Comparing language-specific and cross-language acoustic models for low-resource phonetic forced alignment

Eleanor Chodroff, Emily P. Ahn, and Hossep Dolatian

May 8, 2024

Abstract

Phonetic forced alignment can greatly expedite spoken language analysis by providing automatic time alignments at the word- and phone-levels. In the case of low-resource languages, it remains an open question whether phone-level forced alignment will be more successful with a small language-specific acoustic model or a high-resource cross-language acoustic model. The present study directly compared language-specific and cross-language acoustic models in forced alignment performance using the Urum and Evenki datasets from the DoReCo Corpus. We evaluated six language-specific acoustic models trained with 5, 10, 15, 20, 25, or approximately 70 minutes of language-specific speech data against four English-based cross-language acoustic models that differed in size and accent homogeneity (large Global English or homogeneous American English of varying data amounts). Acoustic models were developed or obtained from the Montreal Forced Aligner, and evaluated against held-out manually aligned phone boundaries. Overall, the Global English model and the larger language-specific acoustic models were competitive with one another, and outperformed the homogeneous cross-language and smaller language-specific acoustic models. From this analysis, we recommend that researchers use a language-specific model with at least 25 minutes of speech (not just audio) or a large, diverse cross-language acoustic model for low-resource forced alignment.

1 Introduction

Access to diverse, multilingual spoken language data has risen considerably in recent years, in large part due to increased computational capacity. Diverse spoken language data has come from traditional fieldwork sources, compilations of prepared large-scale read speech recordings (e.g., Panayotov et al. 2015; Black 2019), as well as large-scale crowdsourced datasets, where speakers may contribute spoken data via a computer or smartphone (Ardila et al. 2019). Phonetic analysis of spoken language data can benefit substantially from a time alignment of the speech transcript to the audio file at the utterance-level and ideally also the word- and phone-levels. Time alignments facilitate subsequent language documentation at all levels of analysis, allowing researchers to efficiently locate segments of interest. Several methods currently exist for locating segments of interest within an audio file, including phonetic forced alignment, keyword spotting, and automatic speech recognition (ASR; Foley et al. 2018; Le Ferrand et al. 2020; Bird 2021; San et al. 2021; Coto-Solano 2022). Accurately labeling and aligning audio files from fieldwork, prepared, or crowdsourced datasets have tremendous implications for advances in linguistic research, community resource development, and the development of language-related tools (e.g., text and speech technology, language documentation, tools for language learning).

In the present study, we focus on phonetic forced alignment, in which an acoustic model of a language's phone set is used to identify the boundaries of each phone in its expected sequence based on the speech transcript. In contrast to manual segmental alignment, forced alignment is estimated to be anywhere from 200 to 400 times faster (Yuan et al. 2013; Young & McGarrah 2023). This technique has become increasingly important for phoneticians, where acoustic-phonetic analyses are conducted on spoken data collected in the field, in a sound booth, or using crowdsourced methodologies (Leemann et al. 2016; Stuart-Smith et al. 2019; Paschen et al. 2020a; Salesky et al. 2020; Ahn & Chodroff 2022; Hutin & Allassonnière-Tang 2022; Zhao & Chodroff 2022). Resulting analyses advance our empirical and theoretical understanding of phonetic realization across individual talkers, dialects, and languages.

Phonetic forced alignment is widely used for "high-resource" languages, where considerable data exists for training the necessary acoustic models (e.g., English). Moreover, several researchers have demonstrated the efficacy of "cross-language forced alignment", in which high-resource acoustic models such as English are used to align low-resource languages. In cases when available data is particularly low (e.g., <2 hours), as in many fieldwork situations, it is uncertain how language-specific acoustic models compare to higher-resource cross-language acoustic models in terms of performance. While cross-language models benefit from greater amounts of training data, language-specific models may benefit from more precise acoustic exposure. In the present study, we directly compared small language-specific acoustic models against cross-language acoustic models for phonetic forced alignment of low-resource languages. The small language-specific acoustic models represent a common scenario in speech resource development in a fieldwork setting: speech data is available but with limited amounts of transcribed data. The cross-language models include a pretrained Global English model (the english_mfa 2.0.0a model from the Montreal Forced Aligner (MFA): McAuliffe & Sonderegger 2022b) with over 3000 hours of training data, as well as acoustic

models developed using varying amounts of American English, representing a more homogeneous but high-resource source of speech data.

2 Background

2.1 Forced alignment systems

Phoneticians have most extensively relied on forced alignment systems such as the Montreal Forced Aligner (MFA; McAuliffe et al. 2017, Gentle (Ochshorn & Hawkins 2017), MAUS and WebMAUS (Schiel 1999; Kisler et al. 2017), the Prosodylab-Aligner (Gorman et al. 2011), FAVE (Rosenfelder et al. 2022), the Penn Forced Aligner (Yuan & Liberman 2008), the Language, Brain and Behaviour: Corpus Analysis Tool (LaBB-CAT) (Fromont & Hay 2012), or easyAlign (Goldman 2011). Several of these systems use a GMM-HMM architecture, where the expected sequence of phonetic segments (henceforth 'phones') can be modeled with a Hidden Markov Model (HMM), and their corresponding acoustic properties with a Gaussian Mixture Model (GMM). These systems commonly employ automatic speech recognition toolkits such as the Kaldi Speech Recognition Toolkit (Povey et al. 2011) for MFA and Gentle and the HTK Toolkit (Young et al. 2002) for Prosodylab-Aligner and FAVE to create a usable pretrained acoustic model for phonetic forced alignment .[1] Each system minimally offers the user an American English acoustic model, with many offering a slightly wider range of acoustic models for different languages.

Training an acoustic model has traditionally required computational power, speech data, and the know-how to train an acoustic model. For several years, the available acoustic models covered only a handful of the world's languages, and the alignment algorithm was frequently associated with a particular language. For instance, FAVE and the Penn Forced Aligner employed an acoustic model trained on American English from the SCOTUS Corpus (Yuan & Liberman 2008; Rosenfelder et al. 2022); the original version of the MFA relied on an acoustic model trained on American English from the Librispeech Corpus (Panayotov et al. 2015; McAuliffe et al. 2017), and MAUS was originally developed using spoken German data (Schiel 1999).

The availability of language-specific models has recently expanded as cross-linguistic speech corpora have become more readily accessible. Several language-specific acoustic models are now currently available through WebMAUS and the MFA, and have been trained on speech corpora such as GlobalPhone, Common Voice, among others (Strunk et al. 2014; McAuliffe et al. 2017; Kisler et al. 2017; Ahn & Chodroff 2022) Moreover, toolkits such as the MFA or the Prosodylab-Aligner have enabled speech researchers to train—with relative ease—acoustic models for phonetic forced alignment on their own language-specific data using their own computer. In a fieldwork setting, language-specific acoustic models have been developed for Matukar Panau using the MFA (Gonzalez et al. 2018; Barth et al. 2020), Eastern Chatino using the Prosodylab-aligner and eSpeak

[1] The HTK toolkit is no longer compatible with modern computer operating systems, which has resulted in a decline in the use of the Prosodylab-aligner, FAVE, and the Penn Forced Aligner.

(Ćavar et al. 2016), Mboshi using custom tools in STK (Synthesis ToolKit in C++; Cook & Scavone 1999; Mitkov 2014; Godard et al. 2018), and Tongan using the Prosodylab-Aligner (Johnson et al. 2018). Additional language-specific acoustic models have also been developed directly using the Kaldi Speech Recognition Toolkit (Dias et al. 2020) or using more recent end-to-end ASR systems (Biczysko 2022; Zhu et al. 2022).

2.2 Development of acoustic models: how much data is necessary?

McAuliffe (2021) conducted a comparison of training set sizes for forced alignment using the Buckeye Corpus of American English (Pitt et al. 2005). The Buckeye Corpus contains about 16.5 hours of true speech (pauses dropped) from 40 speakers, and has the convenient property of being manually aligned at the phone level. Training sets were varied in terms of number of speakers in the mixture, which also corresponded to the overall duration of the training set. When testing on the training data (not an uncommon situation when a researcher simply needs a best-fit alignment), word boundary errors, as measured by distance to annotated time point, were generally on target from even the smallest training sizes, but variance decreased as the training size increased. In examining phone boundary errors in selected CVC words, the error was variable with smaller training sizes, but was not far off from ceiling around the five to ten speaker mark (around three to five hours of speech). In terms of model generalization to new data, both word and phone boundary errors were fairly high until around 20 speakers (around eight to ten hours of speech).

When fitting directly to the data, it was thus recommended to have three to five hours of speech; for generalizing to similar data, it was recommended to have around eight to ten hours of speech. Though the analysis was quite thorough, the study only investigated the training and generalization on American English speech with limited investigation beyond the Buckeye Corpus. It remains to be seen whether such figures might also hold for other languages.

Critically, however, even three hours of training data may not even exist for some languages. In those cases, should we still attempt to train an acoustic model on the available data? Alternatively, would it be better to use cross-language forced alignment using an acoustic model trained on considerably more data from another language?

2.3 Cross-language forced alignment

An alternative to language-specific acoustic models is cross-language forced alignment. Given limited accessibility to training algorithms or just limited language-specific training data, it has been common to use pretrained acoustic models on other dialects and even language varieties. Language-specific phone labels can be mapped to labels in the high-resource pretrained phone set, and then back again following forced alignment. These studies have benefited from the large amounts of training data available for high resource languages like American English, French, or Italian, and have used these acoustic models as substitute acoustic models for a different language.

The majority of cross-language forced alignment scenarios have involved American English acoustic models. American English has been used to align other varieties of English, including British English (MacKenzie & Turton 2020), Australian English (Gonzalez et al. 2020), and several creoles with varying English influence (North Australian Kriol: Jones et al. 2019; Bequia Creole: Walker & Meyerhoff 2020). More notably, American English has also been used to align typologically distinct languages such as Mixtec (DiCanio et al. 2013), Nikyob (Kempton 2017), Bribri (Solórzano & Coto-Solano 2017), Cooks Islands Maori (Coto-Solano et al. 2018), Yidiny (Babinski et al. 2019), Matukar Panau (Barth et al. 2020), various Uralic languages (Leinonen et al. 2021), Nordic languages (Young & McGarrah 2023), and many others (e.g., Sim & Li 2008; Kurtić et al. 2012; Johnson et al. 2018).

