Speech Prosody 2024
2-5 July 2024, Leiden, The Netherlands

Using role-playing tasks to document intonational tune prototypes in Nasal, an endangered language of Sumatra

Jacob Hakim

University of Hawaii at Mānoa
hakimj@hawaii.edu

Abstract

This paper describes a scripted role-playing task used to elicit a basic inventory of intonational tune types in Nasal (ISO 639-3: nsy; glottocode nasa1239), an endangered Austronesian language of Sumatra currently spoken in three villages by around 3,000 people. This study is a subset of the ongoing prosodic description that forms part of a larger Nasal documentation project. Prosodic description is not often included in language documentation, especially for Austronesian languages; when included, these descriptions are often very limited or based on impressionistic descriptions [1]. I make the case here that prosodic description based on carefully planned experimentation and acoustic measurement is not only achievable but a necessary part of linguistic fieldwork for language documentation. In the case of Nasal, eight participants (four men and four women) were recorded reading scripted lines from role-play dialogues in a variety of real-life scenarios. These recordings were transcribed, labeled according to the HCRC dialogue coding scheme [2], [3], and analyzed with a cluster analysis using the Contour Clustering app [4]. The alignment of tune patterns with different utterance types reveals a preliminary inventory of prototypical intonational tunes, including a falling tone pattern for polar questions, a pattern that is crosslinguistically uncommon.

Index Terms: intonation, Nasal, prosodic documentation, language documentation, Austronesian languages, Indonesian languages

1. Introduction

Much linguistic research in recent decades has been dedicated toward the goal of documentation and description of minoritized or underdescribed languages (see [5] for an overview). A central focus of language documentation is the creation of a maximally useful record of language use that is made available to any interested parties [6]. This focus has led to a trend in language work that is increasingly common in which not only linguistic research data but descriptions of data management have been made available [7]. This trend has resulted in increased access to primary data, often in the form of audio or video recordings, of languages from across the globe via open-access archives. This in turn has led to rich new research in language typology, with typological databases (e.g. StressTyp, described in [8]) providing quick access to linguistic data from a large number of languages.

At the same time, prosodic typology is still a budding field with many unanswered questions, and a vast majority of prosodic research, especially studies founded on acoustic measurements rather than impression, has been conducted only on widely spoken, well-described languages. For example, [9] provides a foundational theory of prosodic typology, but though a few lesser-studied languages are included, the majority included are well-studied languages. This is especially true for Austronesian languages [1]. The limited availability of data from endangered or understudied languages hinders the strength of typological studies, which aim to draw conclusions about the possible structures and processes of prosody in human speech.

In response to this need there has been a steadily growing body of work that focuses on describing the prosody of lesser-known languages (e.g., [10] for three Dene languages, or [11] for Papuan Malay, an Austronesian language). This study follows this line of research, drawing on newer approaches in cluster analysis of pitch tracks to investigate aspects of both word- and utterance-level prosody. [4] demonstrates that a cluster analysis has the potential to reduce researcher bias, an issue in impressionistic description and prosodic annotation (especially for non-native speakers), through its bottom-up method of analysis based on the acoustic signal. It also provides a means to increase both the efficiency and reproducibility of prosodic research. Here, a cluster analysis is applied to recordings of a scripted role-playing task; scripted dialogues allow researchers to explore and identify intonation patterns in a controlled way, and provide a useful starting point for analyzing prosody in archives of spontaneous speech [12]. Together, these tools form a method of prosodic description that is both accessible and efficient, while balancing both control and naturalness.

1.1. Nasal

The present study is focused on the intonation of Nasal (ISO 639-3: nsy; glottocode nasa1239), a previously undocumented Malayo-Polynesian language spoken by around 3,000 speakers in Bengkulu Province, Sumatra, Indonesia [13]. The Nasal documentation project is a long-term collaboration between linguists from both Indonesia and abroad and native speakers of a variety of languages spoken in the Nasal speech community. The project is multi-faceted, featuring both researcher- and community-led efforts that take a variety of forms, most often audiovisual recordings. These recordings can be found in an open-access PARADISEC archive [14] (found at https://catalog.paradisec.org.au/repository/BJM02).

