Forced Alignment of Code-switched Urum-Russian Field Data

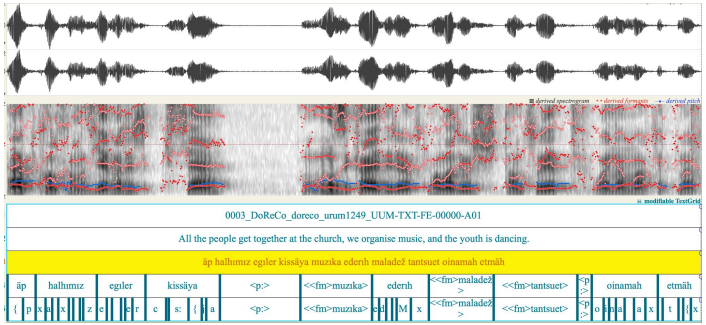
Emily P. Ahn
University of Washington

Phonetics Lab meeting
24 January 2025 | Seattle

Motivation: Our use case

We have some audio recordings and transcriptions for **code-switched** field data, and we want to do phonetic & phonological analyses!

<u>Our goal:</u> perform time-aligned phone segmentation on code-switched speech



Background:
What is code-switching (CS)?

→ Using multiple languages to communicate

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**.”

[Urum, **Russian**]

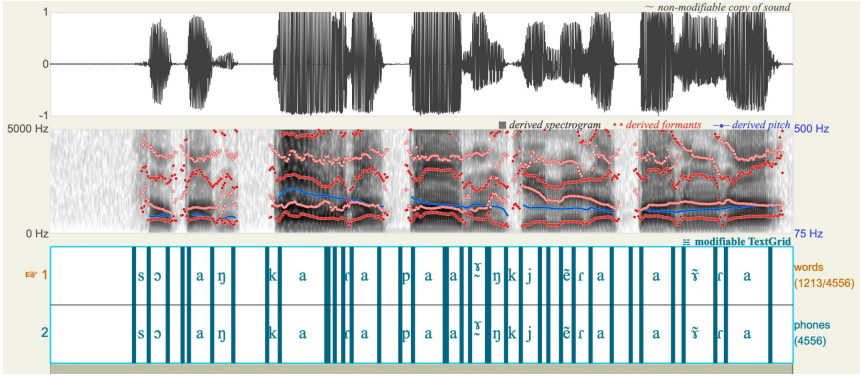
https://awinlanguage.blogspot.com/2017/06/kinds-of-code-switching.html

Background: What is forced alignment?

1) It is a tool that, given speech & text,

Sotanka horapa hamä inkjëmëra Panära

2) does phone segmentation



Motivation: Code-switching is under-studied

- Language of broader communication often used in field data collection, but often ignored
- I have not found any literature on phonetics/phonology of code-switching in field data settings

Motivation:
Code-switching forced alignment?

- Plenty of cross-language forced alignment work
  - e.g. using English model to align Panãra

- 1 paper discussed CS forced alignment
  - Pandey+ (2020): combined model of Hindi + English outperformed monolingual models, but
    - data quantity inconsistent
    - Hindi & English are high-resourced languages

Research Questions

1. Does the inclusion of the language of broader communication, Russian, help or hinder the alignment performance of the target field language data, Urum?
    a. Whether by including Russian in the training data, or using a pretrained Russian model.

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

Data:
Urum overview

Urum (ISO [uum]):

- “Caucasian/Kapchik Urum”
- A Turkic language spoken by ethnic Greeks in Georgia
- Speakers are bilingual in Russian

https://commons.wikimedia.org/w/index.php?curid=2430263

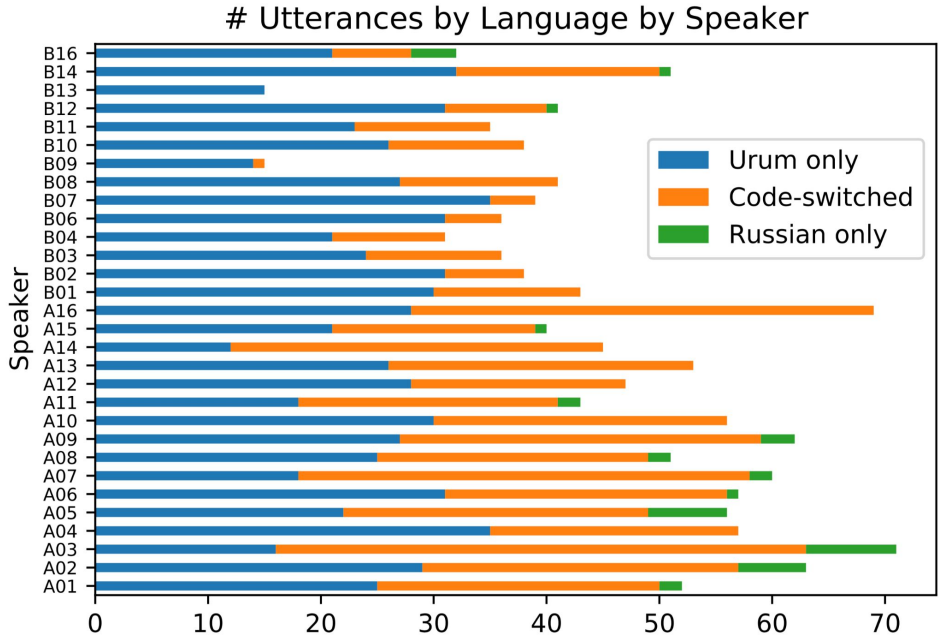
Data:
Urum repository from DoReCo

DoReCo
Language Documentation Reference Corpus

- DoReCo (Paschen+ 2020)
  - field data repository with manual word- and automatic phone-level alignments
- Urum dataset compiled in 2005 (Skopeteas+ 2024)
- 30 speakers (14 male, 16 female)
- 117 minutes of speech, narrative style