In some cases, researchers may identify an alternative high resource language that could serve as a better substitute model for the language at hand. For instance, French acoustic models have been used to align Bribri (Solórzano & Coto-Solano 2017) and Italian acoustic models have been used to align North Australian Kriol (Jones et al. 2019). Bribri alignments generated from an American English-based system were generally more precise compared to the French-based system (Solórzano & Coto-Solano 2017). Though both systems employed an HTK architecture, the exact system architectures (English FAVE vs French EasyAlign), amount of training data, and choice of phone mappings also differed between the two languages, making it difficult to isolate the source of the performance difference. North Australian Kriol was also more accurately aligned with an Italian acoustic model compared to a multilingual acoustic model available through MAUS (Jones et al. 2019). Though North Australian Kriol is an English-based creole, Italian was chosen in this case given its similarly transparent orthography, its similar vowel system, and the similar spontaneous speech style to the data collected in fieldwork. Finally, Tang & Bennett (2019) pooled data from two related Mayan languages, Kaqchikel and Uspanteko, to increase sample size in a variation of cross-language phonetic alignment.

3 Methods

To evaluate the quality of forced alignment produced by language-specific or cross-language acoustic models, the present analysis tested acoustic models from the MFA on data available through DoReCO – Language Documentation Reference Corpora (Paschen et al. 2020b). While a wide range of forced aligners are available, we chose the MFA for a few reasons: first, it builds on a high-quality ASR toolkit, Kaldi, that enables custom acoustic model training; second, it is user-friendly, and third, it includes a very large, pretrained Global English acoustic model that we could use in our cross-language comparisons. Finally, the system performs consistently well against other forced alignment systems (Mahr et al. 2021).

For the data, we employed two of the larger language datasets available through DoReCo, Urum (Turkic; 93 min of speech) (Lorenz et al. 2022) and Evenki (Tungusic; 87 min of speech) (Kazakevich & Klyachko 2023), which both contained recordings of personal or traditional narratives provided in a spontaneous to semi-spontaneous manner. DoReCo proved a useful starting point

for our investigation given its online accessibility, representativeness for standard fieldwork data, and the availability of manually-aligned word boundaries which facilitated the manual phone-level alignments for the gold data.

3.1 Datasets

Urum is a Turkic language spoken in Georgia and southern Ukraine by approximately 171,000 speakers (Eberhard et al. 2023).[2] The full Urum dataset contained 93 minutes of speech from 32 speakers in 132 audio files, and was collected in Georgia around 2005 (Lorenz et al. 2022).[3] Based on the phonetic transcription provided with the corpus, the Urum phonetic inventory has 30 consonants and 9 vowels (Table 1).

<TABLE 1>

Evenki is a Tungusic language spoken in Russia, China, and Mongolia by approximately 16,830 speakers (Eberhard et al. 2023).[4] The full Evenki dataset contained 87 minutes of speech from 23 speakers in 36 audio files (Kazakevich & Klyachko 2023). Based on the phonetic transcription provided with the corpus, the Evenki phonetic inventory has 28 consonants and 13 vowels (Table 2).

<TABLE 2>

In the DoReCo release for each language, foreign material, false starts, filled pauses, unidentifiable material, and prolongations were marked with brackets, followed by an orthographic representation of the spoken segment. Foreign material for both Urum and Evenki was almost always Russian. If any transcription followed the bracketed annotation, it was given a romanized transliteration. We converted these segments to pronunciations using the same grapheme-to-phoneme system as the remainder of the corpus.

The conversion from orthography to pronunciation was done semi-automatically. In the case of prolongations with transcribed material, the automated pronunciation was sometimes incorrect as the transcriber frequently entered the word multiple times in a row, even if it was said only once. These cases accounted for less than 1% of the listed word types, and by token count, were even rarer. In Urum, digits were also given pronunciations by listening to the audio file and using the same phone set as was already in the corpus; these were always in Russian. Critically, while these pronunciations were used for the purposes of training the acoustic model, any bracketed segment or transcribed digit was excluded from the analysis of the test set.

The Urum test set had approximately eight minutes of analyzable speech from 11 speakers (22 minutes of audio). The original amount of test data was originally higher (around 14 minutes);

[2] `https://www.ethnologue.com/language/uum/`

[3] Speech duration was estimated from the sum of the manually aligned word boundaries provided in the DoReCo Corpus. Audio duration was simply the duration of the full audio files.

[4] `https://www.ethnologue.com/language/evn/`

however, bracketed but transcribed segments, such as foreign material and false starts, were ultimately excluded. The test set was manually aligned at the phone level by two trained linguists. The full Urum train set had 79 minutes of speech (including bracketed but transcribed segments) from 25 speakers (124 minutes of audio). Seven of the speakers overlapped between train and test sets. Recording quality was generally marked as "good" (188/246) or "medium" (53/246), with only a few instances of a "bad" (5/346), as indicated in the metadata from DoReCo. Background noise was also minimal (no reported background noise: 130/246, punctual background noise: 98/246, constant background noise: 18/246). The test data contained only "good" (19/26) or "medium" (7/26) sound qualities with a mixture of background noise (none: 10/26, punctual: 15/26, constant: 1/26).

The Evenki test set was created with nine minutes of speech from five speakers (26 minutes of audio). (This was originally 15 minutes including the bracketed but transcribed material.) This was manually aligned at the phone level by two trained linguists. The full Evenki train set had 72 minutes of speech from 18 speakers (153 minutes of audio). None of the speakers overlapped between train and test sets. Recording quality varied considerably from impressionistic markings of "good" (13/36) to "medium" (19/36) to "bad" (4/36), as indicated in the DoReCo metadata. Background noise was present in most recordings (no reported background noise: 9/36, punctual background noise: 26/36, constant background noise: 1/36). The test data contained only "medium" sound qualities with punctual background noise (5/5).

3.2 Preparing the MFA input

For the MFA input, we created a Praat TextGrid file for each wav file with the utterance-level alignment from the DoReCo release and the corresponding transcript. The pronunciation dictionary was obtained directly from the DoReCo data, which included a phonetic transcription for each word. Inspection of the data suggested that the original authors of both datasets used a rule-based grapheme-to-phoneme (G2P) system to convert the orthography to a phonetic transcription. Foreign words (always Russian from our assessment for both Urum and Evenki), false starts, and prolonged words were then assigned phonetic transcriptions using approximately the same G2P mapping for the primary language. In these cases, we found that the Russian transliteration followed a similar G2P mapping as the Urum and Evenki orthographies. Nevertheless, any word with this type of marking was excluded from evaluation: these words were thus included in the training material and present for alignment, but were not included in any of the evaluations. Within the test data, the median proportion of removed segments per utterance ("contamination" in the input utterance) was 0.33 for Urum and 0.20 for Evenki.

3.3 Acoustic models

To compare low-resource language-specific and cross-language forced alignment, we wanted a representative sample of acoustic models. A total of 30 language-specific acoustic models were

created that were trained on 5 to 25 minutes of true speech data in 5-minute increments with six subsets per duration level. The motivation for six subsets was somewhat arbitrary, but allowed for sufficiently different mixtures of the training data at each duration level, with three subsets containing high speaker variability with many speakers present (e.g., 5H1, 5H2, 5H3) and three subsets containing low speaker variability with few speakers present (e.g., 5L1, 5L2, 5L3). The selection of data was mostly created anew for each duration level to best match the targeted duration: for example, 5H1 was not necessarily a full subset of 10H1. The language-specific acoustic models were each trained using the default parameters in the MFA v2 train algorithm (version 2.0.6). To reduce the number of comparisons in evaluation, we selected the median-performing language-specific acoustic models from each duration level: 5, 10, 15, 20, or 25 minutes of training data. That is, among the six acoustic models per duration level (e.g., 5H1, 5H2, 5H3, 5L1, 5L2, 5L3), the chosen acoustic model was ranked third in performance, as assessed via median boundary difference (see Section 3.4 for more details). An additional acoustic model was trained using the full train set for each language, which we refer to as the 70+ minute model (82 minutes for Urum and 70 minutes for Evenki). In total, there were six language-specific acoustic models (one each for the duration levels of 5, 10, 15, 20, 25 or full).

Naturally, the amount of training data per tested phone differed across these acoustic models. The range and median number of tokens per phone can be found in Tables 3 and 4. For Urum, all 37 phones in the test set had a corresponding phone-specific model within each acoustic model ([lː] and [mː] did not appear in the test set). For Evenki, all but two of the 40 phones in the test set were present in the tested acoustic models: the 5-min model was missing a phone-specific model for [eː] and the 20-min model was missing a phone-specific model for [lʲ]. These phones, however, accounted for less than 0.025% of the test set.

Four cross-language acoustic models were evaluated in performance relative to the language-specific acoustic models. The cross-language models were a large Global English model (>3500 hours of Global English), a 100-hour American English model, a 10-hour American English model, and an American English model matched to the number of minutes in the full language-specific models (approximately 75 minutes of speech).

The Global English model was available via the MFA repository of pretrained acoustic models (english_mfa 2.0.0a) and was trained on approximately 3700 hours of English spoken around the world (Global English) (McAuliffe & Sonderegger 2022b). The training data for this dataset comes from Common Voice English v8.0 (2480 hours) (Ardila et al. 2019), LibriSpeech English (982 hours) (Panayotov et al. 2015), the Corpus of Regional African American Language v2021.07 (124 hours) (Farrington & Kendall 2021), Google Nigerian English (6 hours) (Butryna et al. 2020), Google UK and Ireland English (31 hours) (Demirsahin et al. 2020), NCHLT [South African] English (56 hours) (Barnard et al. 2014), and ARU [British] English (7 hours) (Hopkins et al. 2019). Given the dialect variability, the resulting acoustic model has a diverse phone set and representation of speech variation.

The three additional cross-language acoustic models were trained on American English speech from the LibriSpeech ASR Corpus (Panayotov et al. 2015), which contains read speech of audio-

books from mostly American English speakers. The pronunciation dictionary for these corpora was the Global English dictionary from the MFA (McAuliffe & Sonderegger 2022a). At the time of implementation, this was the default English dictionary with the MFA, and the only available English dictionary with IPA symbols.[5] The American English acoustic models were trained using the default parameters in the MFA v2 train algorithm (versions 2.1.7, 2.2.3, and 2.2.11).[6] The first American English dataset contained 100 hours of data from the LibriSpeech "clean" train dataset (Panayotov et al. 2015).[7] The second American English dataset contained 10 hours of speech from the same LibriSpeech "clean" train dataset. The third American English dataset was matched in the number of minutes to the full language-specific acoustic model (approximately 75 minutes). This dataset contained data from the LibriSpeech "clean" dev dataset and had the same number of speakers and gender breakdown as the original Urum or Evenki full dataset. These datasets served as a more homogeneous representation of English, and aimed to deconfound the amount of training data (number of hours or minutes) from the language input (American English vs language-specific). The 100-hour sample represents a large, but phonetically homogeneous cross-language scenario. The 10-hour sample represents the recommended amount of training data for language-specific alignment (McAuliffe 2021), but in a homogeneous cross-language scenario. Finally, the matched sample deconfounds the amount of data from the language input, allowing us to understand the importance of language specificity.

