Automatic Phone Alignment of Code-switched Urum–Russian Field Data

Emily P. Ahn
University of Washington
eahn@uw.edu

Eleanor Chodroff
University of Zurich
eleanor.chodroff@uzh.ch

Gina-Anne Levow
University of Washington
levow@uw.edu

Abstract

Code-switching, using multiple languages in a single utterance, is a common means of communication. In the language documentation process, speakers may code-switch between the target language and a language of broader communication; however, how to handle this mixed speech data is not always clearly addressed for speech research and specifically for a corpus phonetics pipeline. This paper investigates best practices for conducting phone-level forced alignment of code-switched field data using the Urum speech dataset from DoReCo. This dataset comprises 117 minutes of narrative utterances, of which 42% contain code-switched Urum–Russian speech. We demonstrate that the inclusion of Russian speech and Russian pretrained acoustic models can aid the alignment of Urum phones. Beyond using boundary alignment precision and accuracy metrics, we also discovered that the method of acoustic modeling impacted a downstream corpus phonetics investigation of code-switched Urum–Russian.

1 Introduction

Code-switching is a phenomenon where multilingual speakers communicate in more than one language, often within a single utterance.[1] Speakers of languages that are not widely spoken may also speak a *language of broader communication*, or *lingua franca*, in order to communicate with people in the same region or in contact settings. In the language documentation and analysis pipeline, recordings of the target language can be found to be mixed with a language of broader communication. Yet this other language is often overlooked or explicitly ignored if the goal of the fieldwork is to document the language of interest. On the other hand, it may be useful to include the mixture of languages in the analyzed data for methodological or scientific purposes. The extra data could add robustness to the performance of models, or the code-switched speech could better reflect actual usage of the target language.

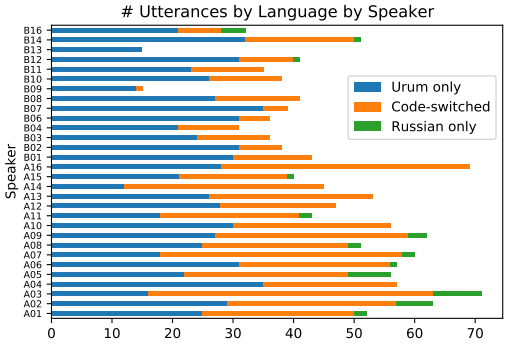

Figure 1: Across the 30 speakers in the DoReCo Urum field repository (Skopeteas et al., 2022), almost all produced code-switched utterances (orange, middle) in addition to monolingual Urum (blue, left) and monolingual Russian utterances (green, right).

Regardless of the analytical use, inclusion of the code-switched language data may benefit processes within the corpus phonetics pipeline. A critical part of this pipeline is phonetic forced alignment, in which a time-aligned phone sequence is identified from the input speech and corresponding transcript, typically using acoustic models of the language-specific phones. Generally, automatic alignment quality correlates with the amount of training data used for the acoustic model (Chodroff et al., 2024). In the case of code-switched speech, there is a question, however, of whether to use *only* the target language data—or to use *all* of the linguistic data—for training the acoustic models. Including code-switched speech during training would result in more data per speaker, which could help build more robust phone-specific acoustic models (as hy-

[1]While *code-switching* can refer to mixing languages or dialects within a whole conversation, we use it to mean switching languages within a single utterance. This finer-grained mixing is also called *code-mixing* in the literature.

pothesized by Chodroff et al., 2024).

Very limited research has included code-switched speech in forced alignment studies, and our work is the first to examine this type of speech in a field data setting. We ask the following research questions (RQs):

1. Does the inclusion of Russian code-switched data in acoustic model training help the alignment of target Urum data?

2. Does the method of acoustic modeling impact a downstream corpus phonetics investigation of code-switched Urum–Russian?

In this paper, we summarize prior work and introduce the Urum language (Section 2), then discuss the methodology of data preparation, acoustic modeling and forced alignment, evaluation and analysis (Section 3). We used the Montreal Forced Aligner (McAuliffe et al., 2017) to train acoustic models from scratch as well as adapt pretrained Russian and English models to our data. With respect to RQ1, we found that the inclusion of code-switched speech and Russian pretrained models improved alignments of Urum (Section 4). To answer RQ2, we tested the impact of acoustic modeling strategies in a bilingual phonetics investigation (Section 5): Are vowels in Urum words pronounced differently in monolingual Urum utterances than in code-switched utterances? After discussion, we conclude with methodological recommendations and areas for future work (Section 6). All code for replicating this work is publicly available.[2]

2 Background

2.1 Phonetic forced alignment

For phonetics research, it can be extremely useful to know the temporal locations of phones within a speech recording. While this can be achieved manually, an automated process can greatly facilitate this, speeding up annotation and enabling analysis of substantially larger speech corpora (Labov et al., 2013). Popular forced alignment tools include the Montreal Forced Aligner (MFA, used in this work; McAuliffe et al., 2017), EasyAlign (Goldman, 2011), and WebMAUS (Kisler et al., 2017). Research has explored a range of strategies to force align low-resource data, including cross-language alignment and manipulation of phone categories (e.g., Ahn et al., 2024; Coto-Solano et al., 2018; DiCanio et al., 2013). However, forced alignment work on low-resource languages that are code-switched has been limited.

2.2 Research on the nature of code-switching

Much of the linguistic literature on code-switching has focused on the syntactic and sociopragmatic aspects of engaging multiple languages at once (Bullock and Toribio, 2009; Muysken, 2000). With respect to the phonetics of code-switching, research has focused on how acoustic properties shift when speakers activate multiple languages in their mind. For example, stop consonant voice onset time and speech rate changed noticeably near language switch boundaries between Spanish and English (Fricke et al., 2016). Relevant to our case study, Seo and Olson (2024) recorded read sentences from Korean–English bilinguals to investigate vowel quality across different syntactic structures. They found that English vowels in code-switched Korean–English utterances were more Korean-like in intra-sentential rather than in inter-sentential code-switched structures. We similarly investigated vowel quality in Urum–Russian code-switched utterances for this paper.

