English-based acoustic models perform well in the forced alignment of two English-based Pacific Creoles

Sam Passmore
Australian National University

Lila San Roque
Australian National University

Kirsty Gillespie
Australian National University

Saurabh Kumar Nath
Australian National University

Kira Davey
Australian National University

Keira Mullan
Australian National University

Tim Cawley
The Defence Science and Technology Group

Jennifer Biggs
The Defence Science and Technology Group

Rosey Billington
Australian National University

Bethwyn Evans
Australian National University

Nick Thieberger
University of Melbourne

Nicholas Evans
Australian National University

Danielle Barth
Australian National University

Abstract

Expanding the breadth of languages used to study sociophonetic variation and change is an important step in the theoretical development of sociophonetics. As data archives grow, forced alignment can accelerate the study of sociophonetic variation in minority languages. This paper examines the application of English and custom-made acoustic models on the alignment of vowels in two Pacific Creoles, Tok Pisin (59 hours) and Bislama (38.5 hours). We find that English models perform acceptably well in both languages, and as well as humans in vowel environments described as 'Highly Reliable'. Custom models performed better in Bislama than Tok Pisin. We end the paper with recommendations on the use of cross-linguistic acoustic models in the case of English-Based Creoles.

1 Introduction

The phonetic analysis of spoken language is a central tenet of our understanding of language variation and change (Labov, 1963). What start as small phonetic changes can develop into salient markers of linguistic difference: among the English-based creoles of the Pacific, the transitivity marker on verbs is *-im* in Tok Pisin from Papua New Guinea (e.g. *planim* 'to plant', from plant him) but is *-em* in Bislama (Vanuatu) and Solomons Pijin (e.g. Bislama *planem* and Solomons Pijin *plandem/planem*). Currently, European languages dominate phonetic research (Tucker and Wright, 2020). To evaluate claims of universal, or at least common patterns of sound change, phonetic research needs to draw from a wider base of linguistic diversity (Tucker and Wright, 2020). Historic hurdles, like data availability, are being broken down through the centralisation and archiving of diverse linguistic data (Thieberger and Harris, 2022; Seifart et al., 2018). The next step is to identify a pipeline to process archived data and facilitate the quantitative analysis of a broader set of languages.

This study assesses whether the pipelines developed in resource-rich languages are sufficient to accelerate the study of lesser-resourced languages or if specific models and approaches must be developed. We specifically assess the domain of automatic segmentation (forced alignment) of vowels comparing the performance of English models to custom made models in English-Based Pacific Creoles.

Vowels are of a particular sociophonetic interest because their pronunciation often correlates with social boundaries within a given language (Labov, 1963). For both theoretical and technical reasons, sociophonetic studies consider the phonological environment of vowels in analysis (Di Paolo and Yaeger-Dror, 2011). By understanding vowel environments, researchers can identify potential drivers of variation and change. For example, the longitudinal study of sound change of English in Philadelphia describes a series of vowel changes, many of which are partly conditioned by the surrounding segments, and so determines the phonetic motivations of the changes (see Labov, 1994, 2001, 2020). From a technical view, we know that segmentabil-

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31172–31183

ity varies based on the vowel's environment (Turk et al., 2012; Gonzalez et al., 2020). Vowels that are next to obstruents (e.g. [t] or [p]) have clear phonological boundaries, but vowels in diphthongs or bordered by approximants are much less clear (e.g. [w] or [l]). We focus on the technical case, with the specific goal of determining whether an automatic alignment approach can get a low-resource language dataset ready for sociophonetic analysis.

We analyse newly archived corpora of Tok Pisin and Bislama, two English-based Pacific Creoles with national language status (Thieberger; Crowley, 1990; Barth, 2023; Smith and Siegel, 2013). Tok Pisin has 3–5 million speakers, including 500,000 primary users, while around 318,000 Ni-Vanuatu speak Bislama (Meyerhoff, 2013). Both languages are growing rapidly as primary languages, often at the expense of linguistic diversity (Kik et al., 2021; Kulick, 2019).

A central component of the automatic forced alignment pipeline is the acoustic model, which contains the information that relates phones to audio signal through Mel-Frequency cepstrum coefficients (MFCCs). State of the art acoustic models of English are trained on more than 3,600 hours of speech, including varieties of English from India, Nigeria, England, and the US/Americas (McAuliffe and Sonderegger, 2024). These large data models allow the quick and reliable creation of transcribed and word or phoneme aligned datasets that can be used for sociophonetic research. For example, Gnevsheva (2020) used forced alignment to illustrate the ethnolect variation in the production of vowels across generations and between monolingual and bilingual speakers of Russian/English in Melbourne, Australia.

Researchers typically rely on one of two strategies depending on the size and cleanliness of the data for alignment, as well as their programming proficiency: creating an acoustic model from a small amount of data (Language-Specific models), or relying on acoustic models from other languages (Cross-Language models, Chodroff et al., In Press).

Language-Specific models have produced good-enough results for further phonetic analysis from as few as 25 minutes of continuous transcribed speech (Chodroff et al., In Press). Urum (Turkic) and Evenki (Tungusic) audio files were aligned while varying the amount of data used for training, finding near ceiling level performance once models were built with more than 70 mins of data (91-96% Data Retention, 52-69% Precision), but good-enough performance with as low as 25 minutes (Data retention: 84-96%; Precision: 50-61%). Similar results are seen for Matukar Panau (Austronesian; Barth et al., 2020), and Nafsan (Austronesian; Billington et al., 2021). To the contrary, a Tongan acoustic model showed around 10% worse performance than human annotation (Johnson et al., 2018). In general, the performance of Language-Specific models is encouraging for those wanting to develop tools for under-resourced languages, but it is time-consuming and technically difficult, hence researcher interest in using existing models cross-linguistically.