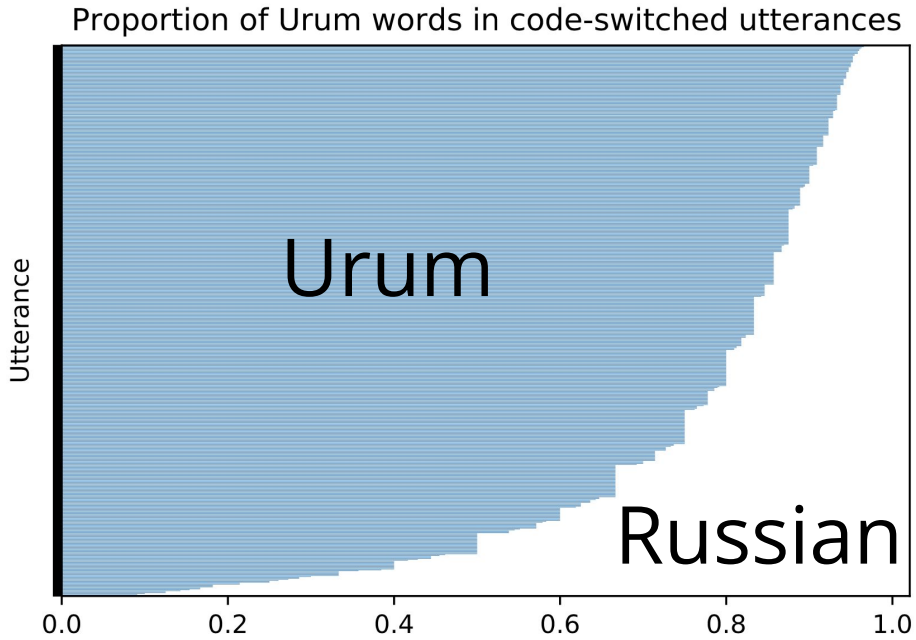
Data:
Urum repository from DoReCo

- 42% all utterances are code-switched
- Almost all speakers code-switch
- Avg utterance
  - Urum: 4.27 sec
  - Code-switched: 6.49 sec



Data:
Urum repository from DoReCo

- The proportion of each language varies across the utterances
  - but most contain more Urum than Russian



Data:
Another Urum-Russian example

aa bizım köv burda urum semyasi yaširıh **gdeto** igirmi beš **semya** öbürlär äp gürjidırlär **a eše** šei **ajar**

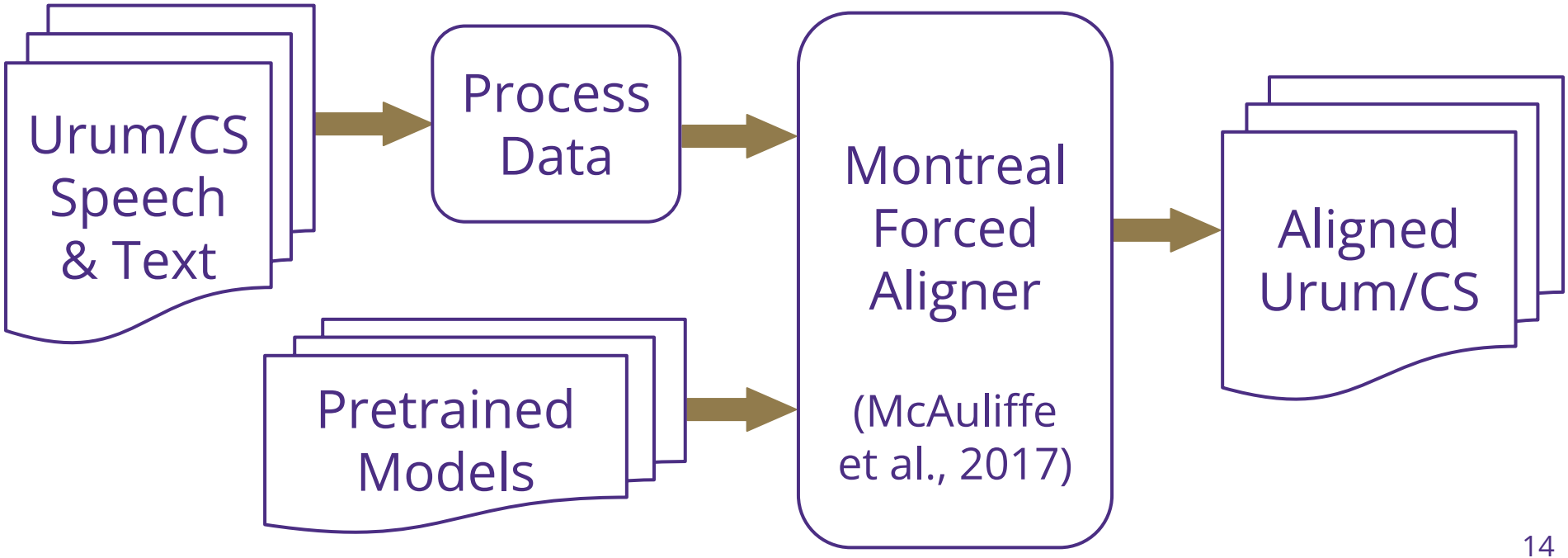
“There are **about** twenty five Urum **families** in our village and **others** are Georgians **and Ajarians**.”

[Urum, **Russian**]

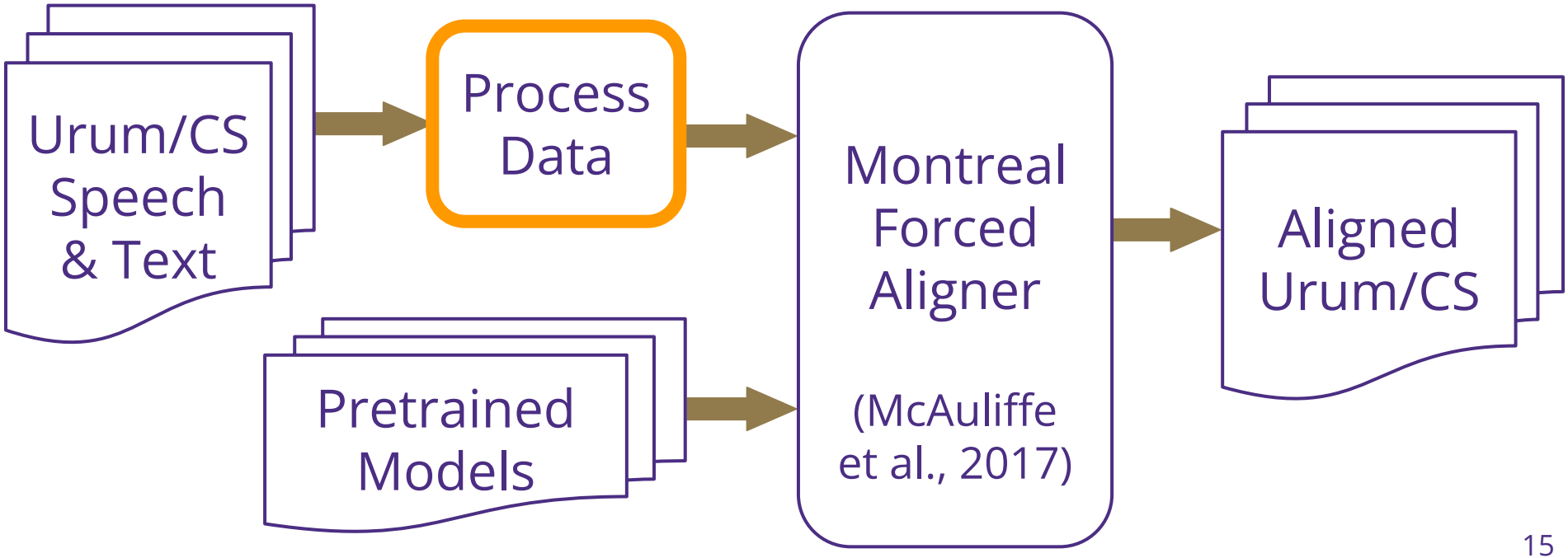
Research Question 1

Does the inclusion of the language of broader communication, Russian, help or hinder the alignment performance of the target field language data, Urum?