It has been observed that a language of broader communication, usually a high-resource language, is often used during the elicitation of a low-resource, target field language. In an overview of methods to bridge language documentation and speech processing technologies, Levow et al. (2017) proposed a language identification task between a high-resource language and a low-resource target language when both are present in field recordings. San et al. (2022) addressed the mixing of high- and low-resource languages by applying state-of-the-art language technologies to detect and transcribe English portions of speech in a dataset documented for the field language Muruwari. In this case, English was largely used in meta-linguistic commentary and questions, such as, *"What is the word for tree?"* This approach helped the annotation process, where authorized people could scan the meta-linguistic content and triage the recordings for later annotators who had more limited access to the corpora. These studies demonstrate that (1) language mixing is prevalent, and (2) applying technology to the higher-resource language can benefit the documentation process.

Developing technologies for code-switching is still a challenging area of research in the Natural Language Processing (NLP) and speech commu-

[2] https://github.com/emilyahn/align_cs

nities. Winata et al. (2023) found over 400 public research papers on code-switching from the ACL anthology and ISCA proceedings over the past few decades. These works focused on tasks ranging from language identification to sentiment analysis to automatic speech recognition (ASR). Among these papers, English mixed with a non-English variety such as Hindi, Chinese, and Spanish, was overrepresented. The authors highlighted a need for work to be done on more diverse non-English language pairs, for which this paper fills a gap.

Forced alignment with code-switched data Two studies incorporated a language of broader communication when training forced alignment systems on field data, though the impacts of mixed language speech input were not explicitly studied. Ahn et al. (2024) included Portuguese speech when developing acoustic models for Panãra, an Amazonian language of Brazil. Chodroff et al. (2024) retained the Russian speech content in the acoustic modeling of Evenki, a Tungusic language, and Urum, a Turkic language, which is also used in this work (Kazakevich and Klyachko, 2022; Skopeteas et al., 2022).

More relevant to the present study is work by Pandey et al. (2020) who compared methods of training and aligning code-switched Hindi–English read speech. Three acoustic models were trained with MFA: Hindi-only, English-only, and Hindi–English mixed, and they discovered that the combined model best aligned English-only speech. It was unclear, however, if the high performance from the Hindi–English mixed acoustic models was due to that model simply having more tokens in its training data than the other models. Our work extends these findings to a low-resource field data scenario with spontaneous speech, and we carefully controlled the variable of training data quantity. We investigated whether including code-switched data could improve the alignment performance of a target low-resource language.

2.3 Urum language

Urum (ISO: uum) is a Turkic language spoken by ethnic Greeks in the Lesser Caucasus of Georgia and in Ukraine. Also known as Caucasian Urum, it is a variety of Anatolian Turkish that is classified as endangered (Campbell et al., 2022). The language variety documented by Skopeteas et al. (2022) and analyzed in this paper has been strongly influenced by Russian since the group's arrival in Georgia in the early 19th century. Notably, most Urum speakers are bilingual in Russian and code-switch often between the two languages (Skopeteas, 2014). Unlike the examples of code-switching being used in purely meta-linguistic commentary, Russian portions of speech in this dataset were part of the narrative content by the speakers. The following shows an example of an Urum–Russian utterance with transliterated Russian displayed in brackets:

> äp halhımız egıler kissäya **[muzıka]** ederıh **[maladež]** **[tantsuet]** oinamah etmäh
>
> *"All the people get together at the church, we organise **[music]**, and the **[youth]** is **[dancing]**."* (Skopeteas et al., 2022)

3 Methodology

3.1 Data source

We utilized the Urum dataset from the DoReCo corpus, which is a field data repository that contains manual word-level and automatic phone-level alignments of speech (Paschen et al., 2020). Traditional and personal Urum narratives were recorded across 30 speakers (16 female, 14 male) and spanned 117 minutes[3] of speech (Skopeteas et al., 2022). Figure 1 presents the distribution of Urum-only, Russian-only, and code-switched utterances among speakers. All but one speaker code-switched. Table 1 reveals that while 42% of the utterances were code-switched, they represent 53% of the repository in minutes.[4] Code-switched utterances averaged 6.5 seconds, which was on average longer than non-codeswitched utterances (Urum-only: 4.5 seconds; Russian-only: 2 seconds). Among the code-switched utterances, Urum word tokens were more frequent than Russian word tokens, as shown in Figure 2.

3.2 Data preparation

Data from the field repository included long-form audio recordings (wav format, sampling rate of 44.1 kHz) and time-aligned transcriptions of the utterances, words, and phones. The audio files were first segmented into utterances using Praat (Boersma and Weenink, 2022), with the corresponding utterance transcripts as Praat TextGrids. Four utterances were removed due to encoding issues.

[3]Time was calculated by summing utterance durations, not file or word durations.

[4]If all utterances with "foreign material" were excluded, as was the protocol in Zhu et al. (2024) over the full DoReCo corpus, we would miss out on half the data.

Utt	Count	Time (min)	Avg (sec)
All	1373	117.6	5.1
Urum	752 (55%)	53.5 (45%)	4.3
CS	581 (42%)	62.9 (53%)	6.5
Russ	40 (3%)	1.3 (1%)	2.0

Table 1: Distribution of utterances across language usage by count and time. Notably, code-switched (CS) utterances had longer durations.

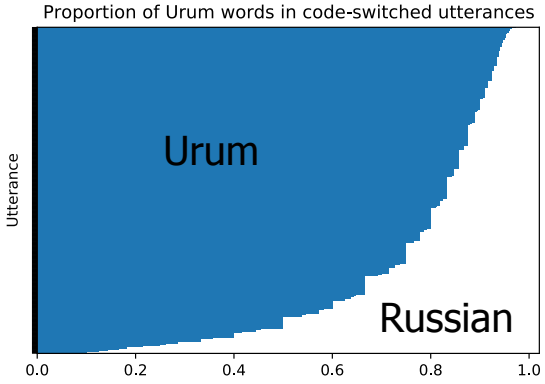

Figure 2: Proportion of Urum (blue, shaded) to Russian (white) word tokens in all 581 code-switched utterances, sorted highest to lowest. The majority of these utterances had more Urum than Russian tokens.