The language-specific pronunciation lexicon was obtained directly from the DoReCo phone transcriptions or from our application of grapheme-to-phoneme conversion. The cross-language pronunciation lexicon was obtained using the Interlingual MFA Toolkit (Dolatian 2023) by mapping the language-specific phones to English phones as specified in Tables 1 and 2. The symbols were then mapped back to their original symbols prior to evaluation.

3.4 Evaluation

The evaluation metrics were the percent data retention, the absolute difference between the aligned and gold boundary for phone onsets (MacKenzie & Turton 2020; Mahr et al. 2021), which we refer to as precision, and finally, the alignment accuracy (Mahr et al. 2021).

Data retention refers to how many segments were aligned; in many cases, phone and word intervals are entirely skipped by the alignment interval if the performance is too poor. The percent data retention was calculated based on the number of successfully merged aligned and gold phone intervals divided by the expected number of gold phone intervals. Data retention was assessed descriptively as a first indicator of potential issues with model performance.

[5] A few dialect-specific English acoustic models and dictionaries have since been made available. Though we could additionally test the influence of the precise phone specifications, we leave this analysis to future work. One advantage of the current approach is that the primary differences between the Global English and American English models is the audio data being used for training.

[6] The MFA v2 train algorithm did not change across these releases.

[7] "Clean" here indicates that the speech from these speakers was recognized with reasonably high accuracy by an automatic speech recognition system.

Precision refers to the difference between the gold and aligned boundary of the phone onset (see also Solórzano & Coto-Solano 2017; McAuliffe et al. 2017; Gonzalez et al. 2020; Mathad et al. 2021 for a similar use of phone onset difference). Precision was assessed descriptively in terms of median rank absolute boundary difference as well as the percent of tokens per model with a boundary within 20 ms of the gold boundary. A linear mixed-effects model was then used to assess variability in the token-level absolute boundary difference in log seconds based on the acoustic model (always compared to the Global English model), input utterance duration, proportion of bracketed segments in the input utterance ("contamination"), natural class of the preceding segment (vowel, approximant, nasal, fricative, stop, or silence), natural class of the target segment (vowel, approximant, nasal, fricative, or stop), as well as the interactions between natural class of the preceding and target segments. A random intercept was included for audio file. Prior to the log transformation, any tokens equal to 0 were converted to 0.001 s to avoid infinite values.

Alignment accuracy was defined as whether the forced aligned interval contains the midpoint of the gold interval (Knowles et al. 2018; Mahr et al. 2021). Accuracy was assessed descriptively in terms of rank, followed by a logistic mixed-effects model for each language. The model structure was the same as for the boundary difference, but without the interactions between the natural classes of the preceding and target segments. (Further interactions led to non-convergence.)

For the linear and logistic mixed-effects regressions, preceding and target natural class were sum-coded. For the six-level natural class of the preceding segment, the held-out level was silence, and for the five-level natural class of the target segment, the held-out level was stops. Acoustic model was always treatment-coded with each model compared to the Global English 3700 hr model (American English 100 hr, American English 10 hr, American English 70+ min, 70+ min, 25 min, 20 min, 15 min, 10 min, 5 min).

4 Results

4.1 Retention

As shown in Figure 1, retention was generally high across all acoustic models for both languages. No major patterns were observed between the cross-language and language-specific models. Retention was, however, consistently lower for the median-performing 5-minute models.

<FIGURE 1>

4.2 Precision

For both languages, the tested language-specific and cross-language acoustic models were fairly competitive with one another in terms of the median boundary difference and the percent of boundaries within 20 ms of the gold test boundary (Figure 2, Figure 3). The consistent exception was

the median-performing 5-min model, which had a noticeably higher and more variable boundary difference for both languages, along with a remarkably low percentage of boundaries within 20 ms of gold (16% for Urum and 14% for Evenki). In addition, precision noticeably worsened across the 15-, 10-, and 5-min models for Evenki, whereas the most noticeable performance drop was from the 10-min model to the 5-min model for Urum. Across both languages, the top three performing models with respect to both precision metrics included the Global English 3700 hr model, the full language-specific model, and the median-performing 25-min model. Ceiling performance as assessed by agreement within 20 ms was at 69% for Urum and 60% for Evenki.

The linear mixed-effects models for Urum and Evenki revealed several significant effects.[8] For Urum, only the full language-specific acoustic model was significantly more precise than the Global English acoustic model. Otherwise, the Global English model significantly outperformed the smaller language-specific models and the American English models. Longer input utterance durations also corresponded to significantly worse precision (Figure 4a). Higher contamination proportions per input utterance—that is, the proportion of bracketed segments due to foreign material, false starts, pauses, etc.—also significantly reduced precision. Several main effects and interactions of the targeted and preceding natural classes also reached significance. Significantly worse precision was observed for preceding vowels, approximants, and nasals, targeted approximants and nasals, as well as vowel–vowel, fricative–fricative, approximant–fricative, and nasal–approximant sequences. Significantly better precision was found for targeted fricatives, vowel–nasal, approximant–nasal, nasal–vowel, nasal–nasal, fricative–vowel, fricative–nasal, stop–vowel, and stop–nasal sequences.

For Evenki, the Global English model outperformed all language-specific models and American English models with respect to precision. As with Urum, longer utterance durations corresponded to significantly worse precision (Figure 4b), as did greater amounts of bracketed material within the input utterance. Several main effects and interactions of the targeted and preceding natural classes were also observed. Similar to Urum, preceding vowels, approximants, and nasals, as well as targeted approximants corresponded to significantly worse precision of the boundary. In contrast to Urum, targeted fricatives as well as preceding fricatives corresponded to significantly worse precision in Evenki. Significantly lower precision was also observed for vowel–vowel, nasal–nasal, vowel–nasal, and stop–nasal sequences. Significantly better precision was observed for targeted vowels, targeted nasals, as well as vowel–approximant, vowel–fricative, approximant–approximant, nasal–fricative, stop–vowel, and stop–fricative sequences.

<FIGURE 2>

<FIGURE 3>

<FIGURE 4>

[8] The full model results for Urum and Evenki precision can be found in the Appendix.

4.3 Accuracy

For both Urum and Evenki, accuracy of each acoustic model strongly mirrored the corresponding precision. Whereas precision corresponded to how far the gold-aligned boundary was from the force-aligned boundary, alignment accuracy reflected whether the force-aligned interval contained the midpoint of the gold interval, indicating a general accuracy of location for a given segment. The language-specific and cross-language acoustic models were overall similar in accuracy, with the primary exceptions of the 5-minute language-specific models, and for Evenki, the 10- and 15-minute models as well (Figure 5). The top three acoustic models always included the Global English model, the full language-specific model, and the median-performing 25-min model.

The pattern of significance observed in the logistic mixed-effects models for both Urum and Evenki was largely comparable to the pattern observed for precision. For Urum, the full language-specific model significantly outperformed the Global English model; otherwise, the Global English model significantly outperformed all other language-specific and cross-language models. Increased utterance durations and higher contamination proportions (bracketed material per utterance) also corresponded to significantly worse accuracy. Significantly worse accuracy was also associated with preceding vowels, approximants, and nasals, as well as targeted vowels, approximants, and nasals. Preceding and targeted fricatives, however, corresponded to significantly better accuracy.

For Evenki, the Global English model significantly outperformed all language-specific and cross-language models in terms of accuracy. As with Urum, longer utterance durations and higher contamination proportions corresponded to significantly decreased accuracy. Preceding vowels, approximants, nasals, as well as targeted approximants were also associated with decreased accuracy. In contrast to Urum, significantly worse accuracy was associated with preceding and targeted fricatives, whereas significantly better accuracy was associated with preceding stops and targeted vowels.

<FIGURE 5>

5 Discussion

Phonetic forced alignment of very low resource languages can present a challenge to researchers: is it worth training a language-specific acoustic model, or should a substitute cross-language acoustic model be used instead? The present study indicates that both paths may be viable, but with a few nuances. First, language-specific acoustic models can be successful with a small, but sufficient amount of data. Based on the results of the present study, the median-performing language-specific acoustic models with approximately 25 minutes of training data reached near-ceiling performance on the measures tested here. (Note that the 25 minutes refers to actual speech, as opposed to just recording duration.) A noticeable drop in performance was consistently observed with only five minutes of language-specific training data, and to a lesser degree with ten minutes of training data. The only cross-language acoustic model to consistently outperform the small language-specific

models was the large-scale Global English acoustic model. Otherwise, the more homogeneous American English models—even with 10 or 100 hours of training data—did not perform as well as a median-performing 25-min model in either of the languages tested. That said, the cross-language models were still competitive with the language-specific models. Indeed, the full language-specific model was always near- or at-ceiling in its performance against the other tested models.

Nevertheless, ceiling performance in the cross-language or low-resource forced alignment approach was not remarkably high. With respect to boundary agreement within 20 ms, the highest percentage was 69% for Urum from the full model with 70+ minutes (66% for the Global English model) and 60% for Evenki from the Global English model (52% for the full model with 70+ minutes). Overall, alignment of Urum was numerically better than that of Evenki, which we suspect might be attributable to the better sound quality in the training and test data. Urum otherwise had slightly longer input utterance durations and greater amounts of false starts or foreign material relative to Evenki. In any case, comparison with similar cross-language or low-resource forced alignment studies suggests this range of performances may be on par with other results, especially taking into consideration that the speech here was connected as opposed to isolated. In cross-language forced alignment, DiCanio et al. (2013) reported a 61% agreement within 20 ms for Yoloxóchitl Mixtec connected speech using an American English Nemours SRL aligner (Yarrington et al. 2008) and 52% using the Penn Phonetics Lab Forced Aligner also trained on American English (Yuan & Liberman 2008). In addition, Kempton (2017) found a 66% agreement within 20 ms for Nikyob isolated words using a Czech phone recognizer. In low-resource forced alignment, Johnson et al. (2018) achieved 63%–71% agreement within 20 ms for Tongan speech using an acoustic model trained on one hour of data using the ProsodyLab-Aligner (Gorman et al. 2011).

