University Library



A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Billington, R;Stoakes, H;Thieberger, N

Title:
The pacific expansion: Optimizing phonetic transcription of archival corpora

Date:
2021-01-01

Citation:
Billington, R., Stoakes, H. & Thieberger, N. (2021). The pacific expansion: Optimizing phonetic transcription of archival corpora. Proceedings of the Annual Conference of the International Speech Communication Association Interspeech, 3, pp.1713-1717. ISCA-INT SPEECH COMMUNICATION ASSOC. https://doi.org/10.21437/Interspeech.2021-2167.

Persistent Link:
https://hdl.handle.net/11343/297124

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

The Pacific Expansion: Optimizing phonetic transcription of archival corpora

Rosey Billington[12], *Hywel Stoakes*[2], *Nick Thieberger*[23]

[1]The Australian National University, Australia
[2]The University of Melbourne, Australia
[3]ARC Centre of Excellence for the Dynamics of Language, Australia

rosey.billington@anu.edu.au, hstoakes@unimelb.edu.au, thien@unimelb.edu.au

Abstract

For most of the world's languages, detailed phonetic analyses across different aspects of the sound system do not exist, due in part to limitations in available speech data and tools for efficiently processing such data for low-resource languages. Archival language documentation collections offer opportunities to extend the scope and scale of phonetic research on low-resource languages, and developments in methods for automatic recognition and alignment of speech facilitate the preparation of phonetic corpora based on these collections. We present a case study applying speech modelling and forced alignment methods to narrative data for Nafsan, an Oceanic language of central Vanuatu. We examine the accuracy of the forced-aligned phonetic labelling based on limited speech data used in the modelling process, and compare acoustic and durational measures of 17,851 vowel tokens for 11 speakers with previous experimental phonetic data for Nafsan. Results point to the suitability of archival data for large-scale studies of phonetic variation in low-resource languages, and also suggest that this approach can feasibly be used as a starting point in expanding to phonetic comparisons across closely-related Oceanic languages.

Index Terms: phonetics, forced alignment, low-resource languages, Nafsan, Oceanic, language documentation

1. Introduction

1.1. The cross-linguistic scope of phonetic documentation

While recent developments in phonological typology have led to a renewed interest in the range of segmental inventories found across the world's languages [1], understanding the diversity of sound systems in spoken languages also requires detailed analyses of phonetic patterns and phonetic typology. Of the 7,000 languages in the world, very few can be considered to have comprehensive phonetic documentation spanning the characteristics of vowels, consonants, and prosodic patterns, and Indo-European languages have remained the focus of descriptive phonetic research in recent decades [2]. Languages of the Pacific are particularly under-represented in phonetic research. For example, Vanuatu is one of the most linguistically diverse regions on the planet [3], but very few languages spoken there have received detailed phonetic investigations. The genetic and typological connections between the languages of Vanuatu remain an active area of research, and the striking phonological diversity across the archipelago is an area of particular interest. Phonetic analysis techniques offer opportunities to empirically compare speech patterns and investigate the intra- and inter-language homogeneity of specific phonological categories, allowing for new insights into the linguistic organisation of these languages. However, such analyses are dependent on the availability of suitable speech data.

1.2. The Nafsan language

Nafsan, also known as South Efate, is a Southern Oceanic language spoken in the villages Erakor, Eratap and Pango on the island of Efate in central Vanuatu, by an estimated 6,000 people. Following earlier historical-comparative work [4, 5], Nafsan has been the focus of detailed language documentation and description, and the compilation of extensive corpus materials [6, 7]. More recent research includes acoustic phonetic analyses based on controlled experimental data [8]. In relation to the sound system, descriptive work has highlighted characteristics of cross-linguistic interest, such as complex phonotactic structures and the presence of labial-velar consonants [9]. The phonetic analyses have also provided evidence for a contrast between short and long vowels in the language [10], raising questions about the historical and typological status of vowel length in languages of central Vanuatu. No targeted phonetic studies of vowels have yet been undertaken for other languages on Efate and adjacent small islands, but for several of these languages there is speech data available through archival sources, which could be drawn on to complement and extend research on these languages and rapidly advance crosslinguistic comparisons.