Urum phone sequences were derived automatically by the repository contributors, so our lexicons (two-column text files with words and their corresponding phone sequences) were gathered from these existing phone sequences. Most of the Russian words had been transliterated into Latin script at the word-level, so we used a simple mapping script to build the lexicon. The Urum phone set from DoReCo included nine vowels and 30 consonants while the transliterated Russian phone set included six vowels and 19 consonants. Only four Russian phones did not exist in the Urum set, as seen in Table 2, and we used the PanPhon tool (Mortensen et al., 2016) to map them to their nearest neighboring Urum phones in the lexicons: ɨ → ɯ, tɕ → tʃ, ʂ → ʃ, ʐ → ʒ. Partially-tagged words such as filled pauses and prolongations were assigned phone sequences in the lexicon and were marked as Urum words.[5]

[5]The repository contributors used tags to transcribe content such as filled pauses, prolongations, and false starts. When a tagged word was partially transcribed (such as in this example of a false start, "<fs>ba"), we manually assigned it a phone sequence ("[b a]") and classified it as an Urum word.

Urum-only	Both	Russian-only
y, æ, œ, ɯ	a, e, i, o, u	ɨ
ɟ, c, dː, tː	b, p, d, t, g, k	
sː, ʃ, ʒ, ɣ, dʒ, tʃ	v, f, z, s, x	tɕ, ʂ, ʐ
l, lː, ɾ, mː	j, r, ɬ, m, n	

Table 2: The phone sets present in the DoReCo transcriptions across Urum and Russian with the middle column representing their overlap.

3.3 Acoustic modeling

We used the Montreal Forced Aligner (MFA version 2.2.17; McAuliffe et al., 2017) to train acoustic models and conduct forced alignment on our data in its default unsupervised manner. Acoustic models learn the probability distributions for all given phone states and their transitions. We split the DoReCo files into mutually exclusive train and test partitions following the same split as Chodroff et al. (2024): 1,097 utterances (100 minutes) in the train set and 273 utterances (16 minutes) in the test set. For this study, we created further subsets of the training data to answer our first research question. First, we summed the minutes of utterances of each language type and found 47 minutes of monolingual Urum utterances and 52 minutes of code-switched utterances. To keep the quantity of Urum-only and code-switched training data the same, we reduced the number of code-switched utterances to 47 minutes, which would equal the Urum-only speech. Our first results compared the alignment performance of a model trained on 47 minutes of purely Urum speech to a model trained on 47 minutes of Urum–Russian speech.[6] Our third training data partition combined both sets to include 94 minutes of Urum-only and code-switched speech. All Russian-only utterances were excluded from training and evaluation.

Since it has been shown to be advantageous to use larger, pretrained models for aligning low-resource languages (e.g., Ahn et al., 2024), we chose two relevant MFA models to continue the experiments. The Russian MFA v3.1.0 model was trained on over 400 hours of data from over 3,000 speakers; this model was selected since Russian was frequently spoken in our dataset (McAuliffe and Sonderegger, 2024). The Global English MFA

[6]While the minutes across the two partitions were the same, the number of utterances was 618 for Urum and 414 for code-switched. However, the number of phones in each partition was roughly 27,000.

v2.2.1 model was trained on over 3,700 hours of data from over 79,000 speakers across the world (McAuliffe and Sonderegger, 2023). This model has previously proven to be effective in aligning the Urum dataset in Chodroff et al. (2024). For cross-language modeling and alignment, we developed the lexicons by applying the PanPhon tool for determining the nearest neighboring phones in cases where the target phone did not exist in the model (Mortensen et al., 2016). Appendix A displays these phone mappings. Each pretrained model was adapted to the same three training data partitions as described in the train-from-scratch settings above.

3.4 Forced alignment evaluation

Our "gold standard" phone alignments for evaluation of the system outputs were obtained from the manually annotated phone boundaries of Urum words in the test partition from Chodroff et al. (2024). For *precision*, we calculated the percent for which the model onset boundary was within 20 ms (the selected threshold) of the manually aligned onset boundary (McAuliffe et al., 2017; MacKenzie and Turton, 2020). For *accuracy*, we utilized a measure that calculated the proportion of model-aligned intervals whose midpoints lay within the respective gold intervals (a similar measure is used in Knowles et al., 2018; Mahr et al., 2021). All evaluation was conducted on the test partition which consisted of 132 Urum utterances and 119 code-switched utterances. The evaluation was conducted only on phones from Urum words and ignored all phones from Russian words.

3.5 Analysis

We conducted mixed-effects regressions in R using the lme4 package to analyze the variables that contributed to both the precision and the accuracy metrics (Bates et al., 2015). We ran two models: the first was a linear model with the dependent variable as log seconds of onset boundary differences, with 0 seconds mapped to 0.001 prior to the log transformation. The second model was a logistic regression with the binary dependent variable being accuracy. Main effects were the language of the test utterance (Urum or CS), the natural class of the current phone, the natural class of the previous phone, the interaction of these two natural classes, the proportion of contaminated (tagged) tokens,[7] the utterance duration (in hectoseconds, seconds / 100, for model convergence), the interactions of model configuration with utterance language, and whether or not the speaker of the test utterance was present in the training set. Random effects were the speaker ID and the file ID of the utterances. The current phone class was sum-coded with the held-out level of stop; the previous phone class was sum-coded with the held-out level of silence. The eight classes analyzed were vowels, approximants, taps/trills, nasals, fricatives, affricates, and stops. Models were treatment-coded, each compared to the train-from-scratch Urum-only (47m) model.

[7] Contamination in an utterance was calculated as the number of tagged tokens, such as false starts or prolongations, divided by the total number of tokens.

