The Use of Phone Categories and Cross-Language Modeling for Phone Alignment of Panãra

Emily P. Ahn[1], *Eleanor Chodroff*[2], *Myriam Lapierre*[1], *Gina-Anne Levow*[1]

[1]University of Washington, USA
[2]University of Zurich, Switzerland

eahn@uw.edu, eleanor.chodroff@uzh.ch, mylapier@uw.edu, levow@uw.edu

Abstract

Automating the time-alignment of phonetic labels in speech facilitates research in language documentation, yet such phonetic forced alignment requires pretrained acoustic models. For low-resource languages, this raises the question as to how and on which data the acoustic model should be trained. To align data from Panãra, an Amazonian indigenous language of Brazil, we investigated three approaches for forced alignment of low-resource languages using the Montreal Forced Aligner. First, we implemented a novel approach of manipulating the acoustic model granularity from phone-specific to increasingly broader natural class categories in training language-specific Panãra models. Second, we trained cross-language English models under two granularity settings. Third, we compared these models to a large, pretrained Global English acoustic model. Results showed that broadening phone categories can improve language-specific modeling, but cross-language modeling performed the best.

Index Terms: forced alignment, phonetics, phonology, low-resource languages

1. Introduction

The alignment of speech to utterance, word, and phone segments is a useful and often necessary step to studying language phenomena. Language research often faces the "transcription bottleneck", where many audio recordings may exist, but there is limited transcribed data [1]. Part of the transcription process includes alignment, and accurate time alignments enable comprehensive language documentation and research, from the acoustic-phonetic level to syntactic and discourse levels. For example, language typologists can utilize alignments to study the universality of sounds across language families [2, 3], and sociolinguists can discover patterns in the usage of vowels among different speaker groups [4]. Overcoming the transcription bottleneck would also increase the amount of usable data for Natural Language Processing (NLP) applications such as machine translation and automatic speech recognition, that could serve endangered language communities.

Obtaining these transcriptions and alignments is time-consuming to produce manually. Automating the placement of phone boundaries, i.e. forced alignment, has been shown to save considerable annotation time [5]. Forced alignment takes speech and its orthographic transcription and automatically produces time-aligned phone segmentation. Forced aligners, such as FAVE [6], WebMAUS [7], and the Montreal Forced Aligner (MFA) [8], rely on acoustic models of a language's phone categories that have been trained on annotated speech data. Pretrained acoustic models are available for many high-resource languages with large amounts of annotated speech data, yet for lower-resourced languages, acoustic models with minimal data need to be trained from scratch.

Alternatively, "cross-language forced alignment" can be implemented in which pretrained acoustic models of high-resourced languages are used to align a lower-resourced language. American English models have been helpful for aligning related English varieties such as British English [9] and North Australian Kriol [10], as well as less related languages such as Cook Islands Maori [11] and Yoloxóchitl Mixtec [12]. Bribri, a Chibchan language of Costa Rica, has also been phone-aligned with English-based FAVE and French-based EasyAlign [13] pretrained models [14].

While cross-language forced alignment benefits from robust amounts of training data for the acoustic model development, it lacks language-specificity for the target language. The phoneme inventory of the target language in a cross-language setting generally needs to be remapped to that of the pretrained acoustic model, as phoneme inventories are rarely one-to-one between languages. There are often sounds in target languages of research that do not exist in high-resourced languages such as English. Another complication is that phonemic units are commonly debated for a given language, and defining a language's inventory is by no means trivial[1] [15]. For example, the Hausa repository in the online PHOIBLE database [16] contains five entries with the stated inventory size ranging from 31 to 46 segments [17].