1.3. Corpus phonetics tools and methods

Phonetic research increasingly draws on large datasets of natural speech, facilitated by innovations in automatic speech recognition which enable (semi)automatic alignment of phonetic annotations to the speech signal, and significantly increase the speed of data processing and analysis [11, 12]. Many software tools emerging from these innovations are based on speech and language models trained on huge corpora for major, well-resourced languages such as English. There is significant interest in adapting or tailoring forced alignment tools and techniques for use with data for low-resource languages, particularly material originating from archives and language documentation projects [13, 14, 15]. Phonetic annotation of such material allows greater inclusion of understudied languages in research on crosslinguistic phonetic patterns [16, 17] and explorations of language-internal patterns of variation and change [18]. However, the limited size and coverage of many language documentation corpora, and level of associated linguistic description, pose challenges for the development of effective speech technologies.

2. Aims of this project

The aim of this project is to quickly and accurately segment narrative speech data for Nafsan based on archival recordings, using a pipeline of novel and existing tools (e.g. [19] [20]), and to extract and analyse phonetic data pertaining to vowels, an area of particular interest in the Nafsan sound system and in cen-

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-2167

tral Vanuatu typology. Given the availability of previous hand-labelled controlled speech data for Nafsan as a 'ground truth' set, a further aim is to compare the phonetic patterns of vowels in the forced-aligned data with those in the hand-labelled data, to establish if the output of the forced alignment shows phonologically plausible patterns. The longer-term goal of this work is to create a framework for crosslinguistic phonetic analyses of a comparative subset of languages spoken across Vanuatu.

3. Materials and methods

3.1. Speech materials

This project draws on a collection of speech material for Nafsan archived in the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). The collection includes ∼130 narratives recorded with close to fifty Nafsan speakers during fieldwork undertaken by the third author, primarily between 1995-–2000 [21]. These are audio-recorded and in some cases video-recorded speech, with associated orthographic transcriptions in ELAN [22] and accompanying metadata. The transcriptions are informed by the contributions of community members and extensive linguistic analysis. Here, we report on results so far using a subset of this corpus, comprising 2h 13m Nafsan speech data for 11 speakers (8 female, 3 male) aged in their 20s-80s at the time of recording. We refer to this as the forced-aligned corpus, and compare it to a smaller hand-labelled corpus gathered by the first author. Both corpora were recorded under field conditions on Efate island in Vanuatu.

Audio and transcription files were prepared to meet the required input formats of the Montreal Forced Aligner [19]. Audio recordings were processed to be homogeneous in terms of sample rate and bit depth (ideally 44.1 kHz, 16bit, mono files). The existing ELAN transcriptions are broadly time-aligned containing segments larger than individual utterances and without demarcation of pauses. The divisions were hand-corrected to correspond more closely to utterances incorporating 1–2 breath units (typically 3–5 second stretches of speech). These text annotations were then exported as Praat Textrgrids [23] containing a single tier with time-aligned utterance level segmentation.

3.2. Tokenised dictionary and G2P

The utterance level tier was used to compile a tokenised dictionary. All unique words in the utterances were extracted, non-speech symbols removed and then converted to lower case. Comments and metalinguistic information were removed before further processing. Then, using grapheme to phoneme (G2P) rules, each head word was converted into a phonemic representation. The final dictionary output is a two column, tab separated value (tsv) text file (UTF-8 NFC LF no BOM, see [24, p. 35]) with the second column containing a phonemic representation generated by G2P rules separated by spaces. This tokenised dictionary was then combined with an existing tokenised master dictionary of Nafsan words [7], and any words not represented in this combined set can be added as they appear in the utterance text.

The orthography of Nafsan is relatively phonetically transparent, and only a few symbols need to be changed in the G2P tokenisation rules. Some symbols that required particular attention were those representing complex articulations such as the labial-velar stops and nasals. The graphemes ⟨p̃⟩ and ⟨m̃⟩ are used in the orthographic representations of the labial-velar stop [k͡p] and labial-velar nasal [ŋ͡m], and in some earlier transcriptions ⟨p$⟩ and ⟨m$⟩ symbols are used. The ⟨g⟩ symbol is used to represent the velar nasal [ŋ], a widespread convention across languages of Vanuatu [25, p. 51]. Symbols or graphemes incorporating diacritics were checked for consistency and normalised to an X-SAMPA trigraph, for example, ⟨p̃⟩ → ⟨k_p⟩ and ⟨m̃⟩ → ⟨N_m⟩. This reduces ambiguity and ensures that these phonemes are modelled separately to plain bilabials and velars (see [24, pp 20–35] for further information).