If large cross-language acoustic models can be used on under-resourced languages, it would open the door to building phonological theory upon a base of diverse empirical and quantitative research by quickly creating phone aligned datasets using the increasing collection of archived data (Michaud et al., 2018). Some researchers working on languages without pre-existing acoustic models have begun using large language models (like English) (Chodroff et al., In Press; Jones et al., 2017; Walker and Meyerhoff, 2020; Solano et al., 2018), to align their data. For example: a model built on Italian has been used to align Australian Kriol (Jones et al., 2017), and an English model has been used to align Bequia Creole (Walker and Meyerhoff, 2020), and Cook Island Māori (Solano et al., 2018). With increased interest in using pre-trained acoustic models cross-linguistically, it is important we evaluate their accuracy on diverse languages. Chodroff et al. (In Press) showed that American English and Global English models performed equally as well as Language-Specific models for Urum and Evenki. Here, we extend the evaluation of English acoustic models to the English-based Pacific Creoles, and extend the comparison with two further possibilities when choosing an acoustic model for alignment. First, is to leverage knowledge of linguistic history to build models of historically similar languages, and secondly fine tuning large language models with data from low-resource languages (called model adaptation).

Although low-resource languages are often not in the position to build Language-Specific acoustic models, it might be possible to leverage existing knowledge of language history to agglomerate datasets across closely related languages. Generally, languages that are more closely related are also more likely to share vocabulary, phonological inventories, and orthography which could be

Table 1: Description of each acoustic model evaluated and the hours of languages they contain. T.P. is Tok Pisin, Bis. is Bislama.

Model	Languages	Training	Reference
Language-Specific	One of Tok Pisin or Bislama	59 (T.P.) 38.5 (Bis)	Barth 2023; Thieberger
Pacific Creole	Tok Pisin and Bislama	97.5	Barth 2023; Thieberger
English	Indian English; Nigerian English; UK English; US English	3,614.2	McAuliffe and Sonderegger 2024
English-Adapted	English Model and one of Tok Pisin or Bislama	3,673.2 (T.P.) 3,652.7 (Bis)	Barth 2023; Thieberger McAuliffe and Sonderegger 2024

combined to create a larger training dataset. We examine the historical dimension by evaluating the performance of English models on English-based Creoles, and evaluate the performance of a general Pacific Creoles model (i.e. not using English data).

Secondly, we look at model adaptation. Model adaptation is designed to improve the performance of an acoustic model when new speakers or acoustic conditions are introduced to a dataset (Lee and Gauvain, 1993). We leverage the idea of English-Creoles speakers as 'new speakers' to adapt existing English models to the nuances of Pacific Creoles and improve the performance of alignment where we have limited data. The precise relationship between phones and MFCCs is unlikely to be exactly the same across languages, but by adapting the large model estimates we might be able to tweak estimates to identify the patterns of the new language.

Overall, this paper will evaluate the performance of four different acoustic models' ability to align two Pacific Creoles, using the Montreal Forced Aligner (MFA; McAuliffe et al., 2017). The four acoustic models are: an acoustic model built using only data from that language (Language-Specific); a model built using all Pacific Creoles data (Pacific Creoles); an existing English acoustic model (McAuliffe and Sonderegger, 2024); and the same existing English model, adapted with either the Tok Pisin or Bislama data (English-Adapted). All models are compared to 5 minutes of hand-corrected boundaries for each language, which we treat as a gold-standard. Our data comes from recently compiled speech corpora of Pacific Creole languages, with a total of 98 hours, comprised of 59 hours of Tok Pisin (Papua New Guinea) and 38.5 hours of Bislama (Vanuatu). The performance of each model is evaluated on its ability to identify the correct phone in the transcription (Data Retention), whether aligned boundaries are within 20 ms of the gold-standard data (Boundary Performance), whether aligned boundaries contain the midpoint of the gold-standard interval (Midpoint Retention), and whether formants extracted from aligned intervals are less than 10% different to the gold-standard formant (Formant Accuracy).

2 Methods

2.1 Data

Recordings for both Creoles were collected between 2023 and 2025. Speech was elicited through inviting a speaker or speakers to introduce themselves and describe their life experiences. Efforts were made to secure two or more speakers within a session to foster a dialogue. In most cases the speaker was also invited to share their experiences of natural disasters, in particular cyclones and floods, expanding to themes such as disaster preparation, response, traditional knowledge, COVID-19 and/or other health issues, and for their views on security and safety. In all cases, recordings were made in the field and contain consistent background noise. These are sub-optimal conditions for phonetic research, but reflect the conditions in which minority language data is often recorded.

The Tok Pisin corpus contains free speech from 147 speakers, totalling 24GB of audio data. There are approximately half men and half women, with an age-range between 19 and 79, with a median age of 42. The Bislama corpus also consists of free speech data from 60 speakers (7.5GB). Approximately half of the participants are men and half are women, with ages from 20 to 70 years old and a median age of 39. Data was collected with informed consent under protocols approved by the ANU Human Research Ethics Committee.

Local community leaders and advisers were consulted as part of the recruitment and recording of participants.

Bislama and Tok Pisin are both analysed as having 5-vowel systems (Crowley, 2004; Smith and Siegel, 2013), corresponding to orthographic <a e i o u>. Transcription of the Bislama and Tok Pisin corpora is orthographic, and thus approximates phonemic transcription. Common variants or contractions (e.g., *blo* as a reduced form of *bilong* in Tok Pisin) are usually transcribed as per their pronunciation. Transcribers were usually native speakers of the languages, but some transcription was also undertaken by non-native speakers (Gillespie, San Roque, Barth, and Thieberger). Instances of code-switching to English or other languages (e.g. Matukar Panau, Nafsan) were demarcated in the transcription and were removed from the analysis.