Methods Overview 1



Methods Overview 1



Methods: Data processing

1) Segment audio by utterances
2) Assign phone sequences to
   - Russian words
   - tagged words (e.g. false starts, prolongations)
3) Partition train/test data
   - create equal sized Urum-only and CS-only sets
   - utilize overall train/test splits from Chodroff+ (2024)

	# utts	time (min)
TRAIN	1097	100.45
Urum	618	47.11
CS (all)	460	52.53
CS (time=Urum)	414	47.10
Russian	19	0.81
TEST	273	16.96
Urum	132	6.17
CS	119	10.15
Russian	22	0.65

Data:
Phone inventories

Phone sets present in the DoReCo transcriptions

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	

Methods:
Data processing - Lexicon creation

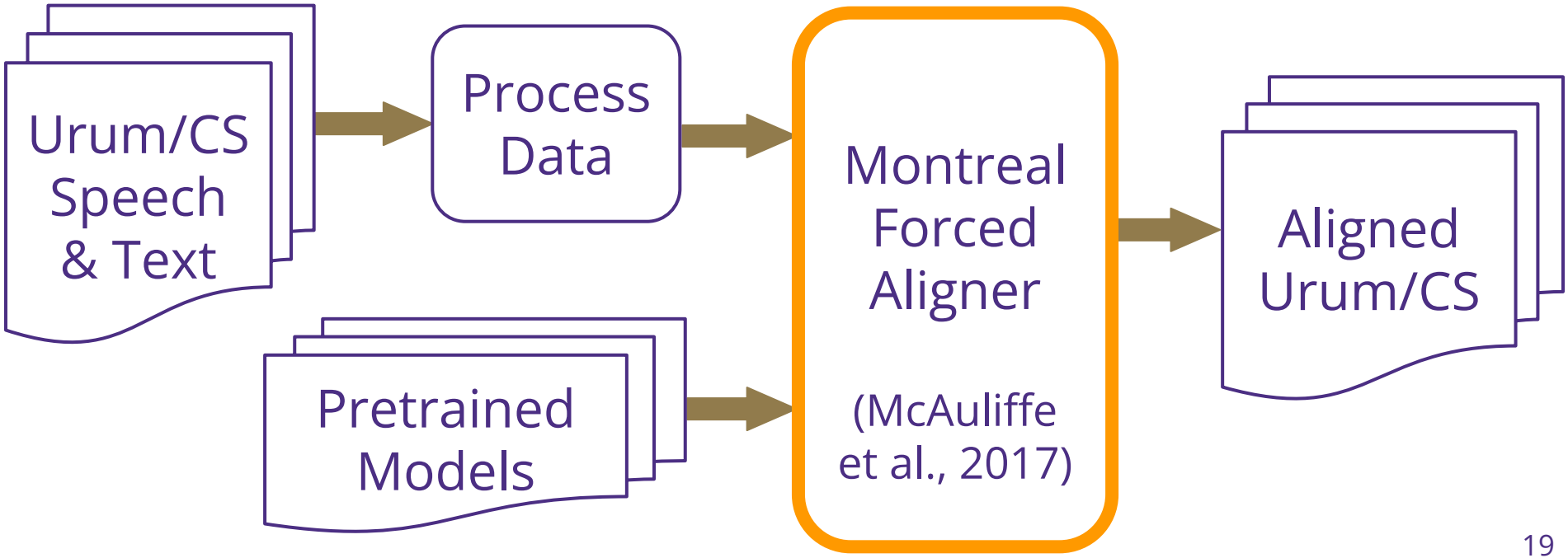
- for Urum words: use DoReCo phone sequences

- for Russian words:
  - convert Latin script to IPA
  - map Russian-only phones to Urum

Examples

kissäya	→	c i s: æ j a
halhımız	→	x a ɬ x ɯ m ɯ z
egıler	→	e g ɯ ɬ e r
_muzıka	→	m u z ɬɯ k a
_maladež	→	m a ɬ a d e zʒ

Methods overview 1

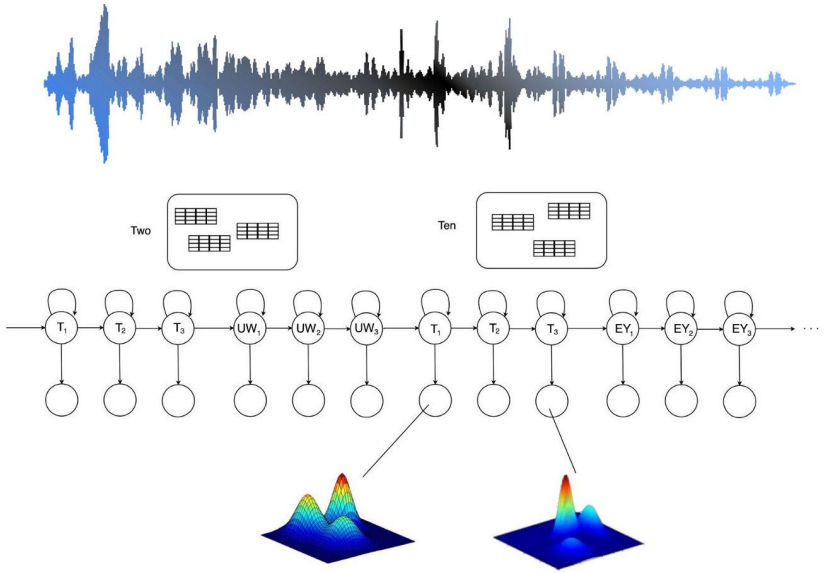


Methods:
Acoustic modeling & forced alignment

Train acoustic models to

1) learn probability distributions for phone states and transitions
2) assign phone boundaries