Given these concerns, we propose an alternative strategy to retain language-specificity in acoustic model training, which increases the amount of data per phone category by broadening phone categories to larger natural classes. Broad class models may prove to be more universal and would not discriminate against a target language that does not have the exact same phone set as the training data. It also may facilitate language-specific model training by expanding the number of instances per phone. As the task of alignment is to identify the transition point from one segment to the next in the acoustic signal, this may still be achievable with a coarser representation of a segment. This strategy was shown to improve cross-language, sentence-level alignment across five European languages in acoustic model development [18]. Alternatively, when comparing two English models' phone-level alignments on Yoloxóchitl Mixtec data, the model with a narrower, "context-sensitive" set of phones outperformed the broader, more phonemic model [12]. To our knowledge, our work is the first to utilize broad phone classes in both language-specific and cross-language modeling for phone-level alignment.

[1]For instance, it took the third author approximately 30 weeks of *in-situ* fieldwork in the Panãra Indigenous land, distributed over 5 years, to determine the phonemic inventory of Panãra.

In this paper, we defined three settings of phonetic granularity and evaluated the forced alignment performance of Panãra, an Amazonian language of Brazil. We analyzed these settings within both language-specific modeling of Panãra (i.e., training a model on just Panãra), and cross-language modeling (i.e., training on English and adapting to Panãra). We address the following research questions:

1. Does broadening phone categories improve alignment performance in language-specific training?
2. Does broadening phone categories improve alignment performance in cross-language training/modeling?
3. Do any of these strategies perform better than using a large, pretrained English model in cross-language alignment?

2. Data

We utilized datasets from two languages: Panãra, an endangered language that is undergoing active documentation by the third author, and English, a high-resourced language, which allowed us to investigate generalizability of the methods.

Panãra (ISO-639-3: kre) is a Jê language spoken in the state of Mato Grosso, Brazil, with approximately 700 speakers. Its phoneme inventory includes typologically less common nasality and length contrasts in both vowels and consonants [19]. Our Panãra dataset consisted of four free-speech recordings from four different speakers—two male and two female; the utterances span 35 minutes of speech. There were a total of 771 utterances that averaged 2.76 seconds each. All data was transcribed orthographically by the third author and corrected with a native speaker of Panãra.[2] Phonetic boundaries from two of the four recordings have been manually hand-corrected by a trained phonetician.

We also selected an English dataset that could be retrained from scratch to incorporate broad phone categories. The motivation for a broad category English model is that it would mimic a more language-independent, multilingual model that would be inclusive of unseen sounds in the target language. The TIMIT English dataset consists of read sentences from 630 speakers across 8 dialect regions in the US [20]. This corpus was selected as it has been manually transcribed at the phonetic level, as well as hand-aligned for phonetic boundaries. After removing some problematic data,[3] we utilized just under 4 hours of TIMIT speech from 519 speakers. This produced 5190 utterances with an average length of 3.08 seconds.

3. Methodology

Our pipeline included the following steps, described in detail below: (i) grapheme-to-phoneme conversion, (ii) lexicon creation, (iii) data management, (iv) acoustic model development and forced alignment, and (v) evaluation of aligned boundaries. As highlighted in Figure 1, the lexicon is the main component we manipulated for our experiments.[4]

[2]Though hours of recording were plenty, our available data was limited by the time-consuming process of transcribing audio into Panãra orthography, which requires approximately 2 hours of work per 1 minute (100 words) of audio. Orthographic transcription is a task that requires expert knowledge and the presence of both the third author and one of the few fully literate speakers of Panãra.

[3]Two of the eight training folders did not run through the Montreal Forced Aligner, so we excluded that data.

[4]All code is publicly available at `https://github.com/emilyahn/force_align`.