3.3. Modelling and forced alignment

The modelling phase and subsequent forced alignment uses the Montreal Forced Aligner (MFA) tool [19]. MFA comprises a set of scripts that extract the phonetic information from the Kaldi speech recognition engine [20] and output a folder of TextGrids. All iterations of MFA were run on a cloud instance of Ubuntu 18.04 (Bionic) comprising 8 virtual CPUs with 32GB of RAM.[1]

Each proposed phoneme listed in the grapheme to phoneme table requires multiple representative examples of the same speech sound within similar linguistic contexts in order to accurately train a complete acoustic model. To streamline the input corpus, accurate utterance-level transcriptions, labelled in a consistent, unambiguous orthography are used to train separate HMMs (Hidden Markov Models) for each linguistically contrastive phoneme computed using MFCC (mel-frequency cepstral coefficients) and extracting distinctive features. As described in [19], thirteen MFCCs are calculated using a 25 ms window size and 10 ms frame shift. Feature calculation has a frequency ceiling of 8 kHz which enables building and using acoustic models that are comprised of various sample rates.

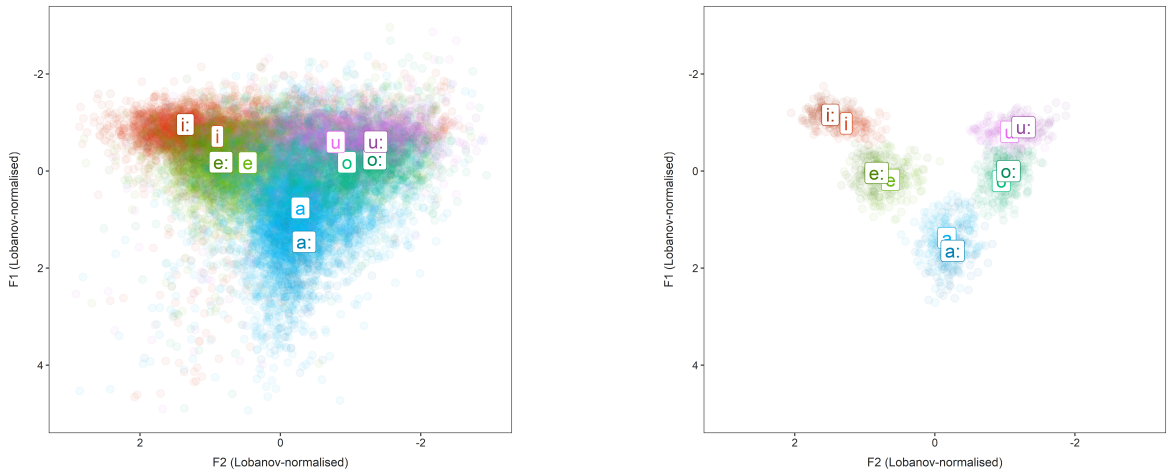
For this study, MFA ran multiple passes of Kaldi recipes starting with monophone models and then iteratively training triphone models (separate models for each group of 3 contiguous phones). The final stages of training apply a Linear Discriminant Method model and finally the model is refined using a Speaker-adapted training (SAT) method. Triphone and speaker adapted models require significantly greater volumes of data from differing linguistic contexts to successfully train.

The resultant Textgrids consist of time-aligned word and phonetic information contained on separate tiers. These were subsequently re-combined with the original utterance level transcriptions and these combined Textgrids form the corpus used in the phonetic analysis below.