Overall, training a language-specific model with just 25 minutes of speech data was preferable to using a *homogeneous* cross-language model. Even though the American English models had considerably more training data than the language-specific ones, the homogeneity of the training data and its mismatch to the target language likely resulted in weaker alignment performance. It could be that American English specifically was a poor match for Urum and Evenki; further research should investigate how overlap between two languages' phoneme inventories and acoustics could influence results. Indeed, previous research has demonstrated worse alignment for phones without a one-to-one match in the cross-language setup (DiCanio et al. 2013; Babinski et al. 2019; Meer 2020). Moreover, researchers may benefit from using a large and diverse acoustic model, such as the Global English model used here for forced alignment. This model was always in the top two tested models, in both languages and for each tested measure.

Phonetic forced alignment can be further improved by minimizing the input utterance duration and appropriately handling the presence of foreign, non-speech, and other extraneous material. Ceiling performance was not always particularly high, indicating a need for alternative approaches to improve retention, precision, and accuracy. Reducing the input interval duration, removing foreign speech or non-speech intervals, and recursive forced alignment approaches would likely improve the overall performance of the forced alignment procedure. Previous studies have demonstrated that recursive forced alignment can substantially improve alignment quality by automatically or semi-automatically reducing lengthy input durations into shorter and shorter intervals (Moreno

et al. 1998; Gonzalez et al. 2018; Barth et al. 2020). The shorter aligned intervals are then recursively fed back into the forced alignment procedure for more precise alignments. Indeed, the present study found a significant influence of input utterance duration on the alignment quality in terms of precision for Evenki and accuracy for both Urum and Evenki.

Given the success of the large, multidialectal acoustic model tested here, these findings also reveal a need to test truly multilingual models of varying training data sizes for low-resource phonetic forced alignment. Multilingual and language-independent acoustic models show considerable promise but have had limited availability. Strunk et al. (2014) included a multilingual model in their release of WebMAUS, which has successfully been used to force align fieldwork recordings in the DoReCo Corpus (Paschen et al. 2020b). The acoustic models employed in the present study did not allow us to deconfound multilinguality (or the presence of diverse accents) from the amount of training data: the large Global English model was trained on an incredibly large amount of data (around 3700 hours) from quite varied accents. It also remains to be seen if a single language acoustic model with considerably more data, e.g., with thousands of hours, could offer a better alternative than a multilingual model of similar size, particularly if the language similarity is high.

In most fieldwork scenarios, however, large quantities of data are just not available. Relatedly, these recordings commonly involve conversations with code-switching, as was the case in the present study. Whether to use code-switched speech for acoustic model training is another important topic for future work. The present study did employ such instances in training, but excluded these for testing. Our speculation is that given the importance of data quantity, it would likely be helpful to include phone categories that overlap between the code-switched and target language, as it would further increase the sample size. An additional point of consideration is the degree to which the recording and speech styles match between the train and test sets, which was relatively high in the present study.

Finally, it will also be important to investigate how newer end-to-end systems such as wav2vec 2.0 (Baevski et al. 2020) or Whisper (Radford et al. 2023) perform against HMM systems that have an explicit representation of a phone. Some work in this direction has found that HMM systems largely outperform systems based on wav2vec 2.0 for word-level alignment (Biczysko 2022), as well as phone-level alignment (Zhu et al. 2022), though the end-to-end approach is still competitive. Further investigating the potential of these systems will be important avenues of research for speech processing and analysis of low-resource languages.

6 Recommendations & Conclusion

Our first recommendation for phonetic forced alignment of low-resource languages is to **use the large Global English acoustic model (or one that is similar) or a language-specific acoustic model with at least 25 minutes of speech**, assuming the MFA or a comparable procedure is used for acoustic model training. This also assumes that the language-specific training data matches the style and general recording environment of the test data. The large Global English model was

consistently competitive with the best-performing language-specific models of the durations tested here. This appears to be a highly reliable model for cross-language forced alignment. In addition, language-specific acoustic models with at least 25 minutes of speech data tended to perform near-ceiling for forced alignment. We must note, however, that we only tested the median-performing models from our initial pilot study, and it is still very possible that one gets unlucky with a particular 25-minute sample. Cross-language acoustic models trained on a relatively homogeneous dataset are unlikely to perform as well as an acoustic model trained directly on the language (assuming at least 25 minutes) or a large and highly diverse acoustic model.

Our second recommendation for phonetic forced alignment more generally is to ensure a short input utterance. **As a heuristic, we recommend an input utterance less than six seconds, and ideally less than two seconds.** In our data, precision generally worsened with increased duration, but as can be seen in Figure 4, alignments could be considerably further off after around 6 seconds. Reducing the input utterance duration can involve a trade-off between manual and automatic processing, but the longer the utterance, the more numerous the errors are that then need to be manually corrected after alignment. Sequences of abnormally low segment durations could also be indicative of a poor alignment; it may be beneficial to target these sequences during manual auditing. For low-resource alignment, recursive alignment to narrow the input utterance duration may also be worth the additional effort for increased precision and accuracy.

To conclude, phonetic forced alignment can substantially facilitate downstream analysis of a spoken language corpus. For languages with no pretrained acoustic model, it has been unclear whether to proceed in phone-level forced alignment using cross-language forced alignment or to train a language-specific acoustic model on the available language-specific data. Our findings indicate that several factors can influence this decision, and results may still vary; however, performance appears to reach ceiling with either a large and diverse multidialectal model (e.g., the large Global English acoustic model) or with a language-specific acoustic model with at least 25 minutes of speech (and not just audio) with comparable style and recording quality. Additional factors such as reducing the input utterance duration can further improve performance. This method was implemented using the MFA; however, a comparison between cross-language and language-specific forced alignment could be extended to any forced alignment system. Further comparisons with multilingual acoustic models or even larger language-specific models could also benefit our understanding of best practices in crosslinguistic forced alignment. As always, the output of automatic phonetic forced alignment should ideally be audited prior to any analysis.

7 Acknowledgments

We wish to thank Michael McAuliffe for his active maintenance of the Montreal Forced Aligner and for his help along the way. We also thank Matt Kelley, Gina-Anne Levow, Richard Wright, Yossi Keshet, Catalina Torres, and the Phonetics and Speech Sciences Group at the University of Zurich for helpful discussion and feedback. This work was supported by the Swiss National Science Foundation Grant PR00P1_208460 to EC.

8 Figures and Tables

Urum phones that stayed the same	Urum phones that were mapped to English phones
a, æ, b, c, d, dʒ, e, f, g, i, j, ɟ, k, l, ɬ, m, n, o, p, ɾ, s, ʃ, t, tʃ, u, v, z, ʒ	dː to d, ɣ to h, lː to l, mː to m, œ to o, r to ɹ, sː to s, tː to t, ɯ to u, x to h, y to u

Table 1: Phone mapping for Urum with the English model

Evenki phones that stayed the same	Evenki phones that were mapped to English phones
a, aː, b, d, e, ə, f, g, h, i, iː, j, k, l, m, n, ŋ, o, p, s, ʃ, t, tʃ, u, v, w, z, ʒ	tʲ to c, eː to e, əː to ə, ɣ to h, dʲ to ɟ, lʲ to l, nʲ to ɲ, oː to o, r to ɹ, sʲ to ʃ, ʃʲ to ʃ, uː to u, ʉ to u

Table 2: Phone mapping for Evenki with the English model

Model	Min	Median	Max
5 min	1	71	401
10 min	5	156	899
15 min	1	248	1465
20 min	1	248	1465
25 min	9	401	2187
70+ min	34	1100	6632

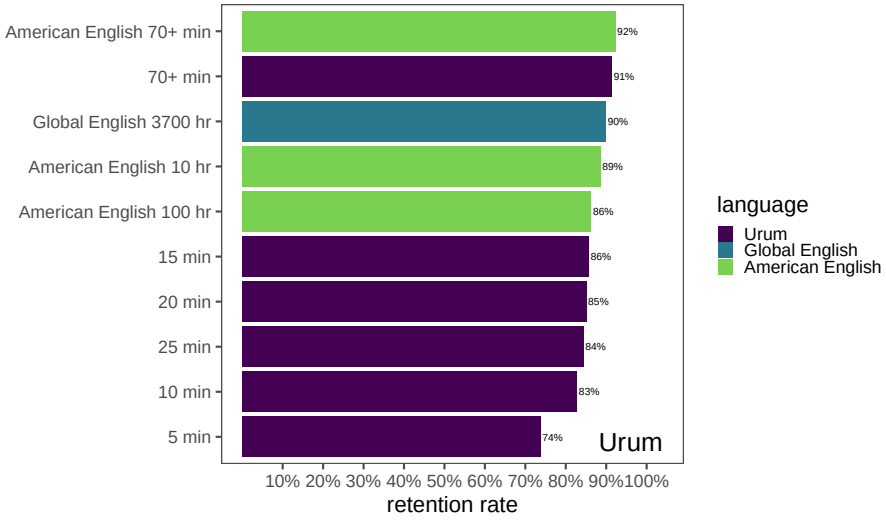
Table 3: Minimum, median, and maximum number of phone-specific training tokens for each Urum language-specific acoustic model. Though the 15-min and 20-min models have the same presented statistics, the distributions were indeed different

Model	Min	Median	Max
5 min	1	50	369
10 min	3	125	699
15 min	1	184	1161
20 min	3	219	1727
25 min	2	236	1494
70+ min	20	776.5	5144

Table 4: Minimum, median, and maximum number of phone-specific training tokens for each Evenki language-specific acoustic model

Figure 1: Segment retention as a percentage of the total number of gold segments for each acoustic model in a) Urum and b) Evenki.