Until this project there has been no dedicated study of the prosody of Nasal or the local languages spoken by the community (with the exception of the Besemah dialect of South Barisan Malay (ISO 639-3: pse); see [15] and [16] for descriptions of stress). It is not clear whether Nasal features lexical stress, though languages in the region, especially Malayic, have often been assumed to feature a pattern of penultimate stress (e.g. [17]). However, much research has been dedicated to the question of whether this assumption reflects a true stress phenomenon, or whether it might be the result of analyzing words in isolation [18], or first-language effects [19], among other methodological issues. The debate over stress in Indonesian in particular is too lengthy to summarize here, but is a clear ex-

10.21437/SpeechProsody.2024-237

ample of the importance of acoustic analysis of the phonetics of the region's languages. Particularly relevant to this research is the conclusion in [20] that boundary tones are a primary method of demarcating phrases in non-stress languages.

In a pilot study of Nasal prosody, [21] showed that Nasal intonation patterns primarily feature pitch movements at the right edge of utterances. The boundary tone shapes described in [21] are outlined further in Section 4. Building on this previous research, this exploratory study uses novel methods of analysis in order to answer the following questions:

- What are the prototypical boundary tone shapes in Nasal?
- How can tune prototypes be efficiently and effectively described as part of language documentation?

These preliminary findings will be a springboard for further work, some of which is already in progress. All participants who were part of these role-play tasks were also recorded playing a Map Task game (based on [2]). Analysis of these recordings will inform a more complete description of Nasal prosody, including both lexical and post-lexical, that is currently underway.

2. Methods

2.1. Participants

To investigate tune types in Nasal, native speakers were recorded reading scripted dialogues as part of a role-play task. For this preliminary study, which is part of a larger work in progress, recordings from four sessions were used, featuring four male (ages 32 - 54) and four female (ages 30 - 53) participants. Participants were recorded in same-gender pairs, and all speakers identify as native speakers of Nasal.

2.2. Procedure

The role-play task features two sets of scripted dialogues, with a separate set for male and female participants.[1] Each set contains three dialogues with different real-life scenarios (for men, meeting on the road, planning to go to the market, and meeting after returning from a trip; and for women, an invitation to make a cake, planning to harvest rice, and planning to collect snails).

Before recording, participants were each given printouts of the three dialogues and instructed to read the lines from one of the two roles (A or B). Participants were given time to practice and discuss the content of the dialogues before recording. Participants were recorded reading through all three dialogues once before switching roles and reading again. Dialogues were numbered and read in the same order each time. In cases of clear disfluency, readers were asked to repeat the previous line(s). Audio was recorded using a Sound Devices Mixpre-3M 3-track audio recorder. Participants were recorded through Shure SM35 headset microphones routed into independent mono channels on the Mixpre-3M. Additional audio was recorded through a Sennheiser ME66 microphone, which was placed on the ground in front of the speakers and was recorded into a separate channel. Audio was recorded at 96kHz and 24-bit. Video was recorded using a Panasonic HC-V875 HD Camcorder set up on a tripod, in AVCHD format at 60fps with 1920 x 1080 resolution.

[1] In a previous study, a single set was used, which was written by male speakers. Female participants quickly informed us that role-playing with these dialogues was awkward, both because of the content and the style of the scripted lines.

2.3. Stimuli

The stimuli used in this experiment were based on a set of dialogues used in the pilot study described in [21]. These dialogues were written in collaboration with three Nasal speakers, who provided the settings for each dialogue, wrote most of the lines and checked them for naturalness. The scripts were intended to contain a variety of sentence types (e.g. polar question, wh-question, imperative, lists; see Table 1); the number of these different types was unbalanced, since naturalness of the exchange between participants was prioritized over experimental symmetry. Additionally, the dialogues were reviewed and edited in an attempt to reduce microprosodic effects (especially near boundaries). Again, this was not possible in every case, and naturalness was prioritized. All utterances were labelled according to the coding format described in [3]. HCRC coding was originally designed to be used with map tasks (see [2]), but the coding has been applied here because of the variety of sentence types it includes, as well as the possibility of comparing this data with data from other experiments (such as map tasks).