For each language, approximately five minutes of conversation was extracted and manually corrected to be used as test data. Manual correction was performed by one coder in Bislama, and two non-overlapping coders in Tok Pisin, after being passed through the Language-Specific acoustic model pipeline. The five minutes of conversation was extracted in 15 second chunks from each conversation, at a random point within the conversation, and excluded from the training data. All code used to perform these experiments is held at https://osf.io/x9scw/. Processed data is available on request, with raw data available on PARADISEC. Tok Pisin is available at http://catalog.paradisec.org.au/collections/3PAC1 and Bislama is available at https://catalog.paradisec.org.au/repository/3PAC3.

2.2 Preparing the data

The identification of vowel boundaries varies based on the linguistic environment, ranging from the easy identification of vowels (such as between obstruents), to avoiding the analysis of vowels due to the ambiguity of their boundary (like next to a voiced fricative). In an automated sociophonetic analysis, it is important to focus analysis on vowels that can be reliably segmented so that subsequent analysis can draw reliable conclusions (Turk et al., 2012). Turk et al. (2012) offer three categories of difficulty when it comes to VC or CV phone boundaries: reliable, sometimes reliable, and avoid (Table 2). These levels of difficulty are based on varieties of English, Finnish, and Japanese. Given the Pacific Creoles' similarity to English, they provide an acceptable comparison.

Table 2: A summary of the segmentability of consonants next to vowels, as they apply to Pacific Creole languages. Adapted from Turk et al., 2012.

Difficulty	**Consonant Sets**
Reliable	Oral Stops [p t k b d g]; Sibilants [s, z]
Sometimes Reliable	Nasal Stops [m n ng]; Voiceless Fricatives [f]
Avoid	Central Lateral Approx. [w l]; Voiced Fricatives [v]

We create five categories of CVC vowel environments from Table 2: Vowels that are between two 'Reliable' boundaries (Highly Reliable); vowels that are between a 'Reliable' boundary and a 'Sometimes Reliable' boundary (in either order; Reliable); a vowel that is between two consonants that are 'Sometimes Reliable' (Moderate), a vowel that is between a 'Sometimes Reliable' consonant and a consonant to be avoided (Unreliable), and finally, a vowel between two consonants that are labelled avoid (Difficult). These categories create a scale of difficulty, while increasing the number of vowels tokens that can sit in each one of the environments. The performance of alignment algorithms is assessed across all vowels observed in test corpora (Bislama = 1,868 tokens; Tok Pisin = 1,878 tokens), and then as the subset of vowels that exist in one of the five difficulty environments (Bislama = 1,432 tokens; Tok Pisin = 1,377 tokens).

2.3 Acoustic Models

We evaluate the viability of four acoustic models within a forced alignment pipeline. They are: a Language-Specific model, a Pacific Creoles model, a pre-trained English model (McAuliffe and Sonderegger, 2024), and an English model that has been adapted to each Creole (the English-Adapted model) (See Table 1). These models can be further considered as two groups: Custom models, which includes the Language-Specific model and Pacific Creoles model (since they are custom built for the datasets), and English models, which include the English and English-Adapted models.

Language-Specific models are trained on the maximum available data for that language, and using the default parameters in the MFA v3.2.1 train algorithm (McAuliffe et al., 2017). The Pacific Creoles model is the sum of these datasets, 98.5

hours. We use the English MFA acoustic model v3.1.0 (McAuliffe and Sonderegger, 2024) as the pre-trained English model. The English-Adapted model uses the pre-trained English model as a base, and is adapted either using the Tok Pisin training data or the Bislama training data. Model adaptation is performed using a maximum a-posterior (MAP) approach (Young et al., 2002), as implemented in Kaldi and MFA (Povey et al., 2011; McAuliffe et al., 2017). All models were run on a MacBook Pro with Apple M1 Pro chip, and 16GB Ram.

2.4 Acoustic Model Performance

Following Chodroff et al. (In Press) we evaluate the performance of our four acoustic models across three metrics: Data Retention (how often does the algorithm return the right phone), Boundary Precision (how often the aligned boundary is within 20 ms of the gold-standard), and Midpoint Retention (how often the aligned phone contains the midpoint of the gold phone). We also examine formant accuracy (whether the F1 and F2 formants from the midpoint of the aligned segment is within 10% of the gold-standard formants). These four performance measures capture three dimensions needed for successfully creating an automatically aligned sociophonetic dataset: identifying phones, aligning boundaries, and extracting formants. Since the English models are trained on phonemic data, but the Creoles are trained on orthographic data, English models contain a wider range of phones (McAuliffe and Sonderegger, 2024) . To make the model more comparable, phones identified as /iː/ and /uː/ were mapped to /i/ and /u/, respectively.

While boundary performance will provide information on how well we can map between automatic and manual methods, formant extraction could feasibly be robust to some conditions of poor boundary estimation. For example, if automatic methods systematically identify wider boundaries than the gold-standard data, then the formant at the midpoint of both segments should be approximately the same.

To evaluate formant extraction performance we extract F1 and F2 using the default settings of Praat (Boersma and Weenink, 2025) for both the gold and automatically aligned datasets. All formant values are normalized by speaker using a variation of the Lobanov standardization, relative to the gold-standard formants.

3 Results

Table 3 shows the performance metrics for the segmentation of all vowel tokens across the four models, for both Tok Pisin and Bislama. In general, English models perform moderately well across the four performance measures. In Bislama, The Language-Specific model slightly outperforms the English models across all performance measures, whereas in Tok Pisin, English models consistently outperform both custom models in all measures except Data Retention.