3.4. Acoustic measurements

The final Textgrids were used to extract phonetic measurements for the narrative speech data via Praat. Statistical analysis and visualisations of vowels were undertaken within the R [26] and Rstudio [27] software environment. The forced-aligned narrative data yielded a total of 17,851 vowel tokens for the 11 speakers represented. A breakdown by vowel quality is shown in Table 1. The unbalanced ratio of vowels represented as short compared to vowels represented as long is notable. While on the one hand, this may reflect the lower functional load of phonemically long vowels in the Nafsan phonological system, it is also very likely that there are phonemically long vowels within the narrative data that have not been represented as such orthographically, given that the status of long vowels, and their inclusion in updates to lexical representations, has only arisen in more recent work.

[1] In this study, version 1.0.1 of MFA is called but it should be noted that the tool has recently been updated to v2.0 which updates the installation process and changes the command but does not significantly alter the back-end structures of the tool.



(a) *Forced-aligned narrative data*

(b) *Hand-labelled controlled data*

Figure 1: *Quality of vowels in Nafsan based on midpoint F1 and F2*

Table 1: *Vowel tokens by type in forced-aligned narrative data*

Phone	Count	Phone	Count
i	3762	i:	49
e	3704	e:	241
a	5817	a:	284
o	2234	o:	102
u	1580	u:	78

The acoustic measurements in the present study include vowel quality, based on formant frequencies, and vowel duration. Measurements of first formant frequency (F1) and second formant frequency (F2) were taken at the midpoint of each vowel token (in Hz); more minimal effects of consonant environment are expected at midpoints, which approximate a vowel target (e.g. [28]). All formant measurements were Lobanov normalised for visualisations [29], and comparisons were undertaken using raw Hz values. Measures of vowel duration are both presented and compared using absolute values (ms).

These are compared to formant and duration measurements from previous work based on controlled speech data from an experiment specifically focused on vowel distinctions in Nafsan [10]. Recordings were made with 7 speakers (3 female, 4 male) producing monosyllabic words in an utterance-medial frame. The controlled speech dataset contains a total of 1,372 vowel tokens roughly balanced according to vowel quality and phonemic length. For the controlled data, initial segmentation was undertaken via the web interface of the Munich AUtomatic Segmentation tool (MAUS) [30] using the language-independent model, and then annotations of the beginning and end of each vowel underwent manual checks and corrections for accuracy.

4. Results

4.1. Vowel quality

Formant values for Nafsan vowel tokens in the forced-aligned narrative corpus reveal a vowel space that looks much as would be expected for a language with five basic vowel quality distinctions, being broadly triangular in shape (see Figure 1a). While some outliers are apparent, most likely related to formant tracking errors, there are clear patterns for each of the five vowel qualities [i], [e], [a], [o], [u]. This is similar to the overall vowel space in the hand-labelled controlled speech (see Figure 1b), as also indicated by statistical tests of Euclidean distances.

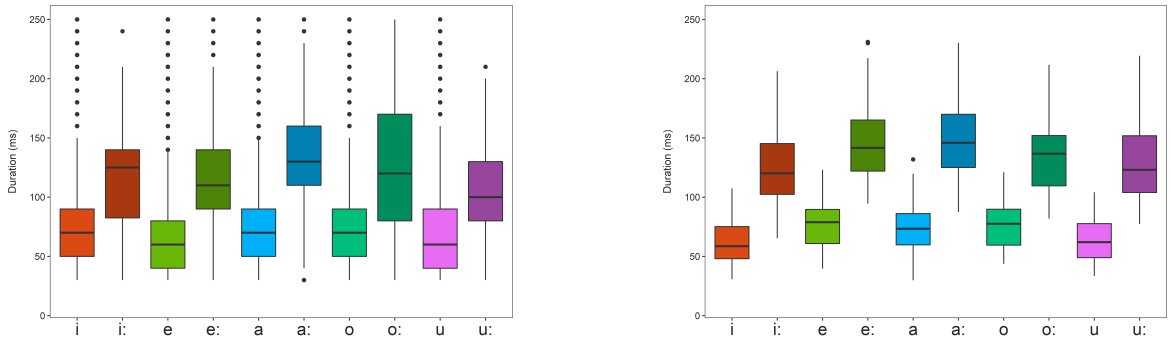
There is some evidence of centralisation of the short vowels compared to the long vowels in the forced-aligned narrative corpus. In comparison, centralisation of short vowels is very minimal in the hand-labelled controlled speech corpus [10]. The more apparent centralisation in the forced-aligned data is most likely because this data is from natural, connected speech produced at a faster rate than the controlled experimental data (e.g. [31]). Furthermore, the controlled data comprised only vowel tokens produced in prosodically prominent monosyllables, and in other research exploring prominence in Nafsan, substantial vowel reduction was only observed for vowels in prosodically weak syllables (at least for the open vowels which were the focus of that study) [8]. As noted, it is likely that the forced-aligned narrative corpus also contains phonemically long vowels which have orthographically been represented as short. If that is the case, we might expect the centralisation of the short vowels to be more substantial after such instances have been identified and re-coded.