2.4. Analysis

The recordings used for analysis were transcribed in ELAN ([22]) by two native speakers, with time-aligned annotations that included direct transcription, discourse transcription, translation, and annotator notes. Annotations were segmented into Intonation Units (IUs; see [23] for detailed definition). IUs were checked and annotations were additionally coded with HCRC dialogue coding before being exported as Praat TextGrids [24]. In Praat, an additional tier was created to further segment the final word of each utterance so that a cluster analysis (see below) could be applied with the utterance-final word(s) as the domain of analysis (this domain was chosen based on the pilot study of Nasal prosody, which noted that the most obvious pitch movements in an utterance overwhelmingly occur over the final word or two). In most cases, only the last word of each utterance was included; however, in some cases it was clear the pitch excursion began in an earlier word, usually in sentences ending in discourse particles, vocatives, or tag questions. Because of this, for all utterances ending in such words the preceding word was also included. Since the dialogues cover a variety of topics, and speakers placed boundaries at different locations in the scripted sentences, the tokens range from one to seven syllables. While an analysis of tokens featuring more limited (or identical) syllable lengths would bring more control to the study, the materials used here focused on emulating natural discourse, and the wide range of syllable lengths for tokens used more organically reflect the many possible alignments of boundary tones that may occur in Nasal speech. During the transcription phase in Praat disfluencies were also marked to be removed from the dataset for analysis.

2.5. Cluster analysis

Cluster analysis was performed in R [25] using the Contour Clustering app described in [4]. Pitch tracks generated by R are grouped through hierarchical clustering, a method by which data are grouped by numerical similarity used for "discovering groups in data" [26, p. 7]. This method of analysis is particularly useful for exploring prosodic patterns, since it does not rely on any *a priori* assignment of categories or labels, which often leads to researcher bias in prosodic analysis [4]. Measurements were taken with a pitch range of 50 - 500 Hz, a time-step of 10 ms, an f0 fit of 0.6, 25 measurement points per token, and

Table 1: *Frequency of occurrences of sentence types by HCRC code.*

HCRC sentence type	# of occurrences
reply-w	143
explain	88
query-w	86
acknowledge	61
instruct	42
clarify	24
query-yn	24
reply-n	23
ready	13
reply-y	7
Total	511

a smoothing bandwidth of 2. (Higher smoothing was chosen to reduce segmental effects in clustering of different tokens). To enable comparison of pitch data between speakers, f0 measurements were normalized using the Octave-Median (OMe) scale, which offers a more natural scale for standardizing pitch by estimating the neutral pitch range for each speaker [27]. Thus pitch measurements are presented as points between -1 and 1 octave from the median of the neutral range of each speaker.

Clusters were generated using Euclidian (L2) distances and a complete linkage criterion. No error correction was applied; from the dataset generated, clusters were first plotted with the maximum possible number of clusters to visually inspect and manually remove pitch tracks that featured obvious pitch errors (e.g. halving or doubling). Next, analyses with different numbers of clusters were plotted, starting with two clusters and increasing the number until adding clusters only split clusters which already featured a small number of tokens, did not reveal any significant change in pitch shape, and did not result in any increase in symmetry. Using this method, an analysis of the data with 10 clusters was chosen. Next, the cluster data were exported back into TextGrids and merged with the original TextGrids. These were then converted to .csvs and imported into R for further exploration and analysis.

An analysis with ten clusters was chosen for this study as it not only illuminates major differences between boundary tone shapes, but also highlights distinctions between pitch patterns that might otherwise be grouped together. The boundary tone shapes grouped into these ten clusters are shown in Figure 1. Decreasing the number of clusters did reduce asymmetry by grouping the tenth cluster with another cluster; however, this cluster reveals a tune shape that seems to occur regularly (if not frequently) in Nasal speech. This number is also appropriate in that it is similar to the number of boundary tones described in [21], though it may not be surprising that a detailed acoustic analysis revealed more variation than did prosodic annotation (by a non-native speaker). There is still quite a bit of asymmetry between clusters (compare the largest, Cluster 3, with 24.7% of the data to the smallest, Cluster 10, with 2.7% of the data), but such asymmetry is expected in conversation, since the sentence types included were also not symmetrical: Table 1 shows the distribution of different sentence types among all dialogues.