3.1 Data Retention

Data Retention is high in all models for both languages (Table 3). The Language-Specific and Pacific Creoles models outperform the English and English-Adapted model for this metric. Custom models return a precision and recall of 100% for vowels across all difficulty conditions for both languages (i.e. Table 2). That is, vowels are always identified correctly and all errors occur in other vowel boundary conditions (such as diphthongs). English and English-Adapted models have 100% recall for all vowel conditions, but precision scores of around 95%. Surprisingly, all English and English-Adapted models make errors in the two most reliable vowel environments (Highly Reliable and Reliable boundaries). Across both Tok Pisin and Bislama, all misclassifications are /e/ vowels as /i/ vowels.

3.2 Boundary Precision

Across all vowels, English and English-Adapted models outperformed the custom models for Tok Pisin data, whereas in Bislama, the Language-Specific model showed the most accurate Boundary Precision (Table 3; Figure 1). If we only look at 'Highly Reliable' vowels in Tok Pisin, English and English-Adapted models increase to ~80% precision, and in Bislama, English and English-Adapted models perform comparably to the Language-Specific model (all around 77%). There is a general trend of model performance declining as vowel difficulty increases.

In both languages, performance is generally better for onset, rather than terminal boundaries (Bislama: 80% of the time, across all models; Tok Pisin: 70%). For Bislama when using the Language-Specific model (the best performing model), onset Boundary Precision is at 100% for the 'Highly Reliable' category, showing a fluctuating decline

Table 3: Performance results across all vowel tokens for all five metrics, and all four acoustic models.

Tok Pisin	**English**	**English-Adapted**	**Language-Specific**	**Pacific Creoles**
Data Retention	0.74	0.75	0.85	0.84
Precision w/n 20ms	0.43	0.45	0.29	0.28
Alignment Accuracy	0.52	0.53	0.34	0.35
F1 Accuracy (10%)	0.69	0.70	0.55	0.57
F2 Accuracy (10%)	0.72	0.73	0.55	0.59
Bislama	**English**	**English-Adapted**	**Language-Specific**	**Pacific Creoles**
Data Retention	0.75	0.74	0.91	0.91
Precision w/n 20ms	0.38	0.38	0.42	0.28
Alignment Accuracy	0.46	0.46	0.50	0.36
F1 Accuracy (10%)	0.71	0.70	0.74	0.61
F2 Accuracy (10%)	0.69	0.69	0.75	0.62

as difficulty increases: 80%, 90%, 85%, and 75% for the most difficult category. Among terminal boundaries, precision drops rapidly from 80% in the two easiest categories, down to 40% in the two most difficult categories. For Tok Pisin, English-Adapted models align 90% of onset boundaries for the two easiest alignment categories, but terminal boundaries drop from 91% in the 'Highly Reliable' category, at 75% in 'Reliable' intervals. Both onset and terminal boundaries follow a similar trajectory downward, at 75% precision for 'Moderate' vowels, to 67%, and to 50% in the most difficult category.

3.3 Midpoint Retention

Midpoint Retention is poor across all models and languages, when considering all vowels (Figure 1). Midpoint Retention is influenced by vowel length, because longer vowels have a wider margin to retain the midpoint. When examining vowels longer than 50 ms, the Tok Pisin corpus using an English-Adapted model increased accuracy to 100% for all phone boundaries, but did not improve performance in midpoint retention for Bislama. We consider vowels of all lengths in the remainder of this section.

Midpoint Retention irregularly decreases as difficulty increases, across all models. Notably, there is a sharp increase in performance of the 'Difficult' category among all models in Tok Pisin. This sharp rise consists of only 12 segments, all of which are greater than 50 ms long, which we have mentioned increases the probability of midpoint retention. The Bislama Language-Specific model also observes an increase in performance in the most difficult environments, which we can also attribute to small samples and longer vowels.

Midpoint Retention errors occur either because the aligned boundaries arrive before the midpoint (which we call undershooting) or after the midpoint (overshooting). Within the Bislama corpus, English models tended to undershoot the midpoint (~80% of errors), whereas Language-Specific models tended to overshoot (~60% of errors). In Tok Pisin, English models overshoot in 80% of errors, but English-Adapted models undershoot ~80% the time. The Language-Specific model, in Tok Pisin, is split evenly between over- and undershooting.

3.4 Formant Accuracy

Across all models and languages, there is no obvious performance difference between F1 and F2, and the best model for F1 accuracy was also always the best model for F2 accuracy. In Tok Pisin the best model is the English-Adapted model (F1 = 70%; F2 = 73%), and the Language-Specific model in Bislama (F1 = 74%; F2 = 75%; Figure 2).

By only examining vowels in 'Highly Reliable' environments, the best Bislama model improves by around 9% (F1 = 82%; F2 = 79%), but there is no change to the Tok Pisin model. Formant Accuracy generally decreased as vowel alignment difficulty increased, which is likely a result of poorer boundary performance. Errors in Formant Accuracy occur more frequently in long (>50 ms) vowels, than in short vowels. In the best performing models of each language (Tok Pisin = English-Adapted and Bislama = Language-Specific), 80% of Tok Pisin formant errors and 71% of Bislama formant errors are in long vowels.

