and then gently fell as the number of hidden units increased.

These results suggest that syllable position can be reliably predicted via the input variables. Indeed, the raw acoustic data suggest that there are reliable correlates of syllable position. Figures 3 and 4 depict vowel duration and midpoint f_0, respectively, for word-initial and word-final syllables. Final syllables are reliably longer in duration than initial syllables, and they are consistently higher in f_0 for all but one participant.

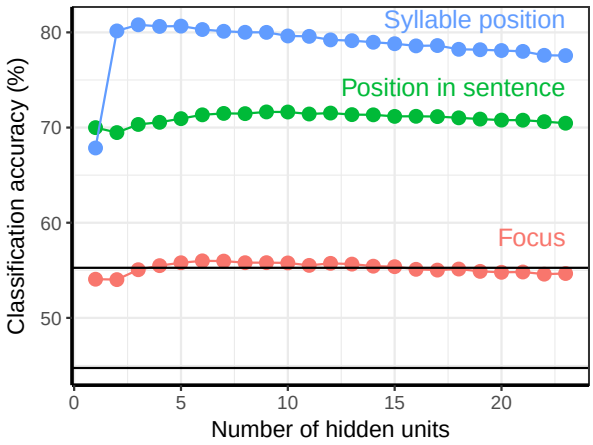

Figure 2: *Mean accuracy of neural networks as a function of number of hidden units. Solid horizontal lines depict boundaries for chance performance.*

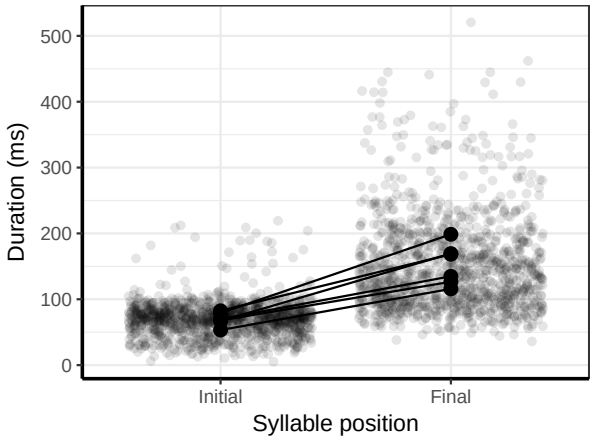

Figure 3: *Vowel duration in word-initial and word-final syllables. Solid points are participant means.*

4. Discussion and conclusion

In this study, an experiment was carried out to examine aspects of prosodic prominence in Besemah. Six native speakers participated in an information gap task to elicit spontaneous productions of bisyllabic target words that varied in sentence position (medial and final) and information status ('in focus' and 'out of focus'). Acoustic measures on the vowels of both syllables (initial and final) were taken.

Ensembles of neural networks succeeded at using the acoustic features to predict syllable position (initial or final) and word position in the sentence (medial or final). The syllable position finding suggests that, at least for the bisyllabic words included in this study, Besemah features a predictable stress system whereby the final syllable is reliably more prominent than the first. Similarly, words in sentence-final position in Besemah appear to undergo phrase-final lengthening, purportedly a universal across languages [14].

In contrast, focus was not reliably predicted beyond chance levels by the neural networks. This finding suggests that either Besemah does not mark focus with prosodic means (to the extent that the current experiment was able to capture the relevant dimensions of interest), or that there is variability in how the individual participants implemented prosodic focus. Both possibilities are attested in the literature—e.g. see K'iche' [15] for the first and Southern British English for the second [16]. We leave this question for further research.

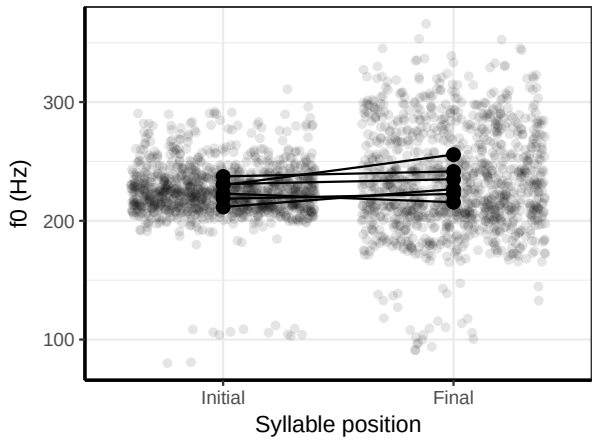

Figure 4: *Vowel midpoint f_0 in word-initial and word-final syllables. Solid points are participant means.*

The results of this study that concern information status and sentence position fall in line with the analysis of other Malayic languages [6, 2]. That is, information status is not predictably marked prosodically, but final boundaries are prosodically marked. The results of this study that concern word level stress provide further support for previous findings on Besemah [10]. That is, Besemah has word-level stress that falls on the final syllable of the word. This finding is indeed surprising given the results of studies in other Malayic languages that find no evidence for word-level stress (e.g., Betawi Malay [6] and Ambon Malay [2]). This suggests that Malayic languages may in fact differ in whether they possess word-level stress or not.