3. Results

Below the pitch patterns that emerged from the cluster analysis are described, incorporating observations of boundary tones

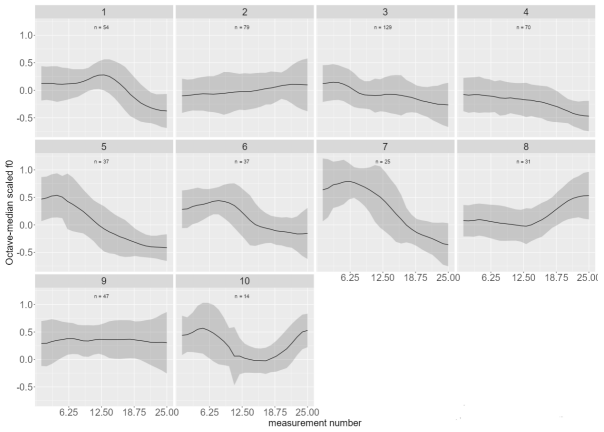

Figure 1: *Plots of pitch patterns at the right edge of utterance with ten clusters.*

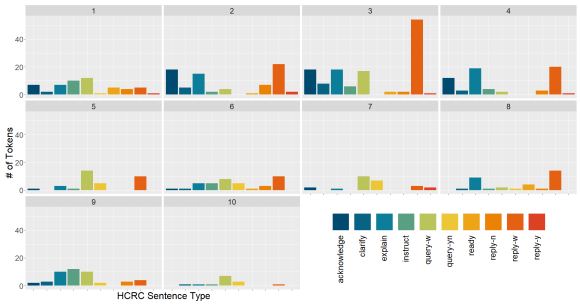

Figure 2: *Distribution of HCRC sentence types by cluster.*[2]

previously identified for Nasal.

3.0.1. Flat

Cluster 9 shows a pitch pattern that starts slightly above average pitch and remains relatively flat over the course of the utterance. This pattern makes up about 9% of the total dataset, and is similar to the no boundary tone (%) pattern previously observed. This cluster is made up of a mix of sentence types, with the majority represented by imperatives, declaratives, and wh-questions.

3.0.2. Rise

Cluster 2 (∼15% of tokens) shows a pitch pattern that starts just below average pitch and gradually rises over the course of the utterance. This corresponds to the H% boundary tone described in [21], and is overwhelmingly made up of declaratives, especially replies to wh-questions and backchanneling responses (acknowledge moves).

3.0.3. Fall

Cluster 4 (∼13% of tokens) shows a falling boundary tone that begins near average pitch and gradually falls to about half an octave below average. This corresponds to [21]'s L% boundary tone and features a similar makeup of sentence types to that of Cluster 2, but with a greater proportion of declaratives (explain) and fewer backchanneling responses.

[2]Palette colors from the PNWColors R package [28].

3.0.4. Rise-fall

Although this pattern was described as one category in [21] (labelled as HL%), no fewer than *four* clusters emerged with pitch trajectories with rise-fall patterns: Cluster 1 (∼10%), Cluster 5 (∼7%), Cluster 6 (∼7%), and Cluster 7 (∼5%). These clusters feature questions, both wh- and yn-questions, as the most common sentence type (with the exception of Cluster 6, whose most common category is responses, followed closely by wh-questions). Thus it seems that in Nasal, all questions, including polar questions, feature a falling tone, a pattern that is crosslinguistically uncommon [29].

Cluster 5 features a pitch pattern that most closely resembles the HL% pattern previously observed: the pitch rises to a relatively high peak near the start of the utterance before falling to a low target well below the average pitch. Cluster 7 is similar, but with a higher pitch peak that is achieved slightly later. Cluster 1, on the other hand, starts with a relatively flat pitch trajectory just above average pitch before rising slightly to a peak around halfway through the utterance, then falling to well below average pitch. Cluster 6 resembles Cluster 1, but a gradual rise to the pitch peak rather than an abrupt rise, and a slightly earlier alignment in the word. These differences may be due to the variable alignment of pitch peaks on the different syllables of the final word(s) in an utterance: so far, no predictable pattern has been identified, and the same token produced by different speakers may feature tones aligned with different syllables.

3.0.5. Fall-rise

Cluster 8 (∼6%) shows a fall-rise pitch pattern in which a clear low target precedes a high boundary tone. In this cluster, pitch starts slightly above the average before dipping down to a trough, then rising sharply to about half an octave above average. This cluster consists mostly of clarify (i.e. adding additional information to a y/n response) and wh-responses. Based on the responses included in this cluster, the rise seems to come from both a continuing intonation pattern and a combination of vocatives and tag questions, which are always realized with a sharp rise at the end of the phrase. Though this pattern is also similar to the unusual rise-dip-rise (H!HH%) boundary tone previously described, the pitch movements here seem to be too sharp to align with that tone type. (In fact, tokens which impressionistically match the rise-dip-rise pattern are scattered throughout other clusters – something to be investigated in future work.)