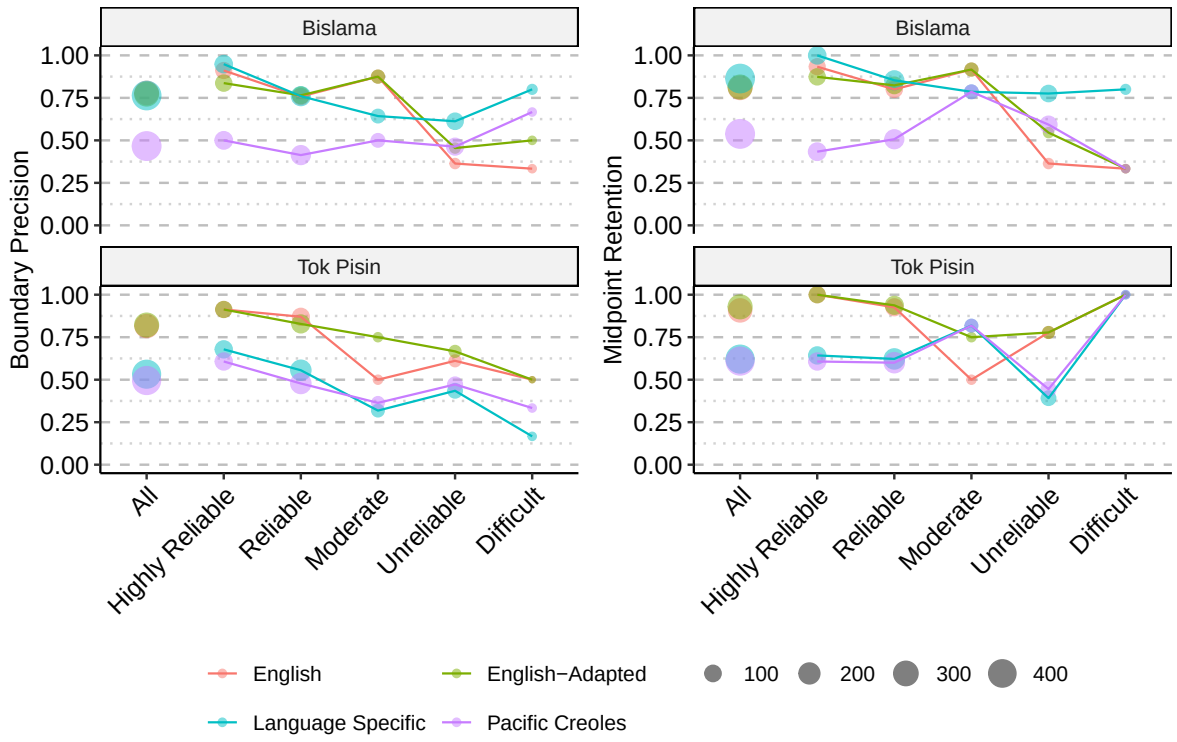

Figure 1: (Left) Boundary precision is measured as the proportion of times a forced aligned boundary is within 20 ms of the gold-standard boundary. (Right) Midpoint Retention is the proportion of tokens where forced aligned boundaries contain the midpoint of the gold-standard interval. Points are sized by the frequency of vowels in that environment.

4 Discussion

As data from more languages becomes readily available, there is a pressure to re-evaluate typological conclusions. In this study, we find that pre-existing acoustic models of English provide good-enough results for a first pass alignment in English-based Pacific Creoles when compared to models built specifically for those Creoles, or when agglomerating data across closely related languages. Adapting English models with data from the new language is a relatively straightforward step researchers can take to incrementally improve results. In Bislama, the custom Language-Specific model showed the best performance in all metrics, although English models performed only marginally worse. In contrast, English-based models performed considerably better than a Language-Specific model in Tok Pisin. Agglomerating data across closely related languages performed poorly across all metrics except data retention, and is not a recommended approach.

The performance of both boundary identification and formant extraction depends on the difficulty of the vowel environment. Vowels that sit between oral stops or sibilants showed higher levels of accuracy than vowels surrounded by central lateral approximants, and voiced fricatives. Research in both manual (Turk et al., 2012) and automatic (Gonzalez et al., 2020) alignment have identified similar patterns. An important limitation of our boundary categorisation is that 'Sometimes Reliable' boundaries are reliable in some situations. Varying reliability may explain the spike in performance within Boundary Precision and formant accuracy. What conditions make *reliable* environments reliable is probably a language-specific phenomenon and linguists should use their knowledge to identify what conditions these might be. Although boundary difficulty is likely to vary between languages, a general rule is that boundaries which humans find difficult to position are also likely to be difficult for automatic approaches. Researchers should utilize their manual alignment experience to determine which vowels are more or less likely to be accurately aligned when using automatic processes, or when prioritising boundaries to correct.