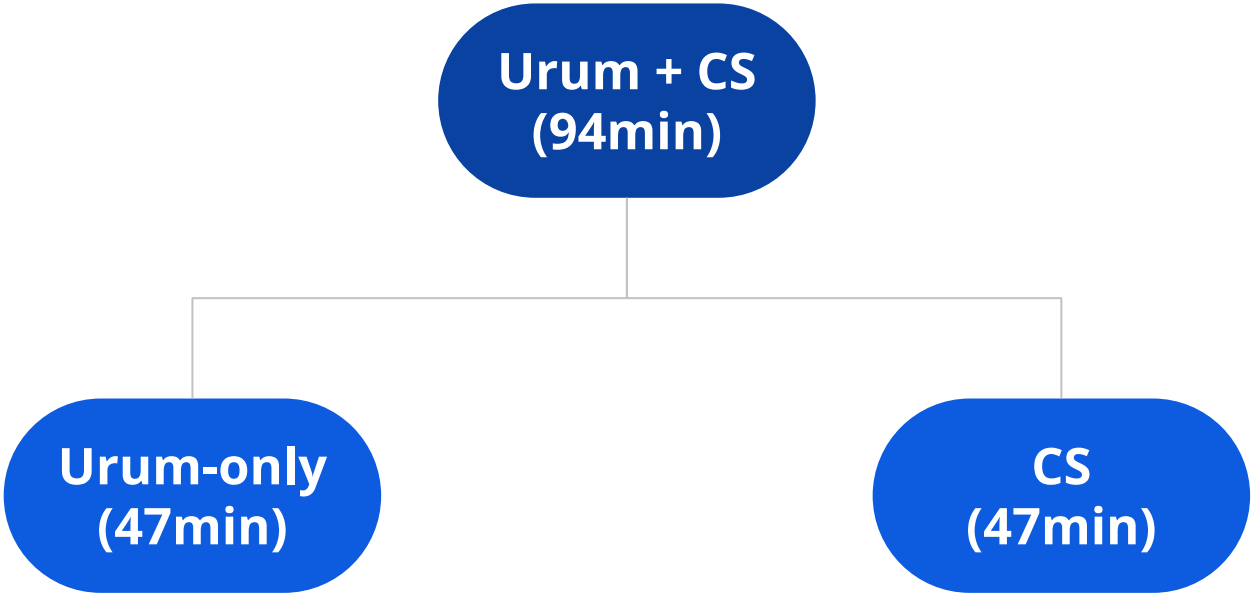
We use the Montreal Forced Aligner (MFA; McAuliffe+ 2017)



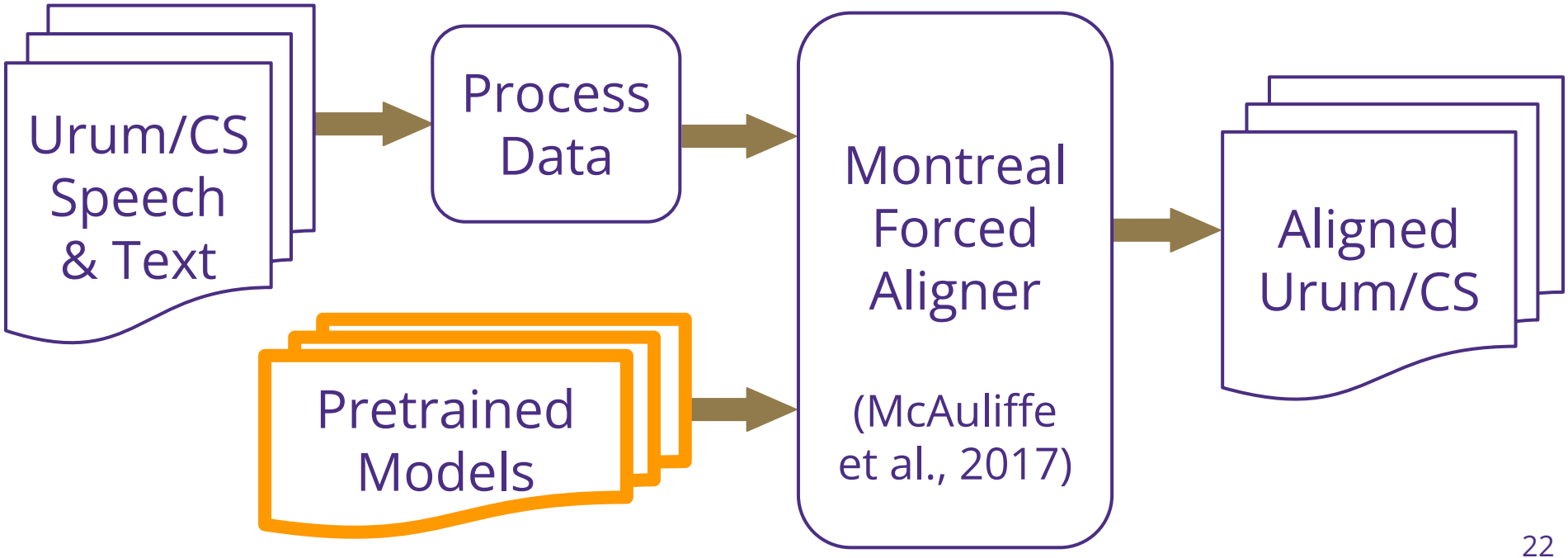
https://miro.medium.com/max/1540/1*YX9aWhQYrVzc-Jf5kQaXkQ.jpeg

Methods: Training-from-scratch

Train models on 3 data partitions:

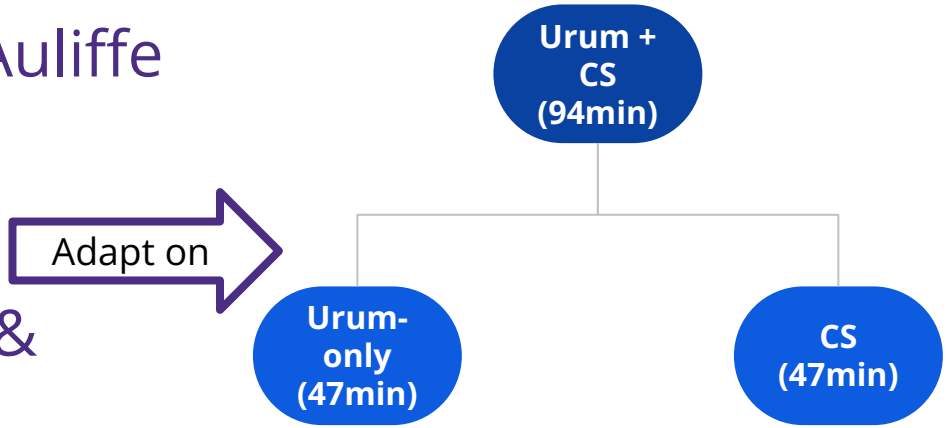


Methods Overview 1



Methods: Pretrained acoustic models

1) Global English MFA (McAuliffe & Sonderegger, 2023)
   - trained on ~4000 hours
2) Russian MFA (McAuliffe & Sonderegger, 2024)
   - trained on ~400 hours