Table 2: *Mean (and s.d.) F1 and F2 of vowel tokens in forced-aligned narrative data*

Phone	F1 (Hz)	F2 (Hz)	Phone	F1 (Hz)	F2 (Hz)
i	403 (123)	2146 (490)	i:	365 (70)	2330 (469)
e	497 (110)	1933 (400)	e:	497 (83)	2150 (407)
a	660 (183)	1574 (304)	a:	765 (233)	1552 (281)
o	499 (116)	1244 (368)	o:	494 (115)	1050 (353)
u	425 (120)	1313 (442)	u:	413 (94)	1124 (359)

4.2. Duration

Duration results for vowels in the forced-aligned narrative corpus show clear distributional patterns associated with phonemic length (see Figure 2a). Vowel tokens labelled as long have higher mean duration values than vowel tokens labelled as short, across all five vowel qualities, as previously established for the hand-labelled controlled speech (see Figure 2b) [10]. In general, the duration differences in long and short vowel pairs for



(a) *Forced-aligned narrative data*

(b) *Hand-labelled controlled data*

Figure 2: *Duration of vowels in Nafsan*

each vowel quality (/i, i:/, /e, e:/, /a, a:/, /o, o:/, /u, u:/) are similar between the two groups. One visible difference between the two datasets is that for the forced-aligned narrative data in Figure 2a, there is greater overlap in the duration values of the short vowels with the range of duration values for the long vowels. While there are likely durational effects of pre-boundary lengthening in the narrative data, as previously noted for Nafsan in a separate study of intonation and focus [32], this pattern also suggests the likelihood that there are long vowels represented as short in the original data. There are also outliers with duration values as high as 1000 ms (not shown in Figure 2a), which are more likely related to errors in the alignment of phone boundaries. Consequently, for the statistical analysis, and in the summary of mean values shown in Table 3, we excluded duration values higher than 250 ms, the upper limit of values observed in [10]. This removed 512 tokens from the sample.

All vowel quality pairs in the forced-aligned narrative corpus are statistically significantly different when a linear mixed model is applied (duration $\sim$ length) with speaker as both a random intercept and random slope. When comparing between datasets (dur $\sim$ length * dataset) the null hypothesis could not be rejected for the two datasets, indicating that they are not significantly different to each other. However, vowel durations for each vowel quality were significantly different to each other when speaker was included as a random factor for slope and intercept. More detailed statistical investigation is needed, once the larger corpus is finalised, to assess the interactions between speaker-specific and quality-specific vowel length realisations across the two corpora.

In Figure 2a there is evidence of 10 ms binning in the duration measurements. Although the recognition algorithm picks out acoustic landmarks so as to place the initial value, each interval is calculated to the nearest 10 ms value and consequently 10 ms is the shortest observable duration. This effect is common of many forced alignment packages including ProsodyLab-Aligner (as noted by [14, footnote 24, p 92]) and The Munich AUtomatic Segmentation (MAUS) tool [30]. There are a number of possible ways to ameliorate this effect however, for example by introducing pseudo-random values into the signal [33], allowing for direct statistical comparison to hand-labelled data.