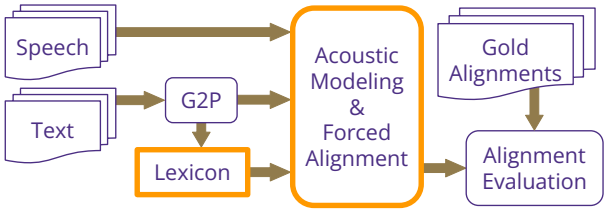

Figure 1: *Our pipeline takes speech and its corresponding text transcriptions as input. We produce a phone sequence with a grapheme-to-phoneme (G2P) system. We then manipulate the lexicon for each model to allow for broadening phone categories and conducting cross-language alignment. We train acoustic models and produce phone-level alignments, which we then evaluate against human-annotated "gold" alignments.*

Table 1: *A subset of the lexicon mapping from the explicit Panãra setting to each of the other language-specific and cross-language settings. Each setting shows the number of distinct phone categories in that model which corresponds to Panãra phones. The Broad category symbols follow the SCA classes in [22]; all other symbols are in IPA.*

Map from	Map to			
Panãra Explicit	**Panãra No-Diacritics**	**Broad (SCA)**	**TIMIT Explicit**	**Global English**
63	29	17	25	30
e	e	E	eɪ	e
eː	e	E	eɪ	eː
ẽ	e	E	eɪ	e
ẽː	e	E	eɪ	eː
n	n	N	n	n
nː	n	N	n	n
ŋ	ŋ	N	ŋ	ŋ
ɲ	ɲ	N	n	n
s	s	S	s	s
sː	s	S	s	s
k	k	K	k	k
kː	k	K	k	k

First, the Panãra data went through an automatic rule-based grapheme-to-phoneme (G2P) system to convert orthography into phones.[5] There were instances of code-switched Portuguese words in the Panãra speech, for which we applied the Epitran G2P in Portuguese [21]. There were only 34 Portuguese words in the entire dataset, and these had been tagged with the language specification directly in the transcription. None of the G2P output was hand-corrected.

Second, we created three settings of phone categories in the lexicon creation stage that informed the language-specific training for Panãra-only acoustic models: an "Explicit", "No Diacritics", and "Broad natural class" setting. The first setting, "Explicit", used the given phone categories, which consisted of 63 phonetic labels from the default output from the G2P system. The second setting, "No Diacritics", had 29 phonetic labels where all length and nasalization markers were ignored. Lastly, we created a "Broad" phone categories setting, which followed the natural classes from the Sound-Class-Based Phonetic Alignment (SCA) tool [22]. Each of our phones was mapped to one of the 28 SCA classes that included 6 vowels, 5 fricatives, 3 plosives, 1 affricate, 2 nasals, 1 laryngeal, 2 approximants, 1 trill/tap/flap, and 7 tones. The Panãra phone set

[5]This G2P system is an internal tool created by collaborator Teela Huff.

had 17 of these sound classes.

For our second research question, we trained the TIMIT English acoustic models under two phonetic granularity settings. The "Explicit" setting had 46 fine-grained phonetic labels.[6] The "Broad" setting also used the SCA classes [22], which included 19 of these sound classes.

For our third research question, we conducted a final comparison of a large pretrained English model's alignment on Panãra data to address whether large cross-language acoustic models may still outperform the smaller acoustic models. The Global English model [24] was the largest pretrained acoustic model available from MFA, consisting of 3770 hours of speech from regions including the US, UK, Nigeria, and India. As this model would have been difficult to retrain from scratch, we did not apply our broad phonetic categories methodology.

To align Panãra data with the Explicit versions of these English models, we also created lexicons that mapped explicit Panãra phones to the respective English phone sets. Table 1 displays a portion of the Panãra phone set and its mappings to the different settings and cross-language phone inventories, while Table 2 displays an example Panãra utterance and the phone sequences it corresponded to using each trained model's lexicon. The main differences between the explicit English phone inventories and Panãra were the following: the Global English model did not have nasalized vowels or lengthened consonants, the TIMIT Explicit model did not contain any nasalized or lengthened sounds, and neither of these had the unrounded back vowels [ɯ] and [ɤ].