All models performed exceptionally well when

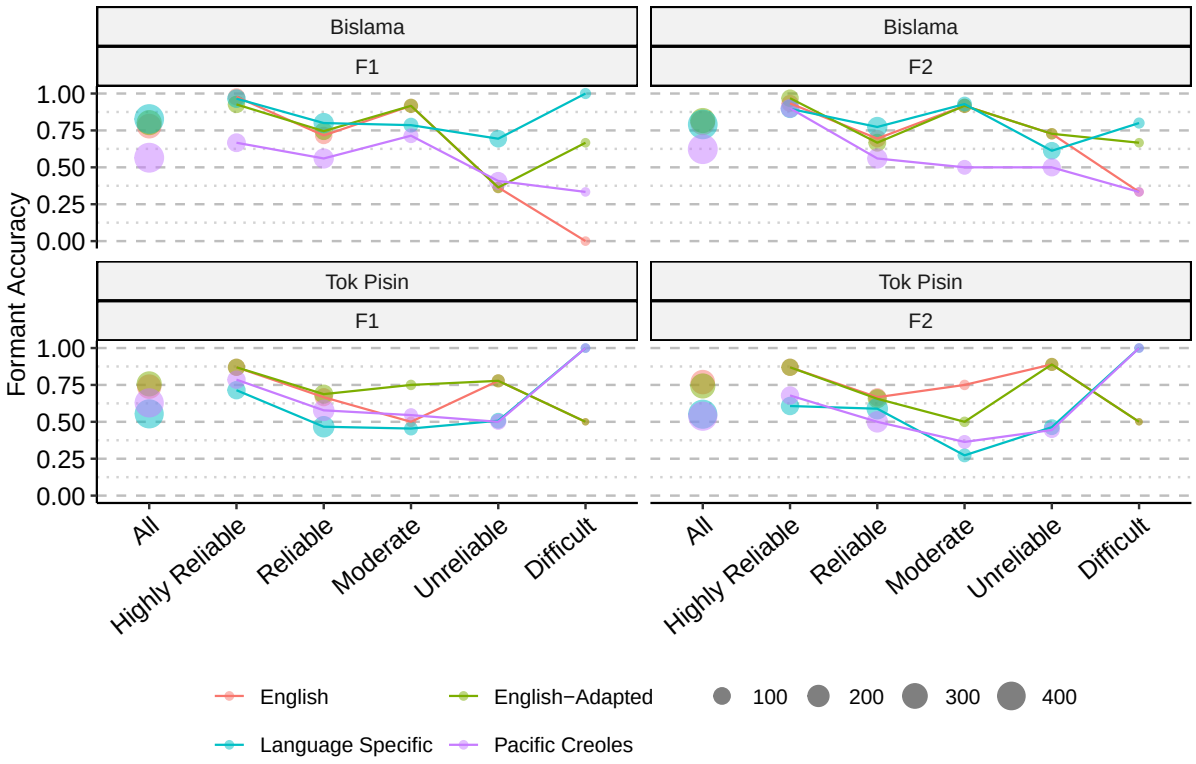

Figure 2: Formant Accuracy is taken as whether the forced aligned boundaries can extract a formant within 10% of the gold-standard. Performance is shown for all vowel tokens (All), and each category of vowel alignment difficulty (Highly Reliable - Difficult). Points are sized by the number of tokens.

identifying vowels. Although English and English-Adapted models performed well, the errors reveal cause for caution when applying English language models to other languages. In both the Tok Pisin and Bislama data, English-based models systematically label /e/ vowels as /i/ vowels, albeit only in around 4% of cases. If the sounding of vowels is systematically different in English that it is in the Pacific Creoles because of, for example, influence from local languages, then there is reason to suspect systematic errors in phone identification caused by the English model. For example: if /e/ vowels are more closed than we expect to find in English, as seen in the data retention errors, then they may be misclassified as /i/. However, the risk appears low.

The performance metrics for Boundary Precision, Midpoint Retention, and Formant Accuracy fall well below the theoretical ceilings of 100%. However, gold-standard alignments are only taken from one coder per language, and disagreement can exist within coders. Most existing work usually considers human agreement to sit at around 80% (Goldman, 2011; Gonzalez et al., 2020; DiCanio et al., 2013). Considering 80% performance as a more realistic ceiling (DiCanio et al., 2013), then, across the identified vowel environments, these acoustic models are producing comparable performance to humans in Boundary Precision (Bislama = 78%; Tok Pisin = 81%), but Midpoint Retention could be improved upon. Performance across all vowel tokens is about 40% worse than human alignment, leaving significant room for improvement compared to similar studies (Chodroff et al., In Press; Billington et al., 2021; Barth et al., 2020).

5 Conclusions and Recommendations

This paper sought to determine how close an automatic-alignment pipeline could get to developing a dataset suitable for phonetic analysis. Researchers whose aim is a completely automatic pipeline for other English–based Creoles should consider using an English acoustic model. This model is likely to provide the best out-of-the-box performance. For further sociophonetic analysis of the dataset, we recommend only analysing vowels that are in reliably segmented phonological environments. The more difficult the phonological environment, the higher the chance of a boundary, or formant extraction error.

Adapting the English model with language specific data provides an incremental performance increase, and is not difficult to implement (Young et al., 2002; McAuliffe et al., 2017). Models built specifically for the dataset at hand showed mixed success. In Tok Pisin, English models outperformed the Language-Specific model. This was despite the fact that Tok Pisin models were trained on more data than Bislama. We do not recommend agglomerating datasets across closely related languages, although future research might examine conditions where this approach may be viable.

If a researcher requires more confidence in the alignment, they should consider a manual review of boundaries. Our results suggest prioritising the review of more difficult vowel environments, and to focus on the alignment of terminal boundaries. If formant extraction is an important output, then researchers should also prioritize the review of vowels longer than 50 ms. This is particularly true if researchers are interested in specific vowel environments, which might dictate the segmentation of difficult vowels rather than vowels in general.

6 Limitations

This project is limited in at least two ways. First, the amount of training data varies between languages. This makes a comparison of performance in Language-Specific models unfair, but it is a constraint of our dataset. Future research could consider downsampling Tok Pisin to match Bislama for transcribed hours. Secondly, It is not clear why the Language-Specific model only performed well in Bislama, and not in Tok Pisin. Some possibilities are that there was a difference in audio quality between the two field sites, with Tok Pisin having worse quality on average, or that Tok Pisin speakers' way of speech is more difficult to parse computationally than Bislama speakers. Since we have no measure of sound quality, we are limited in drawing a strong conclusion. To improve this, future work could consider a qualitative description of sound quality across recordings.

6.1 Review of Errors

For researchers who may want to use automatic pipelines to extract vowel information on minority languages using English-based acoustic models, we offer an in-depth look at the types of errors that occur in the most stable CVC environments - the 'Highly Reliable' and 'Reliable' environments (48 vowel tokens in Bislama and 50 tokens in Tok Pisin). We describe the errors within Boundary Precision, Midpoint Retention, and Formant Accuracy, but we note that errors in one measurement tend to follow with errors in the others.

6.1.1 Boundary Precision Errors

When using the English model on Bislama and Tok Pisin, vowel boundary errors are concentrated in phonetically challenging environments. In Bislama there are 12 vowels with boundary errors, all at the final boundary, with five also showing onset errors. Most of these errors (11 of 12 finals and all onsets) involve the model undershooting the gold-standard, and two involve the word *kasem*, which is completely missed (errors are >2 seconds). Eight of these vowels occur before nasal consonants, where boundaries between modal phones are often unclear, while the remaining errors involve fast speech, breathiness, or frication that masks boundary cues. Background noise in some recordings further complicates alignment. Similarly, the Tok Pisin model shows 10 vowel boundary errors: one with both boundaries undershot, five with undershot final boundaries, and four with onset errors (three overshot, one undershot). Half of these may stem from annotation errors, while the rest involve difficult acoustic environments—nasals, whispered speech, rapid articulation, or noisy backgrounds.