Methods: Pretrained models

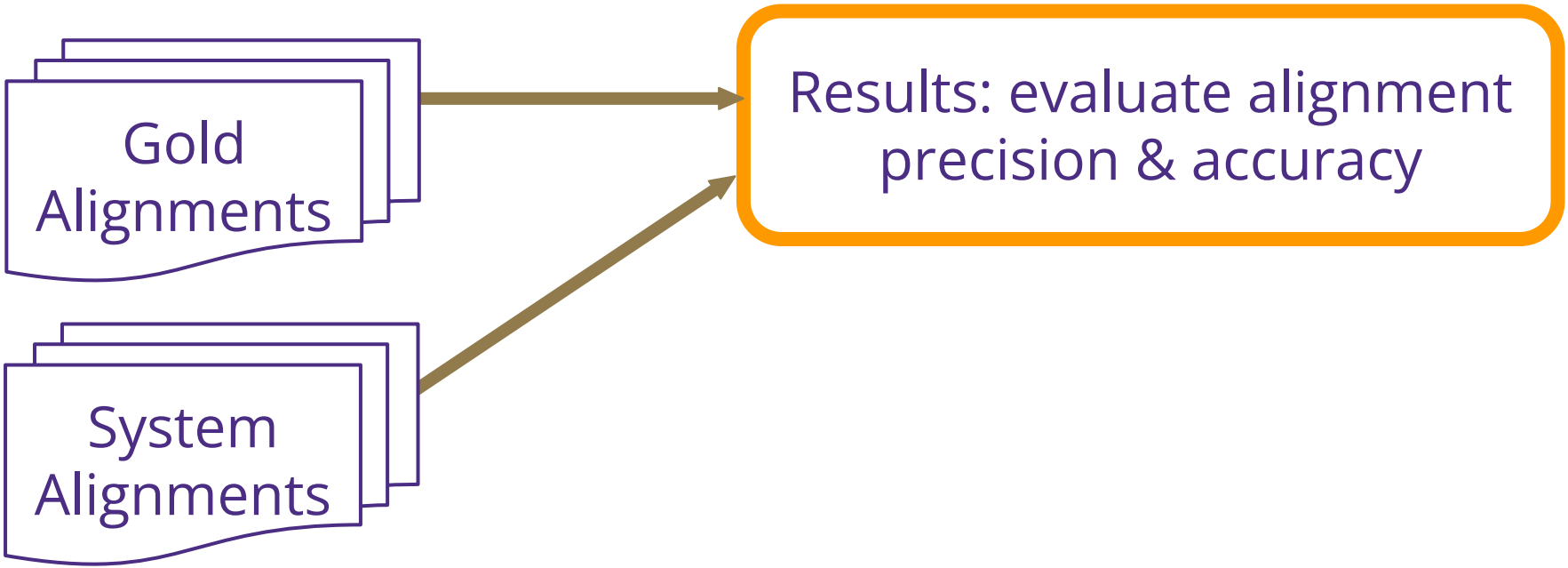
We map Urum/Russian phones to English/Russian phone sets using nearest neighbor calculations with PanPhon (Mortensen+ 2016)

Urum to Global Eng	
d:	d
l:	l
m:	m
r	ɾ
s:	s
t:	t
x	ç
y	ʉ
œ	ɛ
ɣ	ç
ɯ	ə

Russ (CS) to Global Eng	
tɕ	tʃ
ɨ	i
ʂ	ʃ
ʐ	ʒ

Urum to Russian MFA	
ɾ	r
œ	ɛ
ɯ	ɨ
ʃ	ʂ
ʒ	ʐ
d	ɖ
d:	ɖ:
dʒ	dʐ:
l	ɬ
l:	ɬ:
n	ɳ
s	ʂ
s:	ʂ:
t	ʈ
t:	ʈ:
tʃ	tʂ
y	ʉ
z	ʐ