5. Acknowledgements

We would like to thank Katherine Strong, Carrie Honn, Ester Lee, and Hugh Richards for help segmenting vowels, and Matthew Gordon for advice on setting up the experiment. Besemah speakers from the village of Karang Tanding for participating in this study, and Kencana Dewi and Asfan Fikri Sanaf for making fieldwork on Besemah possible. Fulbright-Hays Doctoral Dissertation Fellowship for financial support, Yanti at the Center for Languages & Cultures at the Atma Jaya Catholic University of Indonesia for logistical support, and RISTEK for providing research permissions. All errors are our own.

6. References

[1] R. W. N. Goedemans and E. van Zanten, "Stress and accent in Indonesian," in *Prosody in Indonesian Languages*, V. J. van Heuven and E. van Zanten, Eds. Utrecht: LOT, 2007, pp. 35–62.

[2] 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, 2016.

[3] E. van Zanten, R. W. N. Goedemans, and J. J. Pacilly, "The status of word stress in Indonesian," in *The Phonological Spectrum: Suprasegmental Structure*, J. van de Weijer, V. J. van Heuven, and

H. van der Hulst, Eds. Amsterdam: John Benjamins, 2003, pp. 151–175.

[4] N. Adisasmito-Smith and A. C. Cohn, "Phonetic correlates of primary and secondary stress in Indonesian: A preliminary study," *Working Papers of the Cornell Phonetics Laboratory*, vol. 11, pp. 1–15, 1996.

[5] E. van Zanten and V. J. van Heuven, "Word stress in Indonesian: Fixed or free," *NUSA, Linguistic studies of Indonesian and other languages in Indonesia*, vol. 53, pp. 1–20, 2004.

[6] V. J. van Heuven, L. Roosman, and E. van Zanten, "Betawi Malay word prosody," *Lingua*, vol. 118, pp. 1271–1287, 2008.

[7] B. McDonnell, "Symmetrical voice constructions in Besemah: A usage-based approach," PhD dissertation, University of California, Santa Barbara, 2016.

[8] K. A. Adelaar, "Structural Diversity in the Malayic Subgroup," in *The Austronesian Languages of Asia and Madagascar*, K. A. Adelaar and N. P. Himmelmann, Eds. New York: Routledge, 2005, pp. 202–226.

[9] B. McDonnell, "A conservative vowel phoneme inventory of Sumatra: The case of Besemah," *Oceanic Linguistics*, vol. 47, no. 2, pp. 409–432, 2008.

[10] ——, "Acoustic correlates of stress in Besemah," in *NUSA: Linguistic studies of languages in and around Indonesia, Studies in Language Typology and Change*, T. McKinnon and Yanti, Eds., 2016, vol. 60, pp. 1–28.

[11] C. Roberts, "Information structure in discourse: Towards an integrated formal theory of pragmatics," *Semantics and Pragmatics*, vol. 5, no. 6, pp. 1–69, 2012.

[12] R. Schwarzschild, "GIVENness, AVOIDF and other constraints on the placement of accent," *Natural Language Semantics*, vol. 7, pp. 141–177, 1999.

[13] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed. New York: Springer, 2002, iSBN 0-387-95457-0. [Online]. Available: http://www.stats.ox.ac.uk/pub/MASS4

[14] S. Nakai, S. Kunnari, A. Turk, K. Suomi, and R. Ylitalo, "Utterance-final lengthening and quantity in Northern Finnish," *Journal of phonetics*, vol. 37, no. 1, pp. 29–45, 2009.

[15] R. S. Burdin, S. Phillips-Bourass, R. Turnbull, M. Yasavul, C. G. Clopper, and J. Tonhauser, "Variation in the prosody of focus in head-and head/edge-prominence languages," *Lingua*, vol. 165, pp. 254–276, 2015.

[16] S. Peppé, J. Maxim, and B. Wells, "Prosodic variation in southern British English," *Language and Speech*, vol. 43, no. 3, pp. 309–334, 2000.