Table 2: *An example utterance in Panãra including the original orthography, the phone sequence mapping per lexicon and acoustic model, and the English translation.*

Panãra Orthography	Haa mämä jynkjân rasu hapôô
Panãra Explicit	h a: m ɤ̃ m ɤ̃ j ɯ ŋ k j ɤ n ɾ a s u h a p o:
Panãra No Diacritics	h a m ɤ m ɤ j ɯ ŋ k j ɤ n ɾ a s u h a p o
Broad (SCA)	H A M E M E J I N K J E N R A S Y H A P U
TIMIT English	h a m ə m ə j ɪ ŋ k j ə n ɾ a s u h a p oʊ
Global English	h a: m o m o j u ŋ k j o n ɾ a s u h a p o:
Translation	"Well, this time you arrived (where I am)"

For a fair comparison, the Panãra dataset was split into 3 speakers for training and 1 speaker for testing. Since we have manually-annotated "gold" alignments for two speakers (1 male and 1 female), we implemented two train/test configurations that each held out one of these two speakers for testing. This cross-validation scheme allowed us to not contaminate the test set with a speaker from the train set, yet utilize all of our manually-annotated data. This resulted in the Panãra train set totaling either 27 or 31 minutes of speech, and the aggregated test set including 12 minutes of speech. For TIMIT, we created two versions of the train set; the "full" version had 224 minutes and 495 speakers, and the "small" version had 26 minutes and 51 speakers. This "small" TIMIT train set was devised to have a comparable duration to the Panãra train set. The resulting acoustic models were then used to align the Panãra test set.

After preparing speech audio files, their corresponding phone sequences, and the lexicons, we passed these through the Montreal Forced Aligner (MFA) Version 2.2.17, a tool easily configurable to train and adapt acoustic models from scratch [8]. For our experiments, we applied the 3 different granularity settings to a Panãra-only train and test scenario, and then applied two granularity settings to train the TIMIT English models. All English models were adapted using the Panãra train set, which slightly modified the acoustic model parameters. We used the default MFA settings for all model training and adaptation, which employed a triphone GMM-HMM architecture with MFCCs.

[6]While TIMIT originally had 61 phone labels, they are not typically all used [23]. We collapsed several categories such as "b-closure" with [b], and [ʌ] with [ə].

Table 3: *Alignment performance across various systems on the Panãra test set, as measured by phone boundary onset displacement within 20ms. English-trained models have been adapted to the Panãra train set.*

Trained Dataset	**Trained Settings (# Phone Categories)**	**Accuracy (%) ↑**
Panãra	Explicit (63)	60.20
	No Diacritics (29)	62.35
	Broad (17)	61.92
TIMIT English	Explicit Full (46)	62.65
	Explicit Small (46)	66.09
	Broad Full (19)	56.07
	Broad Small (19)	61.14
Global English	Explicit (100)	**69.82**

Our methods for measuring alignment performance used phone onset boundary differences between system and manually-annotated 'gold' versions. We define boundary accuracy as the percentage of system onsets that are within 20 milliseconds of the corresponding gold onsets (higher is better) [8, 9]. For the case of fine-grained phonetic and phonological analysis, researchers will likely use forced alignment as a first-pass to then manually adjust the phone boundaries.

4. Results

For our first research question related to the influence of broadening phone categories on alignment performance in language-specific training, we examine the top three rows in Table 3. These show that for language-specific training, i.e., the Panãra-trained models, broadening the phone categories beyond the Explicit setting was beneficial. Between removing diacritics and using the broad SCA natural classes, the No Diacritics model performed the highest at an accuracy of 62.35% for the onset boundary to be within 20 milliseconds of the gold boundary. The Broad SCA Panãra model performed slightly worse at 61.92%, suggesting that broadening the categories too much lost specificity for the acoustic model phone categories. The Explicit Panãra model had 63 phone categories, which was likely too many for its training data size, given that its performance on the test set was only 60.2%.

As for our second research question related to cross-language acoustic model training and alignment, we examine the middle four rows in Table 3. For the TIMIT English-trained models that were adapted on Panãra, using broad phone classes degraded performance; Broad SCA models for each data size (Full and Small) performed worse than their respective Explicit models. Even though the broad TIMIT models were aimed at mimicking language-agnostic, multilingual models, this did not help in alignment of Panãra data. The best TIMIT model was the Explicit Small, which yielded a 66.09% accuracy.