In the English-Adapted results for Bislama and Tok Pisin, vowel boundary errors largely mirror those found in the English model. The Bislama results show 17 vowels with boundary errors, including all 12 vowels that appeared with errors in the English model. As before, all 17 involve final boundary errors, with four also showing onset errors. Most of the errors (14 final and all four onset) involve the model undershooting the gold-standard. Ten vowels appear before nasals, while others involve unusually short or long vowel durations or occur in recordings with droning background noise that obscures spectrographic cues. The Tok Pisin results show 14 vowels with errors, including one with both onset and final boundary errors (both undershot), 10 with only final errors (nine undershot, one overshot), and three with only onset errors (two undershot, one overshot). Nine of the 10 errors seen in the English Tok Pisin model persist in the adapted version. Six of the 14 errors may reflect mistakes in the gold-standard segmentation, while the remaining eight are tied to challenging acoustic conditions: four involve nasal-vowel tran-

sitions, one involves a likely phonological variation in *nogut* pronounced as [/noŋgut/], two are whispered, and one is both fast-spoken and noisy.

6.1.2 Midpoint Retention Errors

Midpoint Retention errors show similar patterning to Boundary Precision errors. When using an English model on Bislama data, all vowel midpoints are consistently undershot. Two tokens, from within the words *pikinini* and *tank*, involved nasal segments following the vowel that were not clearly distinguishable. Four errors came from *kasem* and *putum*, with fast speech contributing to errors in *putum*, while in *kasem*, although audio quality was good and boundaries seemed accurate, the labels were likely swapped (/a/labelled as /e/ and vice versa). In *pam*, a nasal carryover effect from the preceding /n/ likely influenced segmentation due to shared place of articulation with the following /p/ and /m/. In *paama*, phone-level segmentation followed expected patterns. Eight tokens 'Highly Reliable' or 'Reliable' vowels had alignment issues tied to difficult acoustic environments. For Tok Pisin, three tokens showed boundary errors—two overshot (*disla* and *nonap*) and one undershot (*pinga*). The segmentation in *disla* was acceptable despite unclear audio, *pinga* was affected by rain noise, and *nonap* was complicated by surrounding nasals.

When using the English-Adapted model for Bislama, vowel mid-points remained consistently undershot. Again, eight 'Highly Reliable' or 'Reliable' tokens were identified, two of which differed from the English model. Those errors repeated from the English model are likely for the same reasons. The two new tokens appeared in *bakegen*, which was segmented clearly as /p/ /a/ /g/ /e/ /n/, and *tek*, where segmentation was complicated by slight noise. For the Tok Pisin model, the error pattern reversed compared to the English model—two tokens were undershot (*disla* and *nogut*), and one was overshot (*pinga*). As before, *disla*'s segmentation was acceptable despite poor vowel clarity, *pinga* suffered from background rain noise, and *nogut* had a noisy signal where the vowel closing boundary may need refinement.

6.1.3 Formant Errors

When using the English acoustic model for the Bislama dataset we find 12 formant errors. Seven tokens have errors for both F1 and F2 measurements. One of these is /i/, which in this case is a short vowel measured close to the release of the preceding plosive /k/, with errors caused by poor boundary alignment and weak acoustic signal. The remaining dual errors are /a/, five which are caused by misalignment (including cases of incorrect phone identification), which relate to the boundary errors described above such as poor audio quality that disrupts formant tracking, while one error is a result of a formant extraction errors in Praat. There are two F1 errors, attributed to misalignment and incorrect formant analysis by Praat. The remaining three errors are in F2, two of which occur on well-aligned tokens. These errors are ascribed to issues with formant tracking in Praat due to weak F2 signalling, where the final token formant measurement is attributed to the vowel being taken very close to the release of the preceding bilabial plosive /p/ where formants are in transition. In the Tok Pisin data, four formant errors were identified: one token with only an F1 error, one with only an F2 error (linked to weak formant bands and background noise), and two with both F1 and F2 errors caused by misalignment and unclear formant structures, particularly in whispered or noisy recordings. Overall, these errors highlight the challenges of accurate formant measurement in low-resource, noisy, or phonetically complex contexts.

In the Bislama and Tok Pisin datasets aligned with the English-adapted acoustic model, most formant errors mirror those found in the unadapted model, with additional errors arising from misalignment and formant extraction issues. The Bislama data contains 13 formant errors, 11 of which are identical to those in the unadapted model and caused by the same factors, such as misalignment and poor audio quality. Two errors are new, with one /e/ token with an F1 error, which was misidentified due to complete misalignment and one /a/ token with an F2 error due to extraction issues from weak acoustic signal. The Tok Pisin data shows six errors, three of which match those in the unadapted model and stem from the same causes. Of the three new errors, one /o/ token shows inconsistent F1 extraction due to weak formant structure and background noise despite good alignment, one /i/ token has an F2 error caused by misalignment near a plosive release, and another /o/ token has alignment issues that place F2 measurements in a transitional region following a nasal.

In general, the patterns we observe reflect the shared challenges of aligning vowels in complex phonetic contexts across both models. In general,

we recommend that when correcting an automatic alignment, researchers pay attention to the alignment of final boundaries, vowels that are not spoken in a regular speaking voice or at a regular speed, and particular attention to vowels occurring before nasal consonants.

Acknowledgements

We thank all the speakers of Bislama and Tok Pisin who contributed to this corpus, as well as the community members who facilitated its construction. Thanks to Eleanor Chodroff for advice on phonetic and forced alignment procedures. This research is supported by the Commonwealth of Australia.