3.0.6. Rise-fall-rise

Cluster 10, the smallest of the clusters with only ∼2.6% of tokens, shows an early rise to a pitch peak high above the average pitch, a dip to a trough around average pitch, and a sharp rise to the same height as the first peak. This cluster consists almost entirely of questions (of both types). An examination of the specific utterances included in this cluster show that almost all tokens in this cluster feature a vocative or tag question at the end of the utterance, similar to those in Cluster 8. The main distinction here seems to be the presence of these token types in questions rather than declaratives. This highlights an important interaction between intonation and morphosyntax that may be explored with further dedicated study.

3.0.7. Rise-fall-rise-fall

Cluster 3 shows the rise-fall-rise-fall pitch pattern that is the most common in this dataset, making up around ∼25% of the data. That is around the same percentage that is made up by replies to wh-questions, and this is clearly reflected in the distribution data (see Figure 2). Interestingly this pattern, which apparently features four tone targets, seems to be the default intonation pattern for presenting new information.

4. Discussion

Based on the descriptions in Section 3, it is clear that the outcomes of the cluster analysis were closely in line with those based on manual annotations of pitch tracks in previous work. The main distinction between these methods is, again, that manual prosodic annotation requires a top-down assignment of labels and categories to continuous data (and is subject to non-native transcriber bias), while a cluster analysis takes a bottom-up approach to grouping pitch patterns based on similarity of acoustic data. This is not to suggest that such an analysis take the place of careful description based on other methods, such as manual annotation; rather, it can serve as a very effective starting point (or supplement) to more traditional, time-intensive analytical methods. The cluster analysis was also able to illuminate what seem to be important fine-grained differences in pitch trajectories and alignment (specifically for the rise-fall boundary tone pattern) that were not captured in previous work on describing Nasal intonation. The crucial distinction is the amount of time required for analysis: manual prosodic annotation is not required at all for cluster analysis, reducing the initial time spent in the transcription phase by a huge amount. Because of the relative ease of preparing the data, cluster analysis can also be applied to recordings of spontaneous speech (as noted in [4]). This has important implications for projects in language documentation, including the Nasal project, whose primary goal often features a large archive of conversational recordings. The findings from controlled studies like this one can directly inform further analysis featuring large sets of naturalistic data from language archives to form a picture of prosody in use across a speech community.

This study also shows the usefulness of role-play dialogues in prosodic description: role-playing tasks provide a means to control material in an utterance, and compare similar material between speakers, while also eliciting material that retains similarity to natural speech. The naturalness of the outcome of these tasks depends, of course, on the participants involved; in this case, participants were excited to engage in an activity slightly more involved than reading repetitive carrier sentences. Having consistent tokens between speakers also aids in comparability of results, particularly with methods such as a cluster analysis, where the placement of the same token in different clusters may reflect differences in speaker production, or elucidate an important distinction between different boundary tones. It is my hope also that the accessibility and efficiency of methods described here will help pave the way toward more regular and widespread documentation in other minority languages, both in Indonesia and farther afield.

5. Acknowledgements

This work was conducted with support from the National Science Foundation. I am also grateful for the support of the Bilinski Foundation, and the American Institute for Indonesian Studies, as well as Pusat Kajian Bahasa dan Budaya at Atma Jaya Catholic University of Indonesia. Finally, none of this research would be possible without the dedicated work and support of Nasal speakers, especially Johan Safri and Wawan Sahrozi.

6. References

[1] N. P. Himmelmann and D. Kaufman, "Prosodic systems: Austronesia," in *The Oxford Handbook of Prosody*, C. Gussenhoven and A. Chen, Eds. Oxford, England: Oxford University Press, 2020, p. 17.

[2] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert, "The Hcrc Map Task Corpus," *Language and Speech*, vol. 34, no. 4, pp. 351–366, Oct. 1991.

[3] J. Carletta, A. Isard, S. Isard, J. Kowtko, A. Newlands, G. Doherty-Sneddon, and A. Anderson, "HCRC Dialogue Structure Coding Manual," 1996.

[4] C. Kaland, "Contour clustering: A field-data-driven approach for documenting and analysing prototypical f0 contours," *Journal of the International Phonetic Association*, vol. 53, no. 1, pp. 159–188, Apr. 2023.

[5] P. Austin, "Language documentation 20 years on," in *IMPACT: Studies in Language, Culture and Society*, L. Filipović and M. Pütz, Eds. Amsterdam: John Benjamins Publishing Company, Oct. 2016, vol. 42, pp. 147–170.