4 Results

4.1 Alignment precision and accuracy

The following results answer our first research question: Does including Russian code-switched data in acoustic model training help the alignment of target Urum data? The different acoustic model configurations were trained or adapted on subsets of the DoReCo dataset, and they were all tested on the held-out test utterances that included both Urum-only and CS utterances. In the scenario where we trained MFA models from scratch (i.e., no pretrained model was used—note the None column in Table 3), we have two findings. When we kept the training data quantity equal at 47 minutes for both Urum-only speech and code-switched speech, the Urum-only model (47m) outperformed the purely code-switched one (47m). This was expected given that we evaluated only on phones from Urum words. However, combining these two training sets in the Urum + CS (94m) model substantially improved upon either smaller model. This also conforms to expectations given that the combined training set included more Urum tokens and also more data overall.

For the experiments using pretrained models adapted on the various Urum/CS partitions, the Russian MFA model adapted on Urum + CS (94m) produced the best results. Even though the Global English MFA model was trained on nearly 4,000 hours of diverse speech, its alignments did not outperform the smaller Russian MFA model. This is perhaps due to the language similarity of Russian to Urum, or the history of Urum being influenced by Russian contact. All models trained or adapted on the different DoReCo subsets patterned the same where the ranking of best to worst subset was Urum + CS (94m) > Urum (47m) > CS (47m), with the

Train/Adapt Partition	Pretrained model None	Eng	Russ
Precision % ↑			
Urum (47m)	63.2	70.4	71.2
CS (47m)	58.2	70.0	70.4
Urum + CS (94m)	70.9	70.6	**71.3**
Accuracy % ↑			
Urum (47m)	80.6	83.7	84.9
CS (47m)	77.2	83.1	84.4
Urum + CS (94m)	**85.1**	83.6	**85.1**

Table 3: Results revealed that the Russian MFA model adapted on all 94 minutes of Urum and code-switched (CS) data performed the best, with maximal training-from-scratch (i.e., Urum + CS (94m)) on par in terms of accuracy. Highest scores are bolded and shaded.

slight exception of accuracy from the Global English MFA with Urum (47m) > Urum + CS (94m).

4.2 Regression analysis

The mixed-effects regression analysis revealed several factors that influenced alignment performance. We report all significant findings of $p < 0.05$, and full output tables are provided in Appendix B. Except for the train-from-scratch CS (47m) model which performed significantly worse, all other models performed significantly better than the Urum (47m) model. Longer utterance durations and higher contamination amounts were correlated with worse performance. The speaker appearing in the training data had no significant effect. The language of the test utterance also had no effect, with a slight exception of the CS (47m) model performing slightly worse on Urum-only test utterances.

In terms of precision, boundaries around taps/trills were displaced more significantly, while boundaries around fricatives showed higher precision. Boundaries preceding vowels also performed better. Significantly better precision was found for vowel–tap/trill, fricative–vowel, affricate–vowel, affricate–nasal, stop–vowel, and stop–tap/trill sequences. Significantly worse precision was found for vowel–vowel, vowel–approximant, tap/trill–vowel, nasal–nasal, and fricative–approximant sequences.

As for accuracy, which used a logistic mixed-effects regression model, significantly better performance was found for phone intervals preceded by nasals, fricatives, affricates, and stops, as well as for targeted phone intervals that were fricatives and affricates. Significantly worse accuracy was found for phone intervals preceded by vowels, approximants, and taps/trills, as well as targeted phone intervals of these three classes. These results are largely comparable to the mixed-effects regression results from Chodroff et al. (2024).

5 Case Study

Following Babinski et al. (2019), we asked a general phonetics question and observed whether there were significant differences between the outputs of the different model configurations above. In other words, to what degree are we comfortable substituting an automatic alignment for manual alignment, in our quest to answer a question about code-switching phonetics? We investigated the following: Are vowels in Urum words pronounced differently in monolingual Urum utterances compared to in CS utterances? First, we answered this with the manually-annotated "gold" test data.

5.1 Methodology

The Pillai–Bartlett trace, or Pillai score, is a useful metric to measure overlap in vowel category qualities. It takes output from a MANOVA test, which is used for measuring overlap between two distributions across two dependent variables—in our case, the first two formant values. Among four commonly used metrics for vowel overlap, Kelley and Tucker (2020) showed that Pillai scores are among the most reliable. Stanley and Sneller (2023) additionally provided a formula to derive a threshold for determining overlap vs separation based on the exact sample size of the tokens. We followed these recommendations and calculated Pillai scores for formant values extracted from the gold test set. Formants were first extracted with the Linear Predictive Coding (LPC) tool in Praat (Boersma and Weenink, 2022), searching for five formants under 5000 Hz for reported male speakers and 5500 Hz for reported female speakers. The formant value analyzed per vowel was an average of the values extracted from the interval midpoint and ten milliseconds before and after the midpoint.

5.2 Results from manual alignments

The gold test data revealed several instances of within-speaker differences in pronouncing certain Urum vowels. Table 4 shows four instances of a particular vowel being marked as significantly non-overlapping across two conditions. For example, the cell for male speaker A03 /a/ marked with

	Spkr	VOWELS a	e	i	o	u	y	œ	æ	ɯ
Male	A01									
	A03	n=189			n=57					
Female	A02									
	A07	n=13								
	B08									
	B11									
	B16	n=20								

Table 4: Our case study revealed that from the gold data, /a/ for 3 speakers and /o/ for 1 speaker (marked with shaded cells and token counts) in Urum words were pronounced significantly differently in monolingual Urum vs code-switched utterances. For these four instances, Pillai scores indicated that the vowel formants for the two groups in question (Urum vs CS) were significantly non-overlapping.

Spkr	a	e	i	o	u	y	œ	æ	ɯ
A01	X								
A03	X		X	X					
A02									
A07	X								
B08								X	
B11									
B16									

Table 5: The *best-performing* acoustic model (Russian MFA adapted on Urum + CS 94m) yielded 3 true positives (shaded X), 3 false positives (unshaded X), and 1 false negative (shaded empty cell).