(a) Urum



(b) Evenki

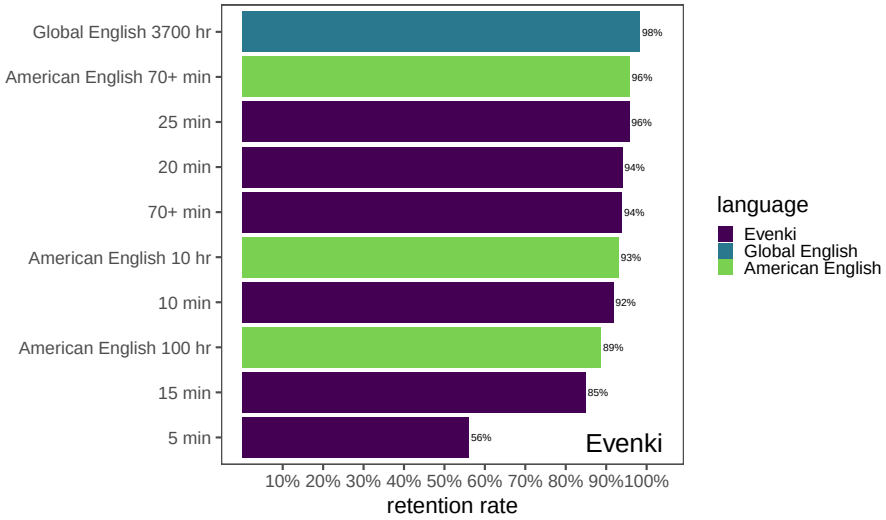
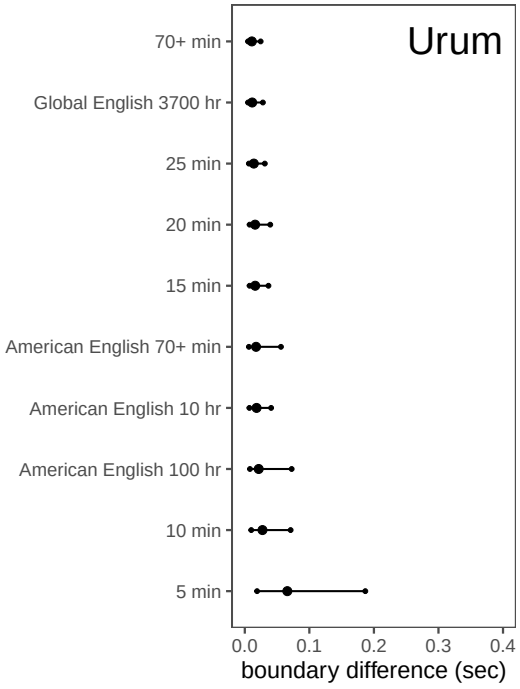


Figure 2: Median and interquartile range of boundary differences in seconds for each tested acoustic model in a) Urum and b) Evenki.

(a) Urum



(b) Evenki

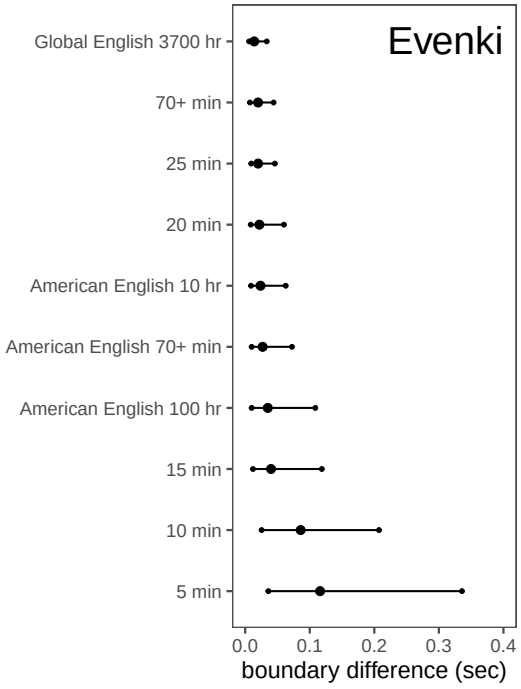
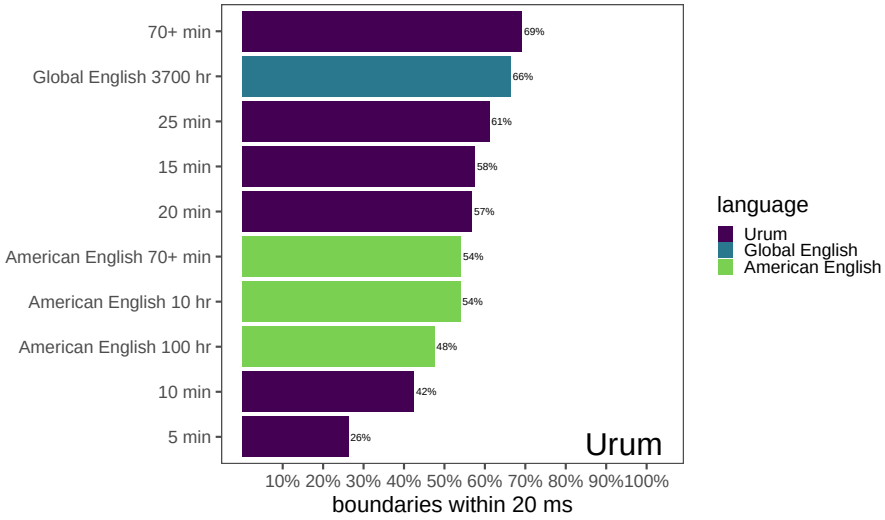


Figure 3: The percent of force-aligned test boundaries within 20 ms of the gold test boundary for each tested acoustic model in a) Urum and b) Evenki.

(a) Urum



(b) Evenki

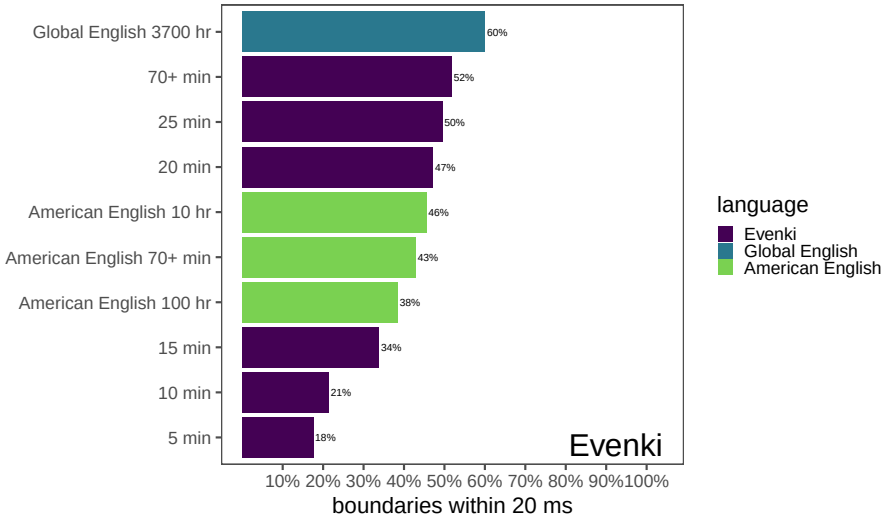
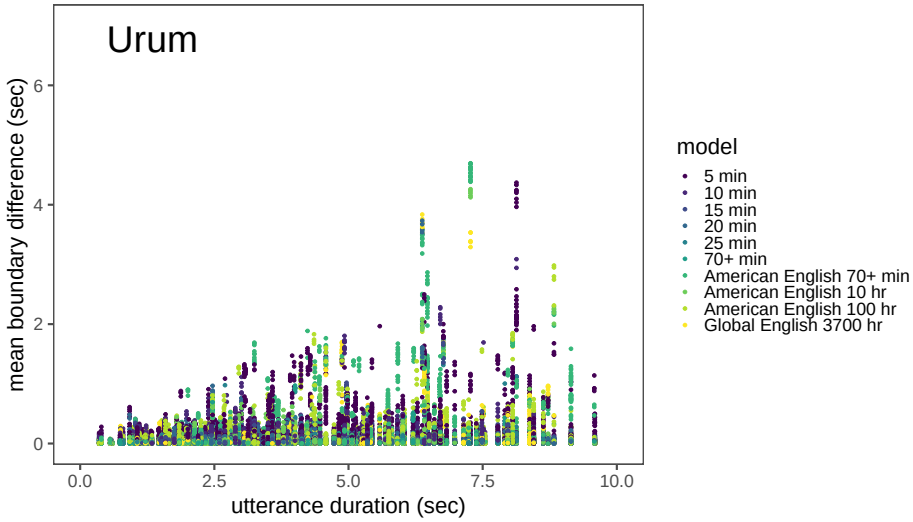


Figure 4: The relationship between precision (mean boundary difference in seconds) and input utterance duration in seconds across the tested acoustic models in a) Urum and b) Evenki. The x-axis has been truncated to 10 sec for better visibility.

(a) Urum



(b) Evenki

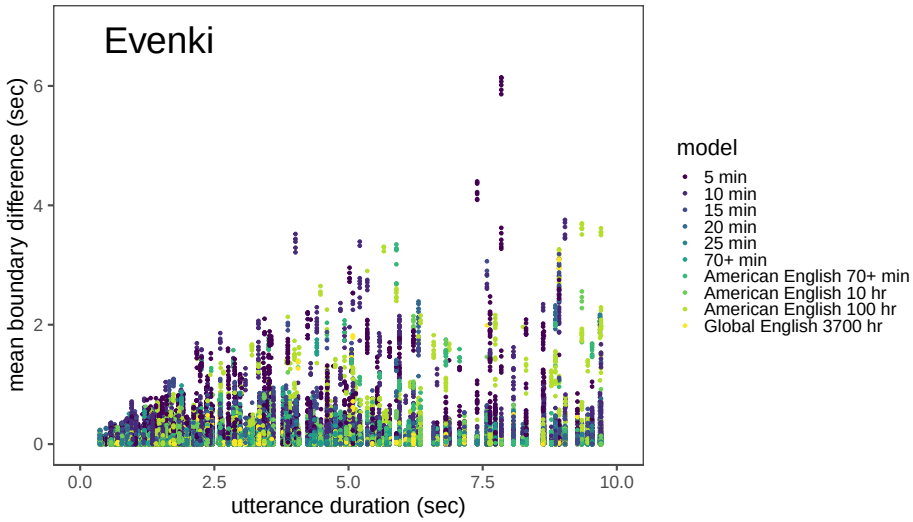
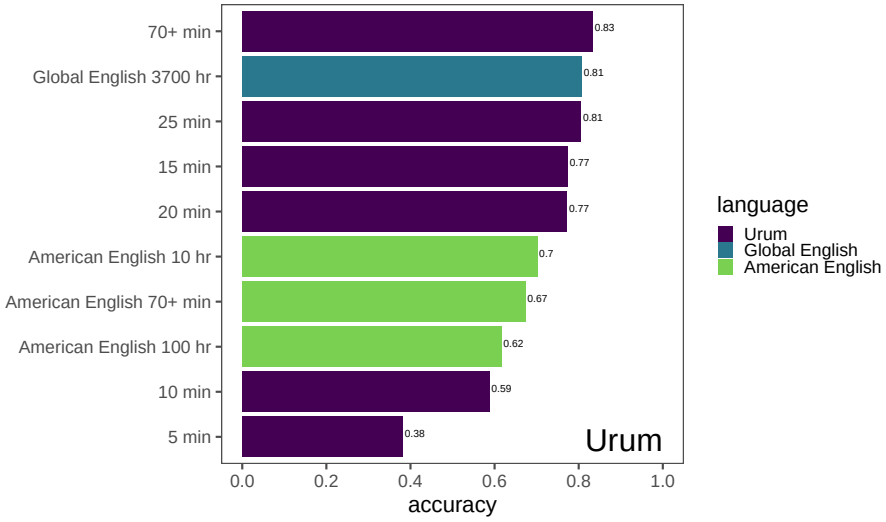
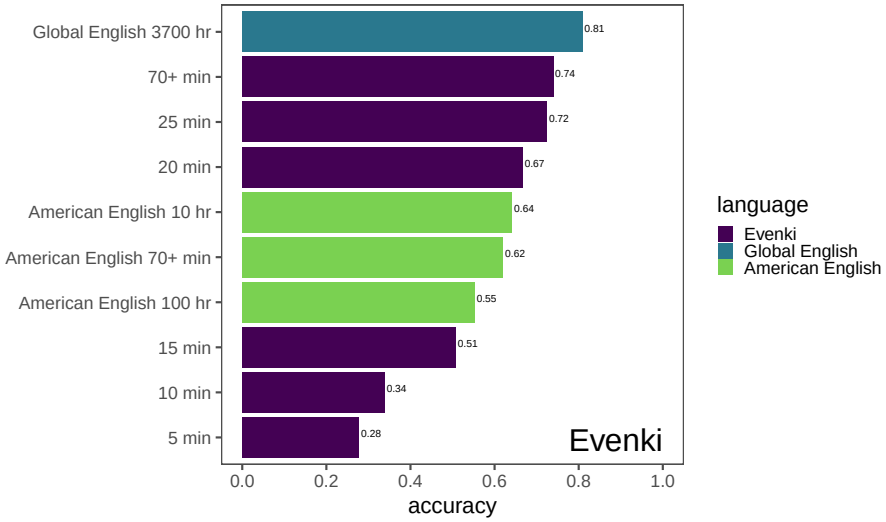