[6] N. P. Himmelmann, "Documentary and descriptive linguistics," *Linguistics*, vol. 36, no. 1, 1998.

[7] A. L. Berez-Kroeker, B. McDonnell, E. Koller, and L. B. Collister, Eds., *The Open Handbook of Linguistic Data Management*. The MIT Press, Jan. 2022.

[8] R. Goedemans and H. van der Hulst, "StressTyp: A database for word accentual patterns in the world's languages," in *Empirical Approaches to Language Typology [EALT]*, M. Everaert, S. Musgrave, and A. Dimitriadis, Eds. Berlin, New York: Mouton de Gruyter, Jan. 2009.

[9] S.-A. Jun, Ed., *Prosodic typology II: the phonology of intonation and phrasing*, ser. Oxford linguistics. Oxford: Oxford University Press, 2014, oCLC: ocn856981095.

[10] J. McDonough, "Documenting intonational prosody: comparison of three Dene/Athabaskan (ISO 639-3) languages using data from different tasks," *University of Rochester Working Papers in Language Sciences*, vol. 7, no. 1, pp. 75–92, 2019.

[11] 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.

[12] S.-A. Jun and J. Fletcher, "Methodology of studying intonation: from data collection to data analysis," in *Prosodic Typology II*, S.-A. Jun, Ed. Oxford University Press, Jan. 2014, pp. 493–519.

[13] B. McDonnell, "Documenting Multilingualism in Southwest Sumatra," in *Indonesian Languages and Linguistics: State of the Field*. Ithaca: SEAP Publications, Cornell University Press., in press.

[14] ——, "Documentation of the multilingual linguistic practices of the Nasal speech community," 2020, publisher: PARADISEC.

[15] ——, "Acoustic correlates of stress in Besemah," *NUSA*, vol. 60, pp. 1–28, 2016.

[16] B. McDonnell and R. Turnbull, "Neural network modeling of prosodic prominence in Besemah (Malayic, Indonesia)," in *9th International Conference on Speech Prosody 2018*. ISCA, Jun. 2018,

[17] R. Blust, *The Austronesian languages*. Pacific Linguistics, Research School of Pacific and Asian Studies, Australian National University, 2013.

[18] M. Gordon, "Disentangling stress and pitch-accent: a typology of prominence at different prosodic levels," in *Word Stress*, 1st ed., H. van der Hulst, Ed. Cambridge University Press, Jun. 2014, pp. 83–118.

[19] E. van Zanten, R. Goedemans, and J. Pacilly, "The status of word stress in Indonesian," in *Current Issues in Linguistic Theory*, J. van de Weijer, V. J. van Heuven, and H. van der Hulst, Eds. Amsterdam: John Benjamins Publishing Company, 2003, vol. 234, pp. 151–175.

[20] E. van Zanten and R. Goedemans, "Prominence in Indonesian Stress, phrases, and boundaries," *Wacana, Journal of the Humanities of Indonesia*, vol. 11, no. 2, p. 197, Oct. 2009.

[21] J. Hakim, "A look at prosody in Nasal," in *2020 proceedings: Selected papers from the twenty-fourth college-wide conference for students in languages, linguistics & literature*. University of Hawaii at Mānoa: National Foreign Language Resource Center, 2021,

[22] "ELAN," Nijmegen, 2023.

[23] J. W. Du Bois, S. Cumming, S. Schuetze-Coburn, and D. Paolino, "Discourse Transcription," *Santa Barbara Papers in Linguistics*, vol. 4, 1992.

[24] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2024.

[25] R Core Team, "R: A Language and Environment for Statistical Computing," Vienna, Austria, 2019.

[26] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*, 5th ed. Chicester, Chichester, West Sussex: Wiley, Wiley, 2010, oCLC: 1058143621.

[27] C. D. Looze and D. Hirst, "The OMe (Octave-Median) scale: a natural scale for speech melody." in *Speech Prosody 2014*. ISCA, May 2014,

[28] J. Lawlor, "jakelawlor/PNWColors: Initial Release," Aug. 2020.

[29] C. Gussenhoven, "Paralinguistics: Three Biological Codes," in *The Phonology of Tone and Intonation*, ser. Research Surveys in Linguistics. Cambridge University Press, 2004, pp. 71–96.