5. Discussion and Conclusions

This paper has shown that a functional acoustic model can be trained with minimal speech material, providing that the input is carefully curated based on the contributions of community members and linguists to the original transcriptions and analyses, and undergoes data cleaning and pre-processing to assist modelling performance. The hand-labelled ground truth dataset used as a point of comparison shows very similar patterns for vowel duration and formant frequencies to the forced-aligned dataset, despite the different characteristics of the source corpora, indicating that the forced-aligned data comprises reliable material for closer linguistic analyses. The model described here was trained on less than 3 hours of transcribed utterance-level data to give word and phone aligned speech. The output can be used to train larger less controlled models incorporating many more speakers, and allowing new interrogations of rich language documentation corpora. As the Nafsan language has had relatively comprehensive documentation, there are large amounts of further resources to draw upon in modelling, but this study demonstrates that promising results can be obtained even where the scope of the data is more limited. In future work it is hoped that this model may be used to inform alignment of languages spoken nearby which share similar phonologies (such as Eton, Lelepa and Nguna). This would allow cross-linguistic comparison of label types to look for phonetic differences within corresponding phonemes, taking into account individual speaker differences and potentially uncovering patterns of sociophonetic variation. This would also greatly expand our knowledge of the phonetics of the languages of Vanuatu and aid documentation and maintenance of languages in the region.

Table 3: *Mean (and s.d.) duration of vowel tokens in forced-aligned narrative data*

Phone	Dur (ms)	Phone	Dur (ms)
i	74 (40)	i:	124 (35)
e	70 (38)	e:	127 (42)
a	79 (41)	a:	138 (39)
o	77 (43)	o:	129 (45)
u	69 (38)	u:	118 (37)

6. Acknowledgements

We thank all the Nafsan speakers who have contributed to and facilitated this work over many years. Thanks also to our collaborator Prof. Janet Fletcher. Funding support from The University of Melbourne [ECR1322020] and the ARC Centre of Excellence for the Dynamics of Language [CE140100041] is gratefully acknowledged.

7. References

[1] S. Moran and D. McCloy, *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History, 2019. [Online]. Available: http://phoible.org

[2] D. H. Whalen, C. DiCanio, and R. Dockum, "Phonetic documentation in three collections: Topics and evolution," *Journal of the International Phonetic Association*, vol. FirstView, pp. 1–27, 2020.

[3] A. François, M. Franjieh, S. Lacrampe, and S. Schnell, "The exceptional linguistic density of Vanuatu," in *The languages of Vanuatu: Unity and diversity*, A. François, M. Franjieh, S. Lacrampe, and S. Schnell, Eds. Canberra: Asia-Pacific Linguistics, 2015, pp. 1–21.

[4] R. Clark, "The Efate dialects," *Te Reo*, vol. 28, pp. 3–35, 1985.

[5] J. Lynch, "South Efate phonological history," *Oceanic Linguistics*, vol. 39, no. 2, pp. 320–338, 2000.

[6] N. Thieberger, *A grammar of South Efate: An Oceanic language of Vanuatu*. Honolulu, Hawaii: University of Hawaii Press, 2006.

[7] N. Thieberger and members of the Erakor community, *A dictionary of Nafsan, South Efate, Vanuatu: Mpet Nafsan ni Erakor*. Honolulu, Hawaii: University of Hawaii Press, in press.

[8] R. Billington, J. Fletcher, N. Thieberger, and B. Volchok, "Acoustic evidence for right-edge prominence in Nafsan," *Journal of the Acoustical Society of America*, vol. 147, no. 4, pp. 2829–2844, 2020.

[9] R. Billington, N. Thieberger, and J. Fletcher, "Nafsan," *Journal of the International Phonetic Association*, in press.

[10] ——, "Quantity and quality interactions in the phonetic realisation of Nafsan vowels," under review. [Online]. Available: bit.ly/nafsan-vowels

[11] K. Evanini, S. Isard, and M. Liberman, "Automatic formant extraction for sociolinguistic analysis of large corpora," in *Proceedings of Interspeech 2009*. Brighton: ISCA, 2009, pp. 1655–1658.

[12] M. Y. Liberman, "Corpus phonetics," *Annual Review of Linguistics*, vol. 5, pp. 91–107, 2019.

[13] C. DiCanio, H. Nam, D. H. Whalen, H. T. Bunnell, J. D. Amith, and R. Castillo García, "Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment," *Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 2235–2246, 2013.