Figure 5: Accuracy in identifying the aligned segment location relative to the gold segment location for each acoustic model in a) Urum and b) Evenki.

(a) Urum



(b) Evenki



References

Ahn, Emily P. & Eleanor Chodroff. 2022. VoxCommunis: A corpus for cross-linguistic phonetic analysis. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, 5286–5294. Marseille, France: European Language Resources Association. https://aclanthology.org/2022.lrec-1.566.

Ardila, Rosana, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer,

Predictors	*Estimates*	*CI*	*p*
Intercept	-4.86	-5.00 – -4.72	**<0.001**
70+ min	-0.10	-0.16 – -0.04	**<0.001**
25 min	0.14	0.09 – 0.20	**<0.001**
20 min	0.30	0.24 – 0.36	**<0.001**
15 min	0.29	0.23 – 0.34	**<0.001**
10 min	0.74	0.68 – 0.80	**<0.001**
5 min	1.58	1.52 – 1.64	**<0.001**
American English 100 hr	0.68	0.62 – 0.73	**<0.001**
American English 10 hr	0.35	0.29 – 0.40	**<0.001**
American English 70+ min	0.55	0.49 – 0.60	**<0.001**
Utterance duration (hs)	2.12	1.56 – 2.68	**<0.001**
Contamination amount	0.51	0.42 – 0.59	**<0.001**
Preceding vowel	0.43	0.37 – 0.49	**<0.001**
Preceding approximant	0.54	0.43 – 0.64	**<0.001**
Preceding nasal	0.29	0.23 – 0.35	**<0.001**
Preceding fricative	0.04	-0.10 – 0.17	0.587
Preceding stop	0.09	-0.00 – 0.17	0.060
Vowel	-0.07	-0.16 – 0.02	0.116
Approximant	0.19	0.06 – 0.31	**0.003**
Nasal	0.09	0.01 – 0.18	**0.026**
Fricative	-0.23	-0.35 – -0.12	**<0.001**
Prec vowel × vowel	0.19	0.08 – 0.30	**0.001**
Prec vowel × approx	0.05	-0.09 – 0.20	0.460
Prec vowel × nasal	-0.11	-0.20 – -0.01	**0.023**
Prec vowel × fric	-0.05	-0.18 – 0.08	0.440
Prec approx × vowel	-0.02	-0.16 – 0.11	0.723
Prec approx × approx	-0.21	-0.47 – 0.06	0.126
Prec approx × nasal	-0.25	-0.39 – -0.11	**<0.001**
Prec approx × fric	0.29	0.02 – 0.55	**0.037**
Prec nasal × vowel	-0.14	-0.24 – -0.04	**0.005**
Prec nasal × approx	0.14	0.00 – 0.28	**0.042**
Prec nasal × nasal	-0.17	-0.26 – -0.08	**<0.001**
Prec nasal × fric	-0.04	-0.18 – 0.09	0.553
Prec fric × vowel	-0.57	-0.74 – -0.41	**<0.001**
Prec fric × nasal	-0.51	-0.68 – -0.35	**<0.001**
Prec fric × fric	1.03	0.65 – 1.42	**<0.001**
Prec stop × vowel	-0.40	-0.52 – -0.28	**<0.001**
Prec stop × approx	0.09	-0.11 – 0.29	0.365
Prec stop × nasal	-0.29	-0.41 – -0.17	**<0.001**
Prec stop × fric	-0.10	-0.29 – 0.10	0.339

Table 5: Linear mixed-effects model results for boundary difference (log seconds) in Urum. Each listed model is compared to the Global English 3700 hr model. Utterance duration was entered as hectoseconds (hs) for model convergence (seconds / 100). Preceding and targeted natural class is sum-coded, such that the listed level can be compared to the average. For preceding natural class, the held-out level is silence; for targeted natural class, the held-out level is stops.

Predictors	*Estimates*	*CI*	*p*
Intercept	-4.67	-4.88 – -4.47	**<0.001**
70+ min	0.22	0.17 – 0.28	**<0.001**
25 min	0.31	0.26 – 0.37	**<0.001**
20 min	0.43	0.38 – 0.49	**<0.001**
15 min	0.99	0.93 – 1.04	**<0.001**
10 min	1.63	1.57 – 1.69	**<0.001**
5 min	2.07	2.01 – 2.13	**<0.001**
American English 100 hr	0.91	0.85 – 0.96	**<0.001**
American English 10 hr	0.49	0.44 – 0.55	**<0.001**
American English 70+ min	0.61	0.55 – 0.66	**<0.001**
Utterance duration (hs)	5.72	5.07 – 6.37	**<0.001**
Contamination amount	0.01	-0.00 – 0.02	0.060
Preceding vowel	0.39	0.32 – 0.46	**<0.001**
Preceding approximant	0.60	0.47 – 0.72	**<0.001**
Preceding nasal	0.27	0.15 – 0.40	**<0.001**
Preceding fricative	0.35	0.14 – 0.56	**0.001**
Preceding stop	-0.04	-0.17 – 0.10	0.606
Vowel	-0.15	-0.23 – -0.07	**<0.001**
Approximant	0.73	0.52 – 0.93	**<0.001**
Nasal	-0.20	-0.31 – -0.09	**<0.001**
Fricative	0.21	0.11 – 0.32	**<0.001**
Prec vowel × vowel	0.39	0.27 – 0.52	**<0.001**
Prec vowel × approx	-0.61	-0.82 – -0.41	**<0.001**
Prec vowel × nasal	0.15	0.03 – 0.26	**0.012**
Prec vowel × fric	-0.27	-0.40 – -0.14	**<0.001**
Prec approx × vowel	0.04	-0.10 – 0.18	0.583
Prec approx × approx	-0.56	-0.81 – -0.31	**<0.001**
Prec approx × nasal	-0.11	-0.41 – 0.19	0.477
Prec approx × fric	0.14	-0.20 – 0.47	0.429
Prec nasal × vowel	-0.01	-0.14 – 0.13	0.930
Prec nasal × approx	-0.39	-0.81 – 0.02	0.064
Prec nasal × nasal	0.49	0.32 – 0.65	**<0.001**
Prec nasal × fric	-0.35	-0.58 – -0.12	**0.003**
Prec fric × vowel	-0.21	-0.42 – 0.01	0.064
Prec fric × nasal	-0.08	-0.50 – 0.35	0.730
Prec fric × fric	-0.15	-0.61 – 0.31	0.524
Prec stop × vowel	-0.30	-0.45 – -0.15	**<0.001**
Prec stop × approx	-0.39	-0.82 – 0.05	0.080
Prec stop × nasal	0.50	0.23 – 0.77	**<0.001**
Prec stop × fric	-0.85	-1.11 – -0.59	**<0.001**

Table 6: Linear mixed-effects model results for boundary difference (log seconds) in Evenki. Each listed model is compared to the Global English 3700 hr model. Utterance duration was entered as hectoseconds (hs) for model convergence (seconds / 100). Preceding and targeted natural class is sum-coded, such that the listed level can be compared to the average. For preceding natural class, the held-out level is silence; for targeted natural class, the held-out level is stops.

Predictors	*Log-Odds*	*CI*	*p*
Intercept	1.88	1.70 – 2.07	**<0.001**
70+ min	0.19	0.09 – 0.28	**<0.001**
25 min	-0.04	-0.13 – 0.06	0.469
20 min	-0.25	-0.34 – -0.15	**<0.001**
15 min	-0.24	-0.34 – -0.15	**<0.001**
10 min	-1.13	-1.22 – -1.04	**<0.001**
5 min	-2.05	-2.14 – -1.96	**<0.001**
American English 100 hr	-1.01	-1.10 – -0.92	**<0.001**
American English 10 hr	-0.61	-0.70 – -0.52	**<0.001**
American English 70+ min	-0.74	-0.83 – -0.65	**<0.001**
Utterance duration (hs)	-1.24	-2.08 – -0.41	**0.004**
Contamination amount	-0.49	-0.61 – -0.37	**<0.001**
Preceding vowel	-0.36	-0.45 – -0.28	**<0.001**
Preceding approximant	-0.16	-0.27 – -0.05	**0.003**
Preceding nasal	-0.13	-0.21 – -0.05	**0.002**
Preceding fricative	0.33	0.23 – 0.44	**<0.001**
Preceding stop	0.05	-0.04 – 0.14	0.308
Vowel	-0.08	-0.12 – -0.04	**<0.001**
Approximant	-0.83	-0.89 – -0.77	**<0.001**
Nasal	-0.24	-0.27 – -0.20	**<0.001**
Fricative	0.86	0.78 – 0.93	**<0.001**

Table 7: Logistic mixed-effects model results for accuracy in Urum. Accuracy was defined as 1 if the force aligned segments that contained the midpoint of the manually aligned segment. Each listed model is compared to the Global English 3700 hr model. Utterance duration was entered as hectoseconds (hs) for model convergence (seconds / 100). Preceding and targeted natural class is sum-coded, such that the listed level can be compared to the average. For preceding natural class, the held-out level is silence; for targeted natural class, the held-out level is stops.