Spkr	a	e	i	o	u	y	œ	æ	ɯ
A01	X								
A03			X	X				X	X
A02									
A07	X								
B08									
B11									
B16									

Table 6: The *worst-performing* acoustic model (trained on the CS 47m partition) yielded 2 true positives (shaded X), 4 false positives (unshaded X), and 2 false negatives (shaded empty cells).

$n = 189$ indicates that A03 uttered 189 /a/ vowels, and his F1×F2 values for /a/ in Urum words from monolingual Urum utterances were significantly different than values for /a/ in Urum words from code-switched utterances. The same can be said for speaker A03's /o/ ($n = 57$), speaker A07's /a/ ($n = 13$), and speaker B16's /a/ ($n = 20$).

5.3 Results from automatic alignments

Second, we calculated Pillai scores from the output of the best-performing and worst-performing models and compared these to the gold scores (Tables 5 and 6). From the best-performing model, the Russian MFA model adapted on the Urum + CS (94m) data, it found six instances of significant non-overlap. Three out of the four gold instances were correctly identified (i.e., three true positives and one false negative), while producing three spurious significant findings (i.e., three false positives). From the worst-performing model, trained on the CS (47m) partition, it produced less congruent findings: only two out of the four gold instances were correctly identified (i.e., two true positives and two false negatives), with four spurious significant findings (i.e., four false positives). We used the phonR package in R (McCloy, 2012) to plot vowel ellipses for /i, a, o/ for male speaker A03, over the two language conditions, and across the three types of output (Figure 3). The gold plot reflects our findings that /a/ and /o/ were significantly different between Urum and CS environments while /i/ was not. The ellipses from the best and worst models show divergence from the gold ellipses. Both models found spurious differences for /i/, and although /a/ visually appears significantly different for the worst model, /a/ was a false negative.

Essentially, the automatic alignments did not yield the same findings as those from the gold alignments in our vowel overlap analysis. Output from the best- and worst-performing models tended to

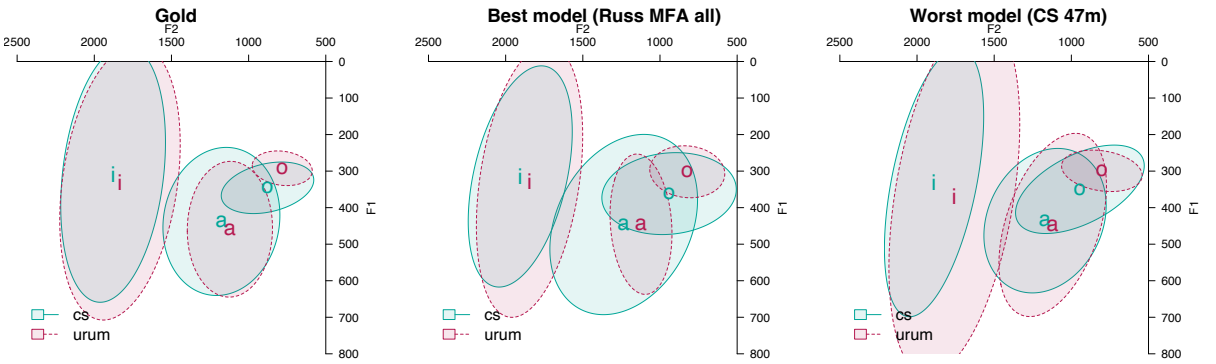

Figure 3: These plots reflect the first two formants (in Hz) for three of the nine Urum vowels, /a, i, o/, for male speaker A03. From left-to-right are formants extracted from the gold alignments, the best model (Russian MFA adapted on Urum + CS 94m) output, and the worst model (CS 47m) output. Vowel labels are positioned at means, and ellipses cover one standard deviation away from the mean.

hallucinate more vowel disparities than the gold output suggested, though the best model's vowel disparity predictions more closely aligned to the gold findings than the worst model's. While the best model's alignments were 11 percentage points higher than the worst model's alignments for precision (and seven percentage points higher for accuracy), these differences can be hard to interpret. This case study allowed us to reveal the nuances of alignment performance, as the downstream output yielded different conclusions.

6 Conclusion

This work tested methodologies of incorporating code-switched data in acoustic model training and alignment in a low-resource, field data scenario. We tested the inclusion of Urum–Russian code-switched utterances in training acoustic models to align Urum phones and found that it was helpful to keep the code-switching to produce a larger train set.[8] The maximally trained-from-scratch model performed roughly on-par with a pretrained Russian model adapted to the same field data. If one is fortunate enough to have 90-some minutes of transcribed data, it should be sufficient to train models (see also the recommendations in Chodroff et al., 2024). Otherwise, utilizing a large, pretrained model performed reasonably, particularly when adapted on target data.

In order to functionally assess the quality of the systems, we tested our best and worst systems' alignment outputs against the gold alignments to answer a bilingual phonetics question (RQ2). Calculating Pillai scores across formant values for individual speakers, we discovered that several speakers pronounced certain Urum vowels significantly differently in monolingual Urum utterances than in code-switched utterances. While not matching the gold alignment results exactly, the best acoustic model yielded more similar results to the gold than the worst acoustic model. We recommend manual adjustment of phone boundaries when conducting phonetic analyses, particularly those involving smaller datasets and temporally sensitive phonetic measurements (e.g., analysis of duration or cases where the boundary determines the measurement location such as onset f0).

As future work, it would be beneficial to conduct a survey study with qualitative and quantitative statistics on the prevalence of code-switching across field data repositories. How are multiple languages used by the elicitors and by the community members of the language being documented?

Further research could also aim to extend the study of phonetics and phonology for code-switched language more broadly. Our case study only scratched the surface to discover the nature of shifting Urum vowel qualities depending on the languages present in an utterance. It would be interesting to discover if the significantly different Urum vowel formants were becoming more Russian-like when surrounded by Russian context, similar to findings on Korean–English by Seo and Olson (2024). Cross-linguistic interference or transfer could be in effect and is worth investigating.