Methods Overview 2



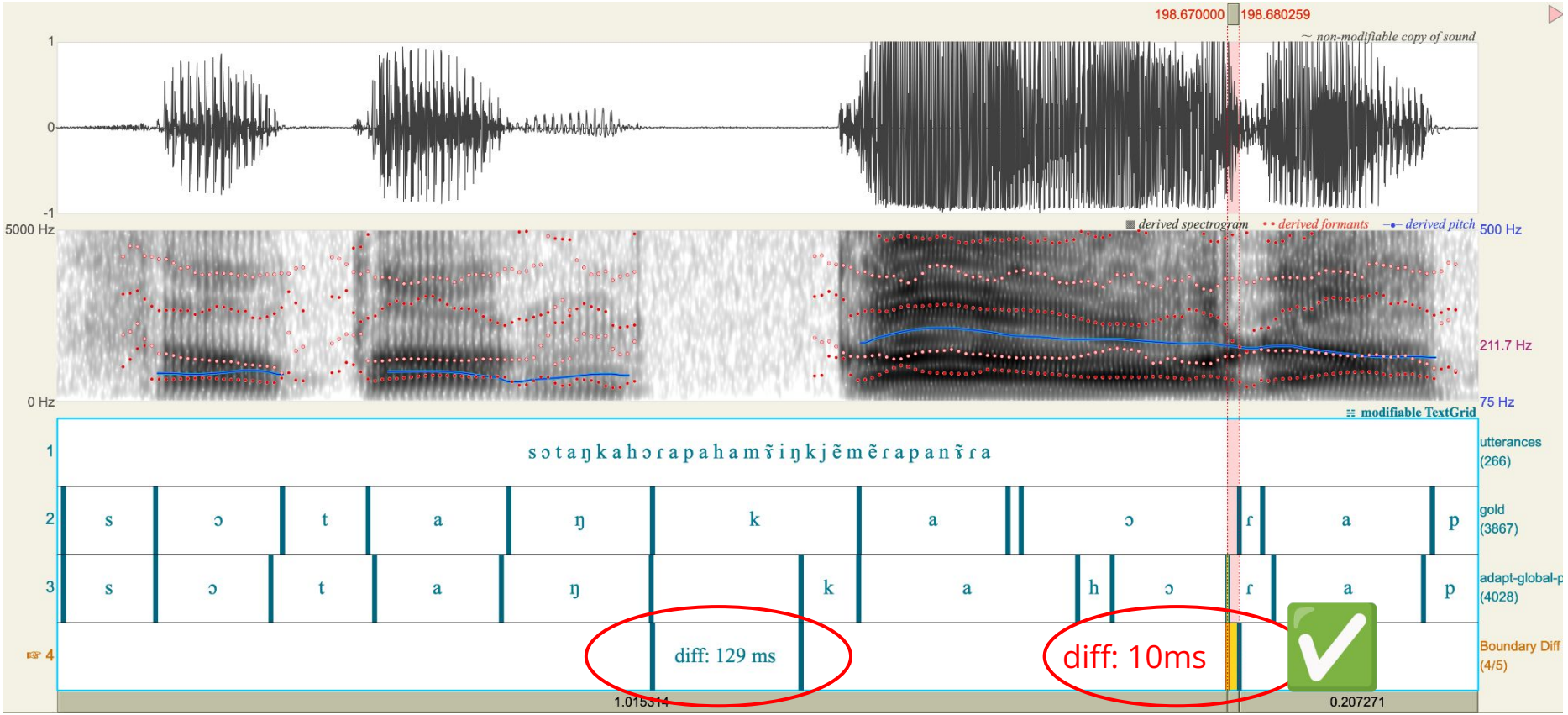
Methods:
Evaluation: Precision

Phone onset boundary precision =

% of system onsets within 20 milliseconds of the corresponding gold onsets (higher ⬆️ is better)

*or: **agreement** between human and system onset boundaries*

(McAuliffe et al., 2017; MacKenzie & Turton, 2020)



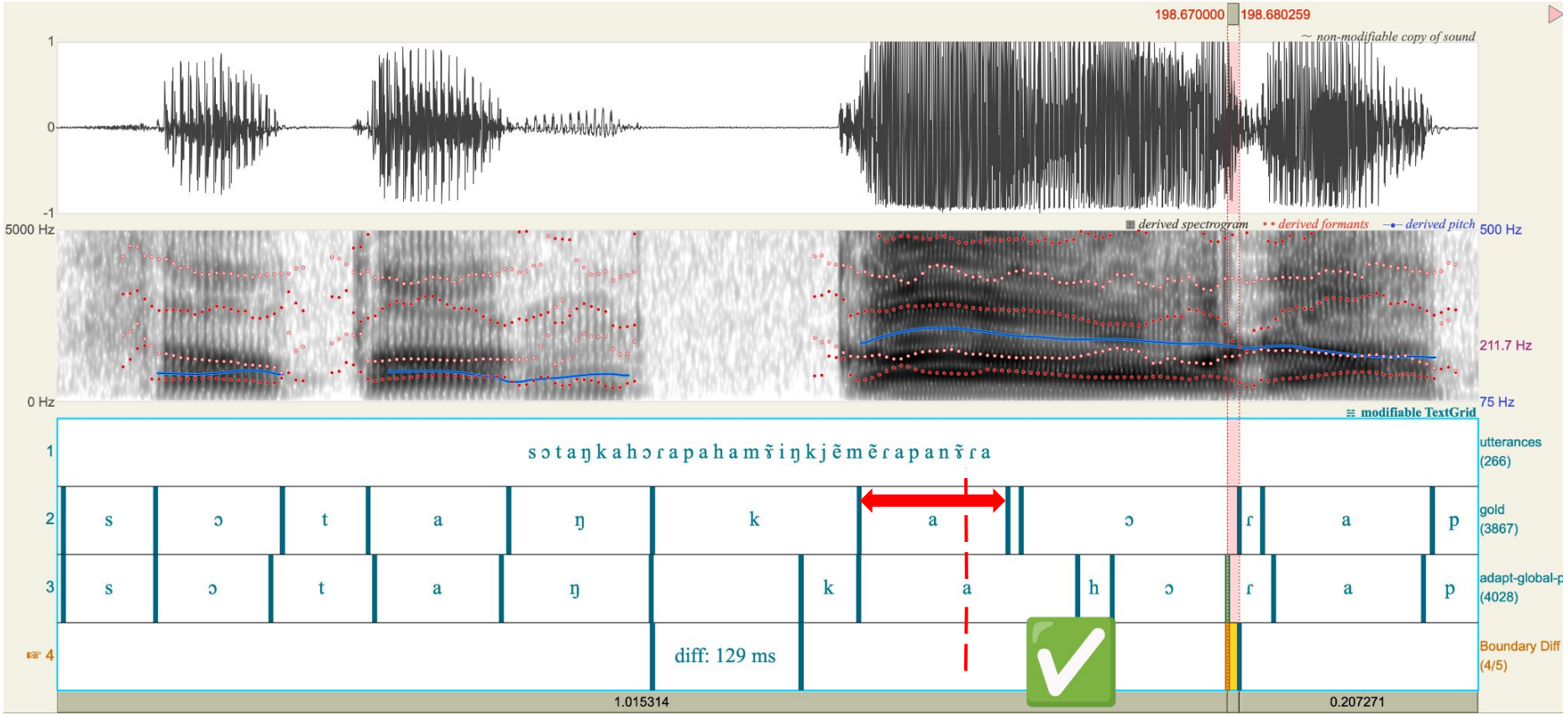
Example of Precision in boundary difference calculations in a Panãra audio file

Methods:
Evaluation: Accuracy

Phone interval accuracy =

% of system midpoints that lie within their corresponding gold intervals (higher ⬆️ is better)

(Knowles+ 2018; Chodroff+ 2024)



Example of Accuracy calculations in a Panãra audio file

Results (RQ 1):

Does including Russian CS data in training help alignment of target Urum data?

Training-from-scratch:

1) Keeping training quantity equal, CS model performs worse than Urum-only model

Precision % (<20ms)

	Train-from-scratch
Urum (47m)	**63.2**
CS (47m)	58.2

Accuracy % (test midpoint w/in gold interval)

	Train-from-scratch
Urum (47m)	**80.6**
CS (47m)	77.2

Results (RQ 1):

Does including Russian CS data in training help alignment of target Urum data?

Training-from-scratch:

1) Keeping training quantity equal, CS model performs worse than Urum-only model
2) Aggregating Urum + CS in training performs the best

Precision % (<20ms)

	Train-from-scratch
Urum (47m)	63.2
CS (47m)	58.2
Urum + CS (94m)	**70.9**

Accuracy % (test midpoint w/in gold interval)

	Train-from-scratch
Urum (47m)	80.6
CS (47m)	77.2
Urum + CS (94m)	**85.1**

Results (RQ 1):

Does using a pretrained Russian (or English) model and adapting on Urum/CS data help alignment of target Urum data?