Predictors	*Log-Odds*	*CI*	*p*
Intercept	1.66	1.50 – 1.82	**<0.001**
70+ min	-0.43	-0.52 – -0.34	**<0.001**
25 min	-0.50	-0.59 – -0.42	**<0.001**
20 min	-0.79	-0.87 – -0.71	**<0.001**
15 min	-1.51	-1.59 – -1.42	**<0.001**
10 min	-2.23	-2.31 – -2.14	**<0.001**
5 min	-2.58	-2.67 – -2.48	**<0.001**
American English 100 hr	-1.30	-1.38 – -1.22	**<0.001**
American English 10 hr	-0.90	-0.99 – -0.82	**<0.001**
American English 70+ min	-1.00	-1.08 – -0.92	**<0.001**
Utterance duration (hs)	-1.69	-2.42 – -0.95	**<0.001**
Contamination amount	-0.21	-0.31 – -0.11	**<0.001**
Preceding vowel	-0.25	-0.32 – -0.18	**<0.001**
Preceding approximant	-0.62	-0.70 – -0.54	**<0.001**
Preceding nasal	-0.37	-0.45 – -0.29	**<0.001**
Preceding fricative	-0.28	-0.39 – -0.17	**<0.001**
Preceding stop	0.12	0.05 – 0.20	**0.002**
Vowel	0.22	0.17 – 0.26	**<0.001**
Approximant	-0.62	-0.67 – -0.57	**<0.001**
Nasal	-0.07	-0.11 – -0.03	**0.001**
Fricative	-0.18	-0.24 – -0.11	**<0.001**

Table 8: Logistic mixed-effects model results for accuracy in Evenki. Accuracy was defined as 1 if the force aligned segments that contained the midpoint of the manually aligned segment. Each listed model is compared to the Global English 3700 hr model. Utterance duration was entered as hectoseconds (hs) for model convergence (seconds / 100). Preceding and targeted natural class is sum-coded, such that the listed level can be compared to the average. For preceding natural class, the held-out level is silence; for targeted natural class, the held-out level is stops.

Reuben Morais, Lindsay Saunders, Francis M. Tyers & Gregor Weber. 2019. Common Voice: A massively-multilingual speech corpus. In Proceedings of the Twelfth Language Resources and Evaluation Conference, 4218–4222. Marseille, France: European Language Resources Association.

Babinski, Sarah, Rikker Dockum, J. Hunter Craft, Anelisa Fergus, Dolly Goldenberg & Claire Bowern. 2019. A Robin Hood approach to forced alignment: English-trained algorithms and their use on Australian languages. Proceedings of the Linguistic Society of America 4(1). 3. http://dx.doi.org/10.3765/plsa.v4i1.4468.

Baevski, Alexei, Henry Zhou, Abdelrahman Mohamed & Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems, 12449–12460.

Barnard, E, MH Davel, C Van Heerden, Febe De Wet & J Badenhorst. 2014. The NCHLT speech corpus of the South African languages. In Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU), http://hdl.handle.net/10204/7549.

Barth, Danielle, James Grama, Simon Gonzalez & Catherine Travis. 2020. Using forced alignment for sociophonetic research on a minority language. In University of Pennsylvania Working Papers in Linguistics, vol. 25, 2.

Biczysko, Klaudia. 2022. Automatic annotation of speech: Exploring boundaries within forced alignment for Swedish and Norwegian.

Bird, Steven. 2021. Sparse transcription. Computational Linguistics 46(4). 713–744. http://dx.doi.org/10.1162/coli$_{a0}$0387. https://direct.mit.edu/coli/article/46/4/713/97329/Sparse-Transcription.

Black, Alan W. 2019. CMU Wilderness Multilingual Speech Dataset. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5971–5975.

Butryna, Alena, Shan-Hui Cathy Chu, Isin Demirsahin, Alexander Gutkin, Linne Ha, Fei He, Martin Jansche, Cibu Johny, Anna Katanova, Oddur Kjartansson, Chenfang Li, Tatiana Merkulova, Yin May Oo, Knot Pipatsrisawat, Clara Rivera, Supheakmungkol Sarin, Pasindu de Silva, Keshan Sodimana, Richard Sproat, Theeraphol Wattanavekin & Jaka Aris Eko Wibawa. 2020. Google Crowdsourced Speech Corpora and Related Open-Source Resources for Low-Resource Languages and Dialects: An Overview. In 2019 UNESCO International Conference Language Technologies for All (LT4All): Enabling Linguistic Diversity and Multilingualism Worldwide, 91–94. http://dx.doi.org/10.48550/ARXIV.2010.06778.

Cook, Perry R & Gary P Scavone. 1999. The synthesis toolkit (STK). In Proceedings of the international computer music conference, Beijing, China.

Coto-Solano, Rolando. 2022. Computational sociophonetics using automatic speech recognition. Language and Linguistics Compass 16(9). e12474. http://dx.doi.org/10.1111/lnc3.12474.

Coto-Solano, Rolando, Sally Akevai Nicholas & Samantha Wray. 2018. Development of natural language processing tools for Cook Islands Māori. In Proceedings of the Australasian Language Technology Association Workshop 2018, 26–33. Dunedin, New Zealand. https://aclanthology.org/U18-1003.

Demirsahin, Isin, Oddur Kjartansson, Alexander Gutkin & Clara Rivera. 2020. Open-source multi-speaker corpora of the English accents in the British Isles. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, 6532–6541. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.804.

Dias, Ana Larissa, Cassio Batista, Daniel Santana & Nelson Neto. 2020. Towards a free, forced phonetic aligner for Brazilian Portuguese using Kaldi tools. In Ricardo Cerri & Ronaldo C. Prati (eds.), Intelligent Systems, vol. 12319, 621–635. Cham: Springer International Publishing. http://dx.doi.org/10.1007/978-3-030-61377-8$_4$4. http://link.springer.com/10.1007/978-3-030-61377-8_44.

DiCanio, Christian, Hosung Nam, Douglas H. Whalen, H. Timothy Bunnell, Jonathan D. Amith & 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. http://dx.doi.org/10.1121/1.4816491.

Dolatian, Hossep. 2023. Interlingual MFA. https://github.com/jhdeov/interlingual-MFA.

Eberhard, David M., Gary F. Simons & Charles D. Fennig. 2023. Ethnologue: Languages of the world. Dallas, Texas: SIL International 26th edn. http://www.ethnologue.com.

Farrington, Charlie & Tyler Kendall. 2021. The Corpus of Regional African American Language http://dx.doi.org/10.7264/1AD5-6T35. https://oraal.uoregon.edu/coraal.

Foley, Ben, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath, František Kratochvíl, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger & Janet Wiles. 2018. Building speech recognition systems for language documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). In Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2018), 205–209.

Fromont, Robert & Jennifer Hay. 2012. LaBB-CAT: an annotation store. In Paul Cook & Scott Nowson (eds.), Proceedings of the australasian language technology association workshop 2012, 113–117. Dunedin, New Zealand. https://aclanthology.org/U12-1015.

Godard, Pierre, Gilles Adda, Martine Adda-Decker, Juan Benjumea, Laurent Besacier, Jamison Cooper-Leavitt, Guy-Noel Kouarata, Lori Lamel, Hélène Maynard, Markus Mueller, Annie Rialland, Sebastian Stueker, François Yvon & Marcely Zanon-Boito. 2018. A very low resource language speech corpus for computational language documentation experiments. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan: European Language Resources Association (ELRA). https://aclanthology.org/L18-1531.

Goldman. 2011. EasyAlign: an automatic phonetic alignment tool under Praat. In Proceedings of InterSpeech, Firenze, Italy.

Gonzalez, Simon, James Grama & Catherine E. Travis. 2020. Comparing the performance of forced aligners used in sociophonetic research. Linguistics Vanguard 6(1). 20190058. http://dx.doi.org/10.1515/lingvan-2019-0058.

Gonzalez, Simon, Catherine Travis, James Grama, Danielle Barth & Sunkulp Ananthanarayan. 2018. Recursive forced alignment: A test on a minority language. In Proceedings of the 17th Australasian International Conference on Speech Science and Technology, 145–148.

Gorman, Kyle, Jonathan Howell & Michael Wagner. 2011. Prosodylab-aligner: A tool for forced alignment of laboratory speech. Canadian Acoustics 39(3). 192–193. https://jcaa.caa-aca.ca/index.php/jcaa/article/view/2476.

Hopkins, Carl, Simone Graetzer & Gary Seiffert. 2019. ARU speech corpus (University of Liverpool). http://dx.doi.org/10.17638/DATACAT.LIVERPOOL.AC.UK/681. http://datacat.liverpool.ac.uk/id/eprint/681.

Hutin, Mathilde & Marc Allassonnière-Tang. 2022. Operation LiLi: Using crowd-sourced data and automatic alignment to investigate the phonetics and phonology of less-resourced languages. Languages 7(3). 234. http://dx.doi.org/10.3390/languages7030234.

Johnson, Lisa M., Marianna Di Paolo & Adrian Bell. 2018. Forced alignment for understudied language varieties: Testing Prosodylab-Aligner with Tongan data. Language Documentation & Conservation 12. 80–123. http://hdl.handle.net/10125/24763.

Jones, Caroline, Weicong Li, Andre Almeida & Amit German. 2019. Evaluating cross-linguistic forced alignment of conversational data in north Australian Kriol, an under-resourced language. Language Documentation & Conservation 13. 281–299. http://hdl.handle.net/10125/24869.

Kazakevich, Olga & Elena Klyachko. 2023. Evenki DoReCo dataset. http://dx.doi.org/10.34847/NKL.5E0D27CU.

Kempton, Timothy. 2017. Cross-language forced alignment to assist community-based linguistics for low resource languages. In Proceedings of the 2nd Workshop on the Use of

Computational Methods in the Study of Endangered Languages, 165–169. Honolulu: Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/W17-0122. http://aclweb.org/anthology/W17-0122.