[8] Our findings echo similar cross-language modeling experiments from other domains such as speech recognition and text-based NLP research, where the inclusion of data from a higher-resource language improved model performance on low-resource language data (e.g., Downey et al., 2024; Farooq and Hain, 2023; Fujinuma et al., 2022).

Limitations

When conducting our regression analyses or case study, we did not take into account code-switching properties at the syntactic or prosodic level. It would be interesting to factor into account whether the code-switched utterance was inter-sentential or intra-sentential (i.e., mixing languages at phrase boundaries or within phrases). When calculating boundary differences, examining how close an Urum word was to a Russian word could have provided useful information. Prosodic factors such as speech rate and pitch would also add insight as, anecdotally, prosody was at times visibly different near the language switch points. Additionally, code-switched words can be confused with loanwords that have a legitimate place in a language's lexicon. All of the Russian words in this repository were explicitly tagged as Russian by the field linguists, but there may be disagreement to the classification of language at the token-level.

The Urum dataset from the DoReCo repository used in this work was particularly well-annotated for both Urum and Russian. However, the quality and quantity of transcriptions here may not be comparable to that in other field data repositories, and replication of our findings on other datasets may be challenging.

Ethics Statement

The dataset in this study has been made publicly available for download and research use. Speech data that is public carries inherent potential harms for misuse in downstream tasks.

Particularly for our methodological approach of including code-switched speech or the language of broader communication in the analysis of field data, we advise some caution. The speech from the non-target language may have been meant to be ignored and not recorded. If sections of the speech data were not explicitly transcribed, they may not have been intended to be used for analysis.

References

Emily P. Ahn, Eleanor Chodroff, Myriam Lapierre, and Gina-Anne Levow. 2024. The Use of Phone Categories and Cross-Language Modeling for Phone Alignment of Panãra. In *Interspeech 2024*, pages 1505–1509.

Sarah Babinski, Rikker Dockum, J. Hunter Craft, Anelisa Fergus, Dolly Goldenberg, and 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*, 4:3:1–12.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-effects Models Using lme4. *Journal of Statistical Software*, 67:1–48.

Paul Boersma and David Weenink. 2022. Praat: Doing Phonetics by Computer (Version 6.0.3).

Barbara E Bullock and Almeida Jacqueline Toribio. 2009. *The Cambridge Handbook of Linguistic Code-switching*. Cambridge University Press.

Lyle Campbell, Nala Huiying Lee, Eve Okura, Sean Simpson, and Kaori Ueki. 2022. The Catalogue of Endangered Languages (ElCat). Database available at http://endangeredlanguages.com/userquery/download/, accessed 2022-08-28.

Eleanor Chodroff, Emily P. Ahn, and Hossep Dolatian. 2024. Comparing Language-specific and Cross-language Acoustic Models for Low-resource Phonetic Forced Alignment. *Language Documentation & Conservation*.

Rolando Coto-Solano, Sally Akevai Nicholas, and Samantha Wray. 2018. Development of Natural Language Processing Tools for Cook Islands Māori. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 26–33, Dunedin, New Zealand.

Christian DiCanio, Hosung Nam, Douglas H. Whalen, H. Timothy Bunnell, Jonathan D. Amith, and Rey Castillo García. 2013. Using Automatic Alignment to Analyze Endangered Language Data: Testing the Viability of Untrained Alignment. *The Journal of the Acoustical Society of America*, 134(3):2235–2246.

C. M. Downey, Terra Blevins, Dhwani Serai, Dwija Parikh, and Shane Steinert-Threlkeld. 2024. Targeted Multilingual Adaptation for Low-resource Language Families. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15647–15663, Miami, Florida, USA. Association for Computational Linguistics.

Muhammad Umar Farooq and Thomas Hain. 2023. Learning Cross-lingual Mappings for Data Augmentation to Improve Low-resource Speech Recognition. In *Interspeech 2023*, pages 5072–5076.

Melinda Fricke, Judith F. Kroll, and Paola E. Dussias. 2016. Phonetic Variation in Bilingual Speech: A Lens for Studying the Production–comprehension Link. *Journal of Memory and Language*, 89:110–137. Speaking and Listening: Relationships Between Language Production and Comprehension.

Yoshinari Fujinuma, Jordan Boyd-Graber, and Katharina Kann. 2022. Match the Script, Adapt if Multilingual: Analyzing the Effect of Multilingual Pretraining on Cross-lingual Transferability. In *Proceedings*

of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1500–1512, Dublin, Ireland. Association for Computational Linguistics.

Jean-Philippe Goldman. 2011. EasyAlign: an Automatic Phonetic Alignment Tool under Praat. In *Interspeech 2011*, pages 3233–3236.

Olga Kazakevich and Elena Klyachko. 2022. Evenki DoReCo Dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin and Lyon. Accessed on 17 Nov 2022.

Matthew C. Kelley and Benjamin V. Tucker. 2020. A Comparison of Four Vowel Overlap Measures. *The Journal of the Acoustical Society of America*, 147(1):137–145.

Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual Processing of Speech via Web Services. *Computer Speech & Language*, 45:326–347.

Thea Knowles, Meghan Clayards, and 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.

William Labov, Ingrid Rosenfelder, and Josef Fruehwald. 2013. One Hundred Years of Sound Change in Philadelphia: Linear Incrementation, Reversal, and Reanalysis. *Language*, 89(1):30–65.

Gina-Anne Levow, Emily M. Bender, Patrick Littell, Kristen Howell, Shobhana Chelliah, Joshua Crowgey, Dan Garrette, Jeff Good, Sharon Hargus, David Inman, Michael Maxwell, Michael Tjalve, and Fei Xia. 2017. STREAMLInED Challenges: Aligning Research Interests with Shared Tasks. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 39–47.

Laurel MacKenzie and Danielle Turton. 2020. Assessing the Accuracy of Existing Forced Alignment Software on Varieties of British English. *Linguistics Vanguard*, 6(s1):20180061.