As for our third research question, if any of these strategies above can outperform a large, pretrained model, our answer is no. The Global English model performed the highest across all models. Even with 100 phone categories in its acoustic model, the 30 categories (see Table 1) that pertained to the

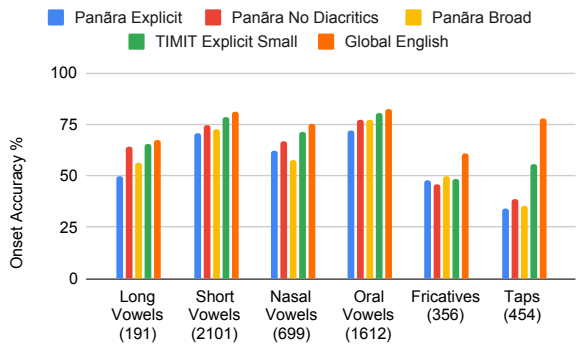

Figure 2: *Onset boundary accuracy within 20ms (y-axis) across a selection of natural classes presented with their token counts (x-axis), on Panãra test data. The colored bars represent 5 of the systems.*

Panãra phone set were language-specific enough to outperform any of the other models.

Beyond these primary questions, we also sought to understand the performance patterns within specific natural sound classes across the different configurations. Figure 2 reveals several findings. Among vowels, short vowels were more accurate than long vowels, and oral vowels were more accurate than nasal vowels. Within Panãra-only systems, the No Diacritics model performed better than the Broad model for long and nasal vowels, implying that combining the vowels into fewer natural class categories lost specificity in the acoustic model training. Some other notable phone-specific patterns involved [h], which performed particularly poorly among fricatives and overall, as well as the tap [ɾ] which performed especially poorly for all Panãra models. The best TIMIT English model outperformed all the other Panãra models except for among long vowels, with which it performed similarly. The Global English model was on par with or outperformed all the other models for each natural class except approximants.

5. Discussion

The above results demonstrated that broadening phone categories can be beneficial on limited data, though cross-language forced alignment using the large Global English model had consistently high performance. Considering cross-language configurations, the Global English model outperformed all others, likely due to its nearly 4000 hours of regionally diverse training data. In addition, the Explicit Small TIMIT model performed better than all remaining cross-language and all language-specific configurations. One potential explanation for this is that the training data for the Explicit Small model was more balanced in its regions of speakers. The higher performance of the Explicit models supports the findings of prior work on Yoloxóchitl Mixtec [12]. Overall, we recommend that language researchers adapt the Global English model, or a similar large pretrained model, to align their own target dataset.

One interesting finding was that long and nasal vowels showed a higher degree of inaccuracy in boundary placement, even for the best-performing Global English model. From a typological angle, long vowels are less common in the world's languages than short vowels, as are nasal vowels compared to oral vowels [25]. The counts of each vowel type in our Panãra data also reflect this imbalance. Despite the diversity of speakers and dialects for the Global English model, typological trends suggest that there were likely fewer instances of vowels being lengthened (which were in the inventory) or nasalized.

In addition, the phonological grammar of Panãra provides some insights into our findings, revealing how language-specific phonological processes make the task of alignment especially challenging. Across all systems, it was more difficult to place the onset boundary for fricatives, nasals, and stops. Fricatives are often distinct in their acoustic and spectral features, yet had low accuracy in our models. As discussed in [26], the glottal fricative [h] can be variably inserted in onsetless syllables, especially those that are prosodically prominent, such as in word-initial or stressed syllables. Under qualitative review of the data, we observed that word-initial [h], which was consistently present in the transcription, was pronounced at times and silent at others. Because the MFA must assign every phone in the given phone sequence a non-zero duration, deleted or invisible segments can throw off the performance.