Kisler, Thomas, Uwe Reichel & Florian Schiel. 2017. Multilingual processing of speech via web services. Computer Speech & Language 45. 326–347. http://dx.doi.org/10.1016/j.csl.2017.01.005.

Knowles, Thea, Meghan Clayards & Morgan Sonderegger. 2018. Examining factors influencing the viability of automatic acoustic analysis of child speech. Journal of Speech, Language, and Hearing Research 61(10). 2487–2501. http://dx.doi.org/10.1044/2018_JSLHR-S-17-0275. http://pubs.asha.org/doi/10.1044/2018_JSLHR-S-17-0275.

Kurtić, Emina, Bill Wells, Guy J. Brown, Timothy Kempton & Ahmet Aker. 2012. A corpus of spontaneous multi-party conversation in Bosnian Serbo-Croatian and British English. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), 1323–1327. Istanbul, Turkey: European Language Resources Association (ELRA).

Le Ferrand, Eric, Steven Bird & Laurent Besacier. 2020. Enabling interactive transcription in an indigenous community. In Proceedings of the 28th International Conference on Computational Linguistics, 3422–3428. Barcelona, Spain (Online): International Committee on Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.coling-main.303. https://www.aclweb.org/anthology/2020.coling-main.303.

Leemann, Adrian, Marie-José Kolly, Ross Purves, David Britain & Elvira Glaser. 2016. Crowdsourcing language change with smartphone applications. PLOS ONE 11(1). e0143060. http://dx.doi.org/10.1371/journal.pone.0143060. https://dx.plos.org/10.1371/journal.pone.0143060.

Leinonen, Juho, Sami Virpioja & Mikko Kurimo. 2021. Grapheme-based cross-language forced alignment: Results with Uralic languages. In Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), 345–350,. Reykjavik, Iceland: Linköping University Electronic Press, Sweden. https://aclanthology.org/2021.nodalida-main.36.

Lorenz, Johanna, Violeta Moisidi, Stefanie Schröter, Stavros Skopeteas & Nutsa Tsetereli. 2022. Urum DoReCo dataset. http://dx.doi.org/10.34847/NKL.AC166N10.

MacKenzie, Laurel & Danielle Turton. 2020. Assessing the accuracy of existing forced alignment software on varieties of British English. Linguistics Vanguard 6(s1). 20180061. http://dx.doi.org/10.1515/lingvan-2018-0061.

Mahr, Tristan J., Visar Berisha, Kan Kawabata, Julie Liss & Katherine C. Hustad. 2021. Performance of forced-alignment algorithms on children's speech. Journal of Speech, Language, and Hearing Research 64(6S). 2213–2222. http://dx.doi.org/10.1044/2020_JSLHR-20-00268.

Mathad, Vikram C., Tristan J. Mahr, Nancy Scherer, Kathy Chapman, Katherine C. Hustad, Julie Liss & Visar Berisha. 2021. The impact of forced-alignment errors on automatic pronunciation evaluation. In Interspeech 2021, 1922–1926. ISCA. http://dx.doi.org/10.21437/Interspeech.2021-1403.

McAuliffe, Michael. 2021. How much data do you need for a good MFA alignment? Tech. rep. https://memcauliffe.com/how-much-data-do-you-need-for-a-good-mfa-alignment.html.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner & Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In Interspeech 2017, 498–502. ISCA. http://dx.doi.org/10.21437/Interspeech.2017-1386.

McAuliffe, Michael & Sonderegger. 2022a. English MFA acoustic model v2.0.0. https://mfa-models.readthedocs.io/pronunciationdictionary/English/EnglishMFAdictionaryv2_0_0.html.

McAuliffe, Michael & Sonderegger. 2022b. English MFA acoustic model v2.0.0a. https://mfa-models.readthedocs.io/acoustic/English/EnglishMFAacousticmodelv2_0_0a.html.

Meer, Philipp. 2020. Automatic alignment for New Englishes: Applying state-of-the-art aligners to Trinidadian English. The Journal of the Acoustical Society of America 147(4). 2283–2294.

Mitkov, Ruslan (ed.). 2014. The Oxford Handbook of Computational Linguistics. Oxford University Press 2nd edn. http://dx.doi.org/10.1093/oxfordhb/9780199573691.001.0001. https://academic.oup.com/edited-volume/42643.

Moreno, Pedro J, Christopher F Joerg, Jean-Manuel Van Thong & Oren Glickman. 1998. A recursive algorithm for the forced alignment of very long audio segments. In International Conference on Spoken Language Processing (ICSLP), vol. 98, 2711–2714.

Ochshorn, RM & Max Hawkins. 2017. Gentle forced aligner. github.com/lowerquality/gentle.

Panayotov, Vassil, Guoguo Chen, Daniel Povey & Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5206–5210. South Brisbane, Queensland, Australia: IEEE. http://dx.doi.org/10.1109/ICASSP.2015.7178964.

Paschen, Ludger, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave & Frank Seifart. 2020a. Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). In Proceedings of the Twelfth Language Resources and Evaluation Conference, 2657–2666. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.324.

Paschen, Ludger, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave & Frank Seifart. 2020b. Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). In Proceedings of the Twelfth Language Resources and Evaluation Conference, 2657–2666. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.324.

Pitt, Mark A, Keith Johnson, Elizabeth Hume, Scott Kiesling & William Raymond. 2005. The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. Speech Communication 45(1). 89–95.

Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer & Karel Veselý. 2011. The Kaldi speech recognition toolkit. In IEEE 2011 workshop on Automatic Speech Recognition and Understanding, .

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey & Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato & Jonathan Scarlett (eds.), Proceedings of the 40th international conference on machine learning, vol. 202 Proceedings of machine learning research, 28492–28518. PMLR. https://proceedings.mlr.press/v202/radford23a.html.

Rosenfelder, Ingrid, Josef Fruehwald, Keelan Evanini, Scott Seyfarth, Christian Brickhouse, Kyle Gorman, Hillary Prichard & Jiahong Yuan. 2022. FAVE: Forced alignment and vowel extraction. Zenodo. http://dx.doi.org/10.5281/ZENODO.593309. https://zenodo.org/record/593309.

Salesky, Elizabeth, Eleanor Chodroff, Tiago Pimentel, Matthew Wiesner, Ryan Cotterell, Alan W Black & Jason Eisner. 2020. A corpus for large-scale phonetic typology. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 4526–4546. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.acl-main.415.

San, Nay, Martijn Bartelds, Mitchell Browne, Lily Clifford, Fiona Gibson, John Mansfield, David Nash, Jane Simpson, Myfany Turpin, Maria Vollmer, Sasha Wilmoth & Dan Jurafsky. 2021. Leveraging pre-trained representations to improve access to untranscribed speech from endangered languages. In 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 1094–1101. Cartagena, Colombia: IEEE. http://dx.doi.org/10.1109/ASRU51503.2021.9688301.

Schiel, Florian. 1999. Automatic phonetic transcription of non-prompted speech. In Proceedings of the 14th International Congress on Phonetic Sciences (ICPhS), 607–610.

Sim, Khe Chai & Haizhou Li. 2008. Robust phone set mapping using decision tree clustering for cross-lingual phone recognition. In 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, 4309–4312. Las Vegas, NV, USA: IEEE. http://dx.doi.org/10.1109/ICASSP.2008.4518608.

Solórzano, Sofía Flores & Rolando Coto-Solano. 2017. Comparison of two forced alignments systems for aligning Bribri speech. CLEI Electronic Journal http://dx.doi.org/10.19153/cleiej.20.1.2.

Strunk, Jan, Florian Schiel & Frank Seifart. 2014. Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 3940–3947. Reykjavik, Iceland: European Language Resources Association (ELRA). `http://www.lrec-conf.org/proceedings/lrec2014/pdf/1176_Paper.pdf`.

Stuart-Smith, Jane, Morgan Sonderegger, Rachel Macdonald, Jeff Mielke, Michael McAuliffe & Eric Thomas. 2019. Large-scale acoustic analysis of dialectal and social factors in English /s/-retraction. In Proceedings of the International Congress of Phonetic Sciences 2019, .

Tang, Kevin & Ryan Bennett. 2019. Unite and conquer: Bootstrapping forced alignment tools for closely-related minority languages (Mayan). In Proceedings of the 19th International Congress of Phonetic Sciences, 1719–1723. Melbourne, Australia.

Walker, James A. & Miriam Meyerhoff. 2020. Pivots of the Caribbean? Low-back vowels in eastern Caribbean English. Linguistics 58(1). 109–130. http://dx.doi.org/10.1515/ling-2019-0037.

Yarrington, Debra, John Gray, Chris Pennington, H. Timothy Bunnell, Allegra Cornaglia, Jason Lilley, Kyoko Nagao & James Polikoff. 2008. ModelTalker Voice Recorder—An interface system for recording a corpus of speech for synthesis. In Jimmy Lin (ed.), Proceedings of the ACL-08: HLT demo session, 28–31. Columbus, Ohio: Association for Computational Linguistics. `https://aclanthology.org/P08-4008`.

Young, Nathan J. & Michael McGarrah. 2023. Forced alignment for Nordic languages: Rapidly constructing a high-quality prototype. Nordic Journal of Linguistics 46(1). 105–131. http://dx.doi.org/10.1017/S033258652100024X.

Young, Steve, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying (Andrew) Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev & Phil Woodland. 2002. The HTK book. Cambridge University Engineering Department.

Yuan, Jiahong & Mark Liberman. 2008. Speaker identification on the SCOTUS corpus. In Proceedings of Acoustics, 5687–5690.

Yuan, Jiahong, Neville Ryant, Mark Liberman, Andreas Stolcke, Vikramjit Mitra & Wen Wang. 2013. Automatic phonetic segmentation using boundary models. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2306–2310.

Zhao, Liang & Eleanor Chodroff. 2022. The ManDi Corpus: A spoken corpus of Mandarin regional dialects. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, 1985–1990. Marseille, France: European Language Resources Association.

Zhu, Jian, Cong Zhang & David Jurgens. 2022. Phone-to-audio alignment without text: A semi-supervised approach. In ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 8167–8171. Singapore, Singapore: IEEE. http://dx.doi.org/10.1109/ICASSP43922.2022.9746112. https://ieeexplore.ieee.org/document/9746112/.

Ćavar, Malgorzata, Damir Ćavar & Hilaria Cruz. 2016. Endangered language documentation: Bootstrapping a Chatino speech corpus, forced aligner, ASR. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 4004–4011. Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1632.