Tristan J. Mahr, Visar Berisha, Kan Kawabata, Julie Liss, and Katherine C. Hustad. 2021. Performance of Forced-Alignment Algorithms on Children's Speech. *Journal of Speech, Language, and Hearing Research*, 64(6S):2213–2222.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech 2017*, pages 498–502.

Michael McAuliffe and Morgan Sonderegger. 2023. English MFA acoustic model v2.2.1. Technical report, https://mfa-models.readthedocs.io/en/latest/acoustic/English/English%20MFA%20acoustic%20model%20v2_2_1.html.

Michael McAuliffe and Morgan Sonderegger. 2024. Russian MFA acoustic model v3.1.0. Technical report, https://mfa-models.readthedocs.io/en/latest/acoustic/Russian/Russian%20MFA%20acoustic%20model%20v3_1_0.html.

Daniel R McCloy. 2012. Vowel Normalization and Plotting with the phonR Package. *Technical Reports of the UW Linguistic Phonetics Laboratory*, 1:1–8.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484.

Pieter Muysken. 2000. *Bilingual Speech: A Typology of Code-mixing*. Cambridge University Press.

Ayushi Pandey, Pamir Gogoi, and Kevin Tang. 2020. Understanding Forced Alignment Errors in Hindi-English Code-Mixed Speech–a Feature Analysis. In *Proceedings of the First Workshop on Speech Technologies for Code-Switching in Multilingual Communities*, pages 13–17.

Ludger Paschen, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave, and Frank Seifart. 2020. Building a Time-aligned Cross-linguistic Reference Corpus from Language Documentation Data (DoReCo). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2657–2666, Marseille, France. European Language Resources Association.

Nay San, Martijn Bartelds, Tolúlopé Ògúnrèmí, Alison Mount, Ruben Thompson, Michael Higgins, Roy Barker, Jane Simpson, and Dan Jurafsky. 2022. Automated Speech Tools for Helping Communities Process Restricted-access Corpora for Language Revival Efforts. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 41–51.

Yuhyeon Seo and Daniel J. Olson. 2024. Phonetic Shifts in Bilingual Vowels: Evidence from Intersentential and Intrasentential Code-switching. *International Journal of Bilingualism*, 0(0):1–16.

Stavros Skopeteas. 2014. Caucasian Urums and Urum Language. *Journal of Endangered Turkish Languages*, 3(1):333–364.

Stavros Skopeteas, Violeta Moisidi, Nutsa Tsetereli, Johanna Lorenz, and Stefanie Schröter. 2022. Urum DoReCo Dataset. In Frank Seifart, Ludger Paschen,

and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin and Lyon. Accessed on 17 Nov 2022.

Joseph A. Stanley and Betsy Sneller. 2023. Sample Size Matters in Calculating Pillai Scores. *The Journal of the Acoustical Society of America*, 153(1):54–67.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023. The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.

Jian Zhu, Changbing Yang, Farhan Samir, and Jahurul Islam. 2024. 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 (Volume 1: Long Papers)*, pages 750–772. Association for Computational Linguistics.

A Cross-language Phone Mappings

Table 7 shows the mappings from Urum or Russian to either English or Russian pretrained MFA models.

B Regression Results

Tables 8 and 9 display the output from the mixed-effects regression models.

Russ (CS) to Eng MFA	Urum to Eng MFA	Urum to Russ MFA
tɕ → tʃ	dː → d	ɾ → r
ɨ → ɯ	lː → l	œ→ ɛ
ʂ → ʃ	mː → m	ɯ → ɨ
ʐ → ʒ	r → ɾ	ʃ → ʂ
	sː → s	ʒ → ʐ
	tː → t	d → d̪
	x → ç	dː → d̪ː
	y → ʉ	dʒ → dʐ
	œ→ ɛ	l → ɬ
	ɣ → ç	lː → ɬː
	ɯ → ə	n → n̪
		s → s̪
		sː → s̪ː
		t → t̪
		tː → t̪ː
		tʃ → tʂ
		y → ʉ
		z → z̪

Table 7: Urum and Russian (code-switched) phones from DoReCo that did not exist in the pretrained English or Russian MFA model lexicons were mapped to their nearest neighboring phones, calculated with the PanPhon tool (Mortensen et al., 2016).