Finally, the poor performance of our models on taps may be explained by the fact that our G2P system did not encode a relevant phonological rule. In particular, [19] described a process of excrescent vowels in complex onsets, such as in the word /krɤ/ → [kˠrɤ] 'thigh'. As a result, the MFA forced an alignment of /krɤ/ to an acoustic signal of the form [kVrɤ], thus encountering an additional, unexpected vowel.

6. Conclusion

This paper investigated the question: how does broadening phone categories of acoustic models affect the temporal degree of accuracy of phone-level forced alignment? We defined three granularity settings of phone categories as input to our acoustic models trained on either Panãra or English, and found mixed results on a Panãra test set: broadening phone categories can be helpful in language-specific training on very limited data, but cross-language alignment with a large Global English model outperformed other configurations. Can our findings for Panãra and English be replicated across other languages?

A main limitation of this work is that because the number of speakers in all of our Panãra data was only four, idiosyncrasies and sociolinguistic factors could strongly affect our results. The two speakers in the Panãra train set were younger than those in the test set, and age is correlated with higher proficiency in discourse skills such as rhythm and intonation. Also, while most Panãra speakers are monolingual, young males are the only group with conversational proficiency in Portuguese as a second language [19]. Although we balanced speaker gender in our data splits, speaker idiosyncrasies could have prevailed. Additionally, the Portuguese and miscellaneous phones present in our Panãra data may have added noise to these results.

One direction for future work is to further investigate the granularity settings of the lexicon. Distinctive features are binary properties in phonology that describe place and manner of articulation, as well as voicing and other properties [27]. As the SCA groupings of natural classes from [22] may not optimally reflect the similarity of sounds in the acoustic representation space, using distinctive features or an unsupervised clustering method to group sounds could aid in identifying more optimal groupings for broader sound classes. A multilingual aligner such as [28], trained on data with diverse phone inventories, could also be compared to these strategies.

7. References

[1] J. Shi, J. D. Amith, R. Castillo García, E. Guadalupe Sierra, K. Duh, and S. Watanabe, "Leveraging End-to-End ASR for Endangered Language Documentation: An Empirical Study on Yolóxochitl Mixtec," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds. Online: Association for Computational Linguistics, Apr. 2021, pp. 1134–1145. [Online]. Available: https://aclanthology.org/2021.eacl-main.96

[2] E. Salesky, E. Chodroff, T. Pimentel, M. Wiesner, R. Cotterell, A. W. Black, and J. Eisner, "A Corpus for Large-Scale Phonetic Typology," in *Association for Computational Linguistics*, 2020, pp. 4526–4546.

[3] E. P. Ahn and E. Chodroff, "VoxCommunis: A Corpus for Cross-linguistic Phonetic Analysis," in *Proceedings of the 13th Language Resources and Evaluation Conference*, 2022, pp. 5286–5294.

[4] R. Coto-Solano, J. N. Stanford, and S. K. Reddy, "Advances in Completely Automated Vowel Analysis for Sociophonetics: Using End-to-End Speech Recognition Systems with DARLA," *Frontiers in Artificial Intelligence*, vol. 4, 2021.

[5] W. Labov, I. Rosenfelder, and J. Fruehwald, "One Hundred Years of Sound Change in Philadelphia: Linear Incrementation, Reversal, and Reanalysis," *Language*, vol. 89, no. 1, pp. 30–65, 2013.

[6] I. Rosenfelder, J. Fruehwald, K. Evanini, and J. Yuan, "FAVE (Forced Alignment and Vowel Extraction) Program Suite," 2011.

[7] T. Kisler, U. Reichel, and F. Schiel, "Multilingual Processing of Speech via Web Services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[8] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. INTERSPEECH*, 2017, pp. 498–502.

[9] L. MacKenzie and D. Turton, "Assessing the Accuracy of Existing Forced Alignment Software on Varieties of British English," *Linguistics Vanguard*, vol. 6, no. s1, p. 20180061, Jan. 2020.