Pretrained models:

1) Russian MFA performs the best
   a) better than English MFA
2) Keeping quantity equal, adapting on CS only is worse than on Urum-only

Precision % (<20ms)

	Train-from-scratch	Eng MFA	Russ MFA
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 % (test midpoint w/in gold interval)

	Train-from-scratch	Eng MFA	Russ MFA
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**

Research Question 2

[Meta question]

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

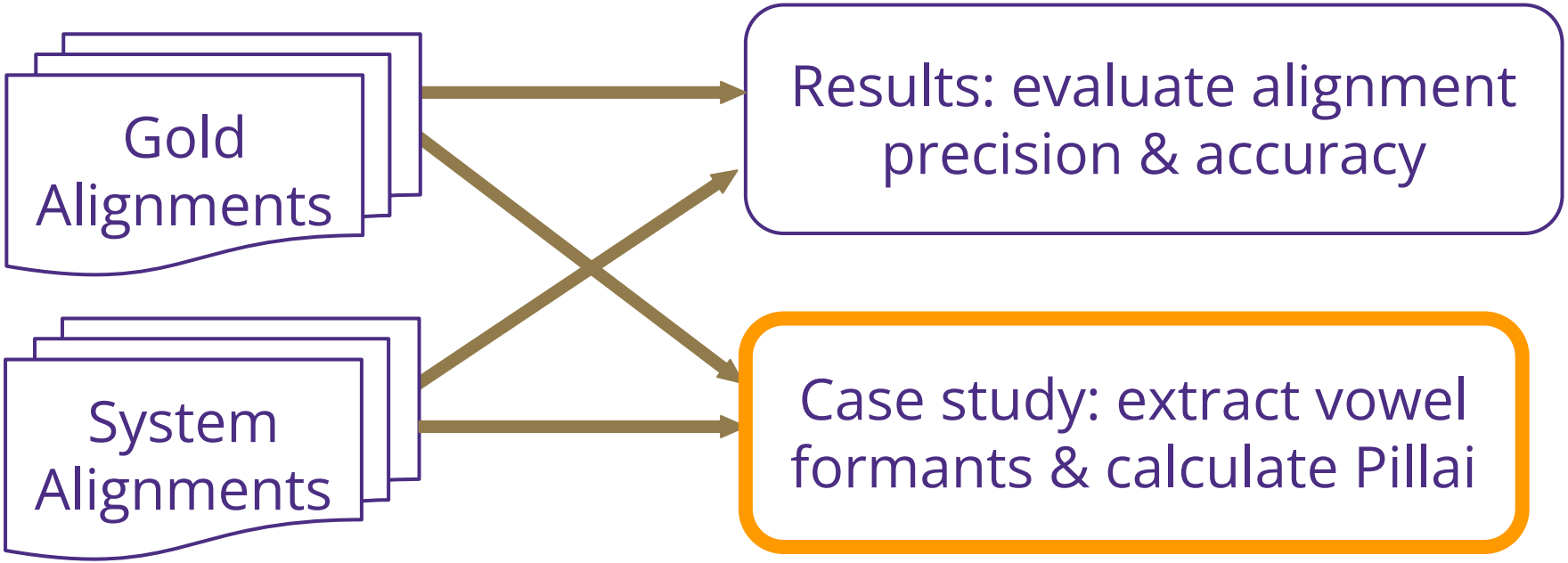
Or: to what degree are we comfortable substituting an automatic alignment for manual alignment, in our quest to answer a question about code-switching phonetics?

Case Study: Specific research question

Are vowels in Urum words pronounced differently in monolingual Urum utterances vs in CS utterances?

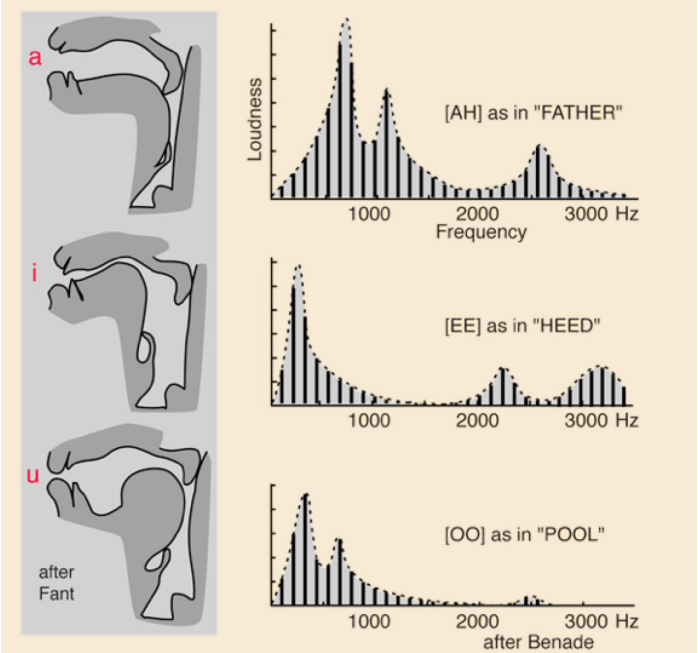
1. **Answer this with gold (manually annotated) test data**
2. **Compare results from best and worst system alignments to gold alignments**

Methods Overview 2



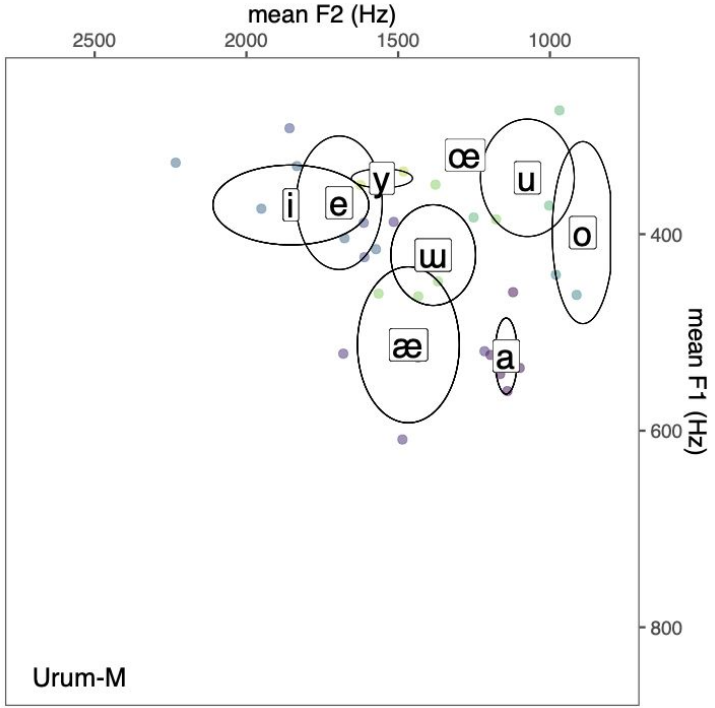
Case Study
Background: Vowel Formant Extraction