Predictors	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	-4.39	0.18	-24.37	**<0.001**
CS (47m)	0.11	0.03	3.35	**<0.001**
Urum + CS (94m)	-0.28	0.03	-8.56	**<0.001**
English + Urum (47m)	-0.22	0.03	-6.83	**<0.001**
English + CS (47m)	-0.20	0.03	-6.02	**<0.001**
English + Urum/CS (94m)	-0.22	0.03	-6.67	**<0.001**
Russian + Urum (47m)	-0.20	0.03	-6.17	**<0.001**
Russian + CS (47m)	-0.18	0.03	-5.42	**<0.001**
Russian + Urum/CS (94m)	-0.21	0.03	-6.46	**<0.001**
Utterance duration	4.03	0.25	16.39	**<0.001**
Contamimation amount	0.63	0.05	13.14	**<0.001**
Speaker seen in training	0.20	0.20	0.98	0.360
Utt is Urum-only	0.01	0.03	0.18	0.859
Prec vowel	-0.11	0.08	-1.39	0.166
Prec approx	0.31	0.55	0.57	0.567
Prec tap/trill	0.30	0.04	7.00	**<0.001**
Prec nasal	-0.17	0.11	-1.56	0.118
Prec fric	-0.22	0.10	-2.28	**<0.05**
Prec affr	0.63	0.39	1.63	0.103
Prec stop	-0.12	0.10	-1.19	0.236
Vowel	-0.39	0.08	-4.79	**<0.001**
Approximant	-0.03	0.09	-0.37	0.712
Tap/trill	0.75	0.38	1.97	**<0.05**
Nasal	-0.18	0.09	-1.95	0.052
Fricative	-0.20	0.08	-2.56	**<0.05**
Affricate	-0.01	0.12	-0.10	0.921
CS (47m) x Utt is Urum-only	0.09	0.05	2.06	**<0.05**
Urum + CS (94m) x Utt is Urum-only	0.04	0.05	0.86	0.390
English + Urum (47m) x Utt is Urum-only	0.01	0.05	0.22	0.828
English + CS (47m) x Utt is Urum-only	0.00	0.05	-0.03	0.974
English + Urum/CS (94m) x Utt is Urum-only	-0.01	0.05	-0.15	0.880
Russian + Urum (47m) x Utt is Urum-only	-0.02	0.05	-0.45	0.651
Russian + CS (47m) x Utt is Urum-only	-0.02	0.05	-0.42	0.674
Russian + Urum/CS (94m) x Utt is Urum-only	-0.01	0.05	-0.23	0.822
Prec vowel x vowel	0.79	0.08	9.31	**<0.001**
Prec vowel x approx	0.39	0.09	4.35	**<0.001**
Prec vowel x tap/trill	-0.83	0.38	-2.17	**<0.05**
Prec vowel x nasal	-0.18	0.09	-1.89	0.059
Prec vowel x fric	-0.10	0.08	-1.24	0.215
Prec vowel x affr	0.24	0.13	1.84	0.066
Prec approx x vowel	0.14	0.55	0.25	0.802
Prec approx x approx	-0.86	0.56	-1.54	0.125
Prec approx x tap/trill	0.68	3.26	0.21	0.836
Prec approx x nasal	0.45	0.56	0.81	0.419
Prec approx x fric	-0.16	0.64	-0.25	0.800
Prec approx x affr	-0.14	0.56	-0.25	0.801
Prec tap/trill x vowel	0.15	0.05	3.06	**<0.01**
Prec tap/trill x approx	0.05	0.07	0.66	0.510
Prec tap/trill x nasal	-0.17	0.11	-1.61	0.108
Prec tap/trill x fric	0.06	0.14	0.40	0.693
Prec tap/trill x affr	-0.13	0.13	-1.05	0.296
Prec nasal x vowel	0.21	0.11	1.94	0.052
Prec nasal x approx	0.09	0.16	0.57	0.572
Prec nasal x tap/trill	-0.24	0.54	-0.45	0.651
Prec nasal x nasal	0.52	0.13	4.03	**<0.001**
Prec nasal x fric	-0.24	0.13	-1.94	0.053
Prec nasal x affr	-0.14	0.19	-0.76	0.445
Prec fric x vowel	-0.25	0.10	-2.58	**<0.01**
Prec fric x approx	0.25	0.11	2.21	**<0.05**
Prec fric x tap/trill	-0.54	0.47	-1.16	0.245
Prec fric x nasal	-0.05	0.15	-0.34	0.736
Prec fric x fric	0.16	0.12	1.36	0.173
Prec fric x affr	0.20	0.20	0.99	0.321
Prec affr x vowel	-1.05	0.39	-2.68	**<0.01**
Prec affr x approx	-0.49	0.43	-1.13	0.257
Prec affr x tap/trill	3.15	2.14	1.48	0.140
Prec affr x nasal	-1.04	0.48	-2.20	**<0.05**
Prec affr x fric	-0.01	0.94	-0.01	0.993
Prec stop x vowel	-0.32	0.10	-3.11	**<0.01**
Prec stop x approx	-0.01	0.12	-0.08	0.933
Prec stop x tap/trill	-0.80	0.38	-2.12	**<0.05**
Prec stop x nasal	0.22	0.14	1.60	0.109

Table 8: Linear mixed-effects regression results for phone onset boundary difference (in log seconds, with 0 seconds mapped to 0.001 prior to the log transformation). Models were treatment-coded, each compared to the train-from-scratch Urum-only (47m) model. The current phone class was sum-coded with the held-out level of stop; the previous phone class was sum-coded with the held-out level of silence. Utterance duration was entered as hectoseconds (seconds / 100).

Predictors	**Estimate**	**Std. Error**	**z-value**	**Pr(>|z|)**
Intercept	2.07	0.27	7.68	**<0.001**
CS (47m)	-0.22	0.04	-5.05	**<0.001**
Urum + CS (94m)	0.34	0.05	7.10	**<0.001**
English + Urum (47m)	0.23	0.05	4.94	**<0.001**
English + CS (47m)	0.18	0.05	3.93	**<0.001**
English + Urum/CS (94m)	0.22	0.05	4.67	**<0.001**
Russian + Urum (47m)	0.33	0.05	6.87	**<0.001**
Russian + CS (47m)	0.29	0.05	6.08	**<0.001**
Russian + Urum/CS (94m)	0.35	0.05	7.17	**<0.001**
Utterance duration	-2.13	0.47	-4.56	**<0.001**
Contamimation amount	-0.94	0.10	-9.46	**<0.001**
Speaker seen in training	-0.08	0.34	-0.24	0.814
Utt is Urum-only	0.06	0.03	1.94	0.053
Prec vowel	-0.31	0.03	-10.18	**<0.001**
Prec approx	-0.16	0.04	-3.94	**<0.001**
Prec tap/trill	-0.26	0.04	-6.28	**<0.001**
Prec nasal	0.10	0.04	2.26	**<0.05**
Prec fric	0.24	0.04	6.12	**<0.001**
Prec affr	0.33	0.10	3.34	**<0.001**
Prec stop	0.15	0.03	4.64	**<0.001**
Vowel	-0.23	0.04	-5.90	**<0.001**
Approximant	-0.96	0.04	-23.66	**<0.001**
Tap/trill	-1.11	0.04	-27.73	**<0.001**
Nasal	0.03	0.04	0.60	0.547
Fricative	0.45	0.05	9.97	**<0.001**
Affricate	1.62	0.17	9.72	**<0.001**

Table 9: Logistic mixed-effects regression results for accuracy. Accuracy is 1 if the midpoint of the system interval lies within the corresponding gold interval. Models were treatment-coded, each compared to the train-from-scratch Urum-only (47m) model. The current phone class was sum-coded with the held-out level of stop; the previous phone class was sum-coded with the held-out level of silence. Utterance duration was entered as hectoseconds (seconds / 100).