[10] C. Jones, W. Li, A. Almeida, and A. German, "Evaluating Cross-linguistic Forced Alignment of Conversational Data in North Australian Kriol, an Under-resourced Language," *Language Documentation & Conservation*, vol. 13, pp. 281–299, 2019. [Online]. Available: http://hdl.handle.net/10125/24869

[11] R. Coto-Solano, S. A. Nicholas, and S. Wray, "Development of Natural Language Processing Tools for Cook Islands Māori," in *Proceedings of the Australasian Language Technology Association Workshop 2018*, Dunedin, New Zealand, 2018, pp. 26–33. [Online]. Available: https://aclanthology.org/U18-1003

[12] C. DiCanio, H. Nam, D. H. Whalen, H. Timothy Bunnell, J. D. Amith, and R. C. García, "Using Automatic Alignment to Analyze Endangered Language Data: Testing the Viability of Untrained Alignment," *The Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 2235–2246, Sep. 2013.

[13] J.-P. Goldman, "EasyAlign: an Automatic Phonetic Alignment Tool under Praat," in *Proc. INTERSPEECH*, 2011.

[14] R. Coto-Solano and S. F. Solórzano, "Comparison of Two Forced Alignment Systems for Aligning Bribri Speech," *CLEI ELECTRONIC JOURNAL*, vol. 20, no. 1, 2017.

[15] C. Anderson, T. Tresoldi, S. J. Greenhill, R. Forkel, R. Gray, and J.-M. List, "Variation in Phoneme Inventories: Quantifying the Problem and Improving Comparability," *Journal of Language Evolution*, 11 2023. [Online]. Available: https://doi.org/10.1093/jole/lzad011

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

[17] S. P. Archive, "Hausa sound inventory (SPA)," in *PHOIBLE 2.0*, S. Moran and D. McCloy, Eds. Jena: Max Planck Institute for the Science of Human History, 2019. [Online]. Available: https://phoible.org/inventories/view/124

[18] S. Hoffmann and B. Pfister, "Text-to-Speech Alignment of Long Recordings using Universal Phone Models," in *Proc. INTERSPEECH 2013*, 2013, pp. 1520–1524.

[19] M. Lapierre, "The Phonology of Panãra: A Segmental Analysis," *International Journal of American Linguistics*, vol. 89, no. 2, pp. 183–218, 2023.

[20] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "DARPA TIMIT Acoustic-Phonetic Continous Speech Corpus CD-ROM. NIST Speech Disc 1-1.1," NASA/STI Recon, Technical Report 93, 1993.

[21] D. R. Mortensen, S. Dalmia, and P. Littell, "Epitran: Precision G2P for Many Languages," in *Proceedings of the 11th Language Resources and Evaluation Conference*, 2018, pp. 2710–2714.

[22] J.-M. List, "SCA: Phonetic Alignment Based on Sound Classes," in *New Directions in Logic, Language, and Computation*, M. Slavkovik and D. Lassiter, Eds. Springer, 2012, pp. 32–51.

[23] S. Fernández, A. Graves, and J. Schmidhuber, "Phoneme Recognition in TIMIT with BLSTM-CTC," *arXiv preprint arXiv:0804.3269*, 2008.

[24] M. McAuliffe and M. Sonderegger, "English MFA acoustic model v2.2.1," https://mfa-models.readthedocs.io/acoustic/English/EnglishMFAacousticmodelv2_2_1.html, Tech. Rep., May 2023.

[25] M. K. Gordon, *Phonological Typology*. Oxford University Press, 2016, vol. 1.

[26] M. Lapierre, "The Phonology of Panãra: A Prosodic Analysis," *International Journal of American Linguistics*, vol. 89, no. 3, pp. 333–356, 2023. [Online]. Available: https://doi.org/10.1086/724988

[27] R. Jakobson, C. G. Fant, and M. Halle, "Preliminaries to Speech Analysis: The Distinctive Features and their Correlates," 1951.

[28] J. Zhu, C. Yang, F. Samir, and J. Islam, "The Taste of IPA: Towards Open-Vocabulary Keyword Spotting and Forced Alignment in Any Language," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2024.