[14] L. S. Johnson, M. Di Paolo, and A. Bell, "Forced alignment for understudied language varieties: Testing Prosodylab-Aligner with Tongan data," *Language Documentation & Conservation*, vol. 12, pp. 80–123, 2018.

[15] S. Babinski, R. Dockum, D. Goldenberg, J. Hunter Craft, A. Fergus, and C. Bowern, "A Robin Hood approach to forced alignment: English-trained algorithms and their use on Australian languages," *Proceedings of the Linguistic Society of America*, vol. 4, no. 3, pp. 1–12, 2019.

[16] E. Chodroff, A. Golden, and C. Wilson, "Covariation of stop voice onset time across languages: Evidence for a universal constraint on phonetic realization," *Journal of the Acoustical Society of America*, vol. 145, no. 1, p. EL109, 2019.

[17] F. Seifart, J. Strunk, S. Danielsen, I. Hartmann, B. Pakendorf, S. Wichmann, A. Witzlack-Makarevich, N. P. Himmelmann, and B. Bickel, "The extent and degree of utterance-final word lengthening in spontaneous speech from ten languages," *Linguistics Vanguard*, in press.

[18] D. Barth, J. Grama, S. Gonzalez, and C. Travis, "Using forced alignment for sociophonetic research on a minority language," *University of Pennsylvania Working Papers in Linguistics*, vol. 25, no. 2, p. Article 2, 2020.

[19] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Proceedings of Interspeech 2017*, F. Lacerda, D. House, M. Heldner, J. Gustafson, S. Strombergsson, and M. Wlodarczak, Eds. Stockholm, Sweden: ISCA, 2017, pp. 498–502.

[20] D. Povey, G. Arnab, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proceedings of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.

[21] N. Thieberger, *Natrauswen nig Efat: Stories from South Efate*. Melbourne, Australia: University of Melbourne, 2011.

[22] ELAN, "Computer software (Version 6.0), the Language Archive, Max Planck Institute for Psycholinguistics," Nijmegen, 2020. [Online]. Available: https://tla.mpi.nl/tools/tla-tools/elan/

[23] P. Boersma and D. Weenink, *Praat: Doing phonetics by computer [Computer Program]*. The University of Amsterdam, 2021, retrieved:01/03/2021 from http://www.praat.org, Version: 6.1.40.

[24] S. Moran and M. Cysouw, *The Unicode cookbook for linguists*. Berlin, Germany: Language Science Press, 2018.

[25] R. Clark, *Leo Tuai: A comparative lexical study of North and Central Vanuatu languages*, ser. Pacific linguistics, S. A. Wurm, Ed. Canberra, Australia: Research School of Pacific and Asian Studies, The Australian National University, 2009, no. 603.

[26] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org

[27] RStudio Team, *RStudio: Integrated Development Environment for R*, RStudio, PBC., Boston, MA, 2021. [Online]. Available: http://www.rstudio.com/

[28] R. J. J. H. van Son and L. C. W. Pols, "Formant frequencies of Dutch vowels in a text, read at normal and fast rate," *Journal of the Acoustical Society of America*, vol. 88, no. 4, pp. 1683–1693, 1990.

[29] B. M. Lobanov, "Classification of Russian vowels spoken by different speakers," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 606–608, 1971.

[30] F. Schiel, C. Draxler, and J. Harrington, "Phonemic segmentation and labelling using the MAUS technique," in *Workshop New Tools and Methods for Very-Large-Scale Phonetics Research*, 2011.

[31] Y. Hirata and K. Tsukada, "Effects of speaking rate and vowel length on formant frequency displacement in Japanese," *Phonetica*, vol. 66, pp. 129–149, 2009.

[32] J. Fletcher, R. Billington, and N. Thieberger, "Prosodic marking of focus in Nafsan," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. Canberra: Australasian Speech Science and Technology Association, 2019, pp. 3787–3791.

[33] M. G. Kendall and B. B. Smith, "Randomness and random sampling numbers," *Journal of the Royal Statistical Society*, vol. 101, no. 1, pp. 147–166, 1938. [Online]. Available: http://www.jstor.org/stable/2980655