References

Danielle Barth. 2023. Tok pisin corpus.

Danielle Barth, James Grama, Simon Gonzalez Ochoa, and Catherine Travis. 2020. Using forced alignment for sociophonetic research on a minority language. *Selected Papers from NWAC47*, 25(2).

Rosey Billington, Hywel Stoakes, and Nick Thieberger. 2021. The pacific expansion: Optimizing phonetic transcription of archival corpora. In *Interspeech 2021*, pages 4029–4033. ISCA.

Paul Boersma and David Weenink. 2025. *Praat: Doing phonetics by computer*.

Eleanor Chodroff, Emily P Ahn, and Hossep Dolatian. In Press. Comparing language-specific and cross-language acoustic models for low-resource phonetic forced alignment. *Language Documentation and Conservation*.

Terry Crowley. 1990. *Beach-la-Mar to Bislama: The emergence of a national language in Vanuatu*. Oxford University Press.

Terry Crowley. 2004. *Bislama reference grammar*. Oceanic linguistics special publication (31). University of Hawaii Press, Honolulu.

Marianna Di Paolo and Malcah Yaeger-Dror. 2011. *Sociophonetics: a student's guide*. Routledge, London.

Christian DiCanio, Hosung Nam, Douglas H. Whalen, H. Timothy Bunnell, Jonathan D. Amith, and Rey Castillo García. 2013. Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *The Journal of the Acoustical Society of America*, 134(3):2235–2246.

Ksenia Gnevsheva. 2020. The role of style in the ethnolect: Style-shifting in the use of ethnolectal features in first- and second-generation speakers. *International Journal of Bilingualism*, 24(4):861–880.

Jean-Philippe Goldman. 2011. Easyalign: an automatic phonetic alignment tool under praat. In *Interspeech 2011*, pages 3233–3236. ISCA.

Simon Gonzalez, James Grama, and Catherine E. Travis. 2020. Comparing the performance of forced aligners used in sociophonetic research. *Linguistics Vanguard*, 6(1).

Lisa M. Johnson, Marianna Di Paolo, and Adrian Bell. 2018. Forced alignment for understudied language varieties: Testing prosodylab-aligner with tongan data. *Language Documentation and Conservation*, 12:80–123.

Caroline Jones, Katherine Demuth, Weicong Li, and Andre Almeida. 2017. Vowels in the barunga variety of north australian kriol. In *Interspeech 2017*, pages 219–223. ISCA.

Alfred Kik, Martin Adamec, Alexandra Y. Aikhenvald, Jarmila Bajzekova, Nigel Baro, Claire Bowern, Robert K. Colwell, Pavel Drozd, Pavel Duda, Sentiko Ibalim, Leonardo R. Jorge, Jane Mogina, Ben Ruli, Katerina Sam, Hannah Sarvasy, Simon Saulei, George D. Weiblen, Jan Zrzavy, and Vojtech Novotny. 2021. Language and ethnobiological skills decline precipitously in papua new guinea, the world's most linguistically diverse nation. *Proceedings of the National Academy of Sciences*, 118(22):e2100096118.

Don Kulick. 2019. *A Death in the Rainforest: How a Language and a Way of Life Came to an End in Papua New Guinea*. Algonquin Books of Chapel Hill, New York, UNITED STATES.

William Labov. 1963. The social motivation of a sound change. WORD, 19(3):273–309.

William Labov. 1994. *Principles of linguistic change, volume 1: Internal factors*. Blackwells, Cambridge, MA.

William Labov. 2001. *Principles of linguistic change, volume 3: Cognitive and cultural factors*, volume 3. Blackwells, Cambridge, MA.

William Labov. 2020. The regularity of regular sound change. *Language*, 96(1):42–59.

C.H. Lee and J.L. Gauvain. 1993. Speaker adaptation based on map estimation of hmm parameters. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 558–561 vol.2.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeeh 2017*, volume 2017, pages 498–502.

Michael McAuliffe and Morgan Sonderegger. 2024. English mfa acoustic model v3.1.0. Technical report, https://mfa-models.readthedocs.io/acoustic/English/English MFA acoustic model v3_1_0.html.

Miriam Meyerhoff. 2013. *Bislama*. Oxford University Press, Oxford.

Alexis Michaud, Oliver Adams, Trevor Anthony Cohn, Graham Neubig, and Séverine Guillaume. 2018. Integrating automatic transcription into the language documentation workflow: Experiments with na data and the persephone toolkit. *Language Documentation and Conservation*.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, and Petr Schwarz. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. Language documentation twenty-five years on. *Language*, 94(4):e324–e345.

Geoff P. Smith and Jeff Siegel. 2013. *Tok Pisin*. Oxford University Press, Oxford.

Rolando Coto Solano, Sally Akevai Nicholas, and Samantha Wray. 2018. Development of natural language processing tools for cook islands māori. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, page 26–33, Dunedin, New Zealand.

Nick Thieberger. Bislama corpus.

Nick Thieberger and Amanda Harris. 2022. When your data is my grandparents singing. digitisation and access for cultural records, the pacific and regional archive for digital sources in endangered cultures (paradisec). *Data Science Journal*, 21(1).

Benjamin V. Tucker and Richard Wright. 2020. Introduction to the special issue on the phonetics of under-documented languages. *The Journal of the Acoustical Society of America*, 147(4):2741–2744.

Alice Turk, Satsuki Nakai, and Mariko Sugahara. 2012. *Acoustic Segment Durations in Prosodic Research: A Practical Guide*, pages 1–28. De Gruyter.

James A. Walker and Miriam Meyerhoff. 2020. Pivots of the caribbean? low-back vowels in eastern caribbean english. *Linguistics*, 58(1):109–130.

Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, and Dan Povey. 2002. *The HTK book*, volume 3.2. Cambridge University Engineering Department, Cambridge.