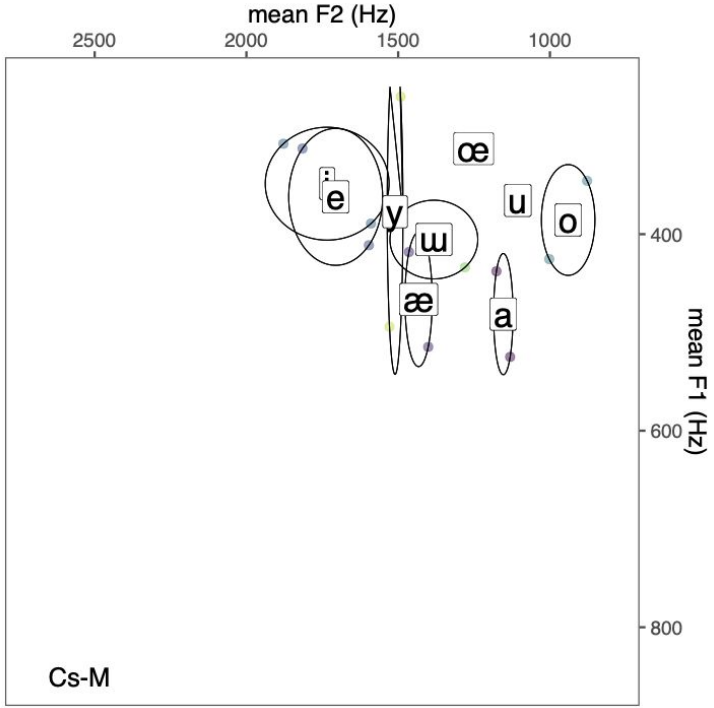
- Formants: high acoustic energy regions that reflect resonant frequencies in the vocal tract
- Algorithm/Tool: Linear Predictive Coding (LPC) in Praat
  - 5 formants under 5000Hz & 5500Hz
  - averaged F1&F2 midpoint + 10ms before/after



http://hyperphysics.phy-astr.gsu.edu/hbase/Music/vowel.html

Case Study: Gold speaker-averaged plots for 2-5 male speakers





Case Study
Background: Pillai scores

Pillai-Bartlett trace:

- output from a MANOVA test, used for measuring overlap between two distributions across two dependent variables (e.g. F1 and F2)

We use thresholds to determine if Urum vowels overlap in production across Urum vs CS

- formula from Stanley & Sneller (2023) utilizes exact sample size

https://joeystanley.com/blog/a-tutorial-in-calculating-vowel-overlap/pillai_example.png

Case Study Results: Gold (manually annotated) data

Significant difference in F1/F2 between Urum and CS utt found in 4 vowels across 3 speakers

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

True Pos
True Neg
False Pos
False Neg

Case Study
Results: System data – “best” model

Russian MFA adapted on Urum + CS yields 3 TP, 3 FP, 1 FN

	VOWELS								
GOLD	**a**	**e**	**i**	**o**	**u**	**y**	**œ**	**æ**	**ɯ**
Male									
A01	X (n=163)								
A03	X (n=188)		X (n=151)	X					
Female									
A02									
A07	X								
B08								X (n=29)	
B11									
B16									

True Pos
True Neg
False Pos
False Neg

Case Study
Results: System data – “worst” model

CS-only model yields 2 TP, 3 FP, 2 FN

		VOWELS								
	GOLD	**a**	**e**	**i**	**o**	**u**	**y**	**œ**	**æ**	**ɯ**
	Male									
	A01	X (n=163)								
	A03			X (n=151)	X				X (n=108)	X (n=85)
	Female									
	A02									
	A07	X								
	B08									
	B11									
	B16									

True Pos
True Neg
False Pos
False Neg

Case Study Conclusion

Good

"Best" model: 3 TP, 3 FP, 1 FN
"Worst" model: 2 TP, 3 FP, 2 FN

- Automatic alignment from even the "best" model doesn't tell the same story as manual aligned output
- The "worst" model captures less than the "best"
  - reveals nuance to precision/accuracy metrics
    - "worst" model (CS-only): 58%/77%
    - "best" model (Russian MFA) 71%/85%

Conclusion: Summary

1. **When aligning the target language, Urum, utilizing Russian was sometimes beneficial**
    a. Including Russian/CS added more data to training, making alignment more robust
    b. Best: pretrained Russian model adapted on all Urum/CS
2. **Automatic alignments may still need hand-correcting**
    a. evidenced by study of Urum vowel F1/F2 overlap across utterances

Conclusion: Future Directions

- **Analyze [code-switched] field data where transcription is limited/unavailable**
  - Can we utilize Automatic Speech Recognition for higher-resourced languages?

- **Dive deeper into how alignment quality affects different phonetic measures**

References

Chodroff, E., Ahn, E. P., & Dolatian, H. (2024). Comparing language-specific and cross-language acoustic models for low-resource phonetic forced alignment. *Language Documentation & Conservation*.

Knowles, T., Clayards, M., & Sonderegger, M. (2018). Examining factors influencing the viability of automatic acoustic analysis of child speech. *Journal of Speech, Language, and Hearing Research.*

MacKenzie, L., & Turton, D. (2020). Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard*, 6(s1).

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Interspeech*.

McAuliffe, M., & Sonderegger, M. (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.

McAuliffe, M., & Sonderegger, M. (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.

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., & Levin, L. (2016). Panphon: A resource for mapping IPA segments to articulatory feature vectors. In *COLING*.

Pandey, A., Gogoi, P., & Tang, K. (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*.

Paschen, L., Delafontaine, F., Draxler, C., Fuchs, S., Stave, M., & Seifart, F. (2020). Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). In *LREC*.

Skopeteas, S., Moisidi, V., Tsetereli, N., Lorenz, J., & Schröter, S. (2024). Urum DoReCo dataset. In Seifart, Frank, Ludger Paschen and Matthew Stave (eds.). Language Documentation Reference Corpus (DoReCo) 2.0.

Stanley, J. A., & Sneller, B. (2023). Sample size matters in calculating Pillai scores. *Journal of the Acoustical Society of America*.

Questions / Feedback?

eahn @ uw . edu

To my Phon Lab audience:

1) What parts of this work interest you most?
2) How would you handle code-switched data?

The End

UNIVERSITY *of* WASHINGTON