

One size does not fit all: Tailoring the Montreal Forced Aligner (MFA) to your data

Michael McAuliffe and Kaylynn Gunter

July 16, 2025

Welcome!

Goals and schedule

- Introduction: Your MFA Toolbox
 - Technical overview of MFA
 - The essentials of alignment - mfa align
 - Leveling up - mfa adapt
- The basics: Levers for improving alignment
- Advanced techniques
 - Part 1: Resourced language example - English
 - Part 2: Adapting models to novel varieties
 - Part 3: Languages without MFA support
- Lunch (12pm-2pm)
- Afternoon session (2pm-5pm)
 - Bring-your-own-data
- <https://github.com/mmcauliffe/mfa-adaptation>

Introduction

- Speech-to-text Alignment
 - Goal is to match the orthographic representation of speech to the time it was spoken in the audio file
 - Sequence-to-sequence problem: map the sequence of audio to the sequence of orthographic text
 - We assume the text represents exactly what is spoken (i.e., the ground truth)

Montreal Forced Aligner

- A command line utility for performing forced alignment
- Converts audio transcriptions to phone and word intervals
- Pretrained models and dictionaries for 20+ languages
- Still gaps in coverage

So what can you do? How can you get quality alignments for your data?

Training vs. Target data

- Training data: used to train model
 - Inherent variability of acoustic data vs.
 - Modeling the variability (e.g., dialect-specific lexicons, phone sets, etc.)
- Target data: the dataset for which you're predicting phone labels
- Training data variability is typically a super-set of target data variability

Key concepts

For MFA training:

- Training data is known, represents a wide variety of input sources
- Target data is unknown

For today:

- Target data is known
- Adapting data (i.e., additional training data) and target data may be the same

Key question for today:

- How well does your target data map to the training data?
- We will demonstrate a few ways to experiment with MFA

Technical overview of MFA

- Progressively more complex models
 - Monophone (Context independent phones)
 - Triphone (Context dependent phones)
 - LDA triphone
 - Feature transform based on phones
 - Speaker-adapted triphone
 - Feature transform based on speakers
 - Pronunciation probabilities
 - How likely is a given pronunciation?
 - How likely is silence to precede/follow a given pronunciation?

- Progressively larger subsets
 - 10k utterances
 - 20k utterances
 - 50k utterances
 - 150k utterances (optional)
 - All utterances
- Subset logic has evolved over versions
 - Split between dialect dictionaries
 - Avoid speakers with single utterances
 - New in 3.3: Manually aligned utterances always included
- New in 3.3: Subset from sub-corpora
 - Prioritize read speech corpora earlier
 - Incorporate noisy corpora like CommonVoice later

**So you have a pretrained model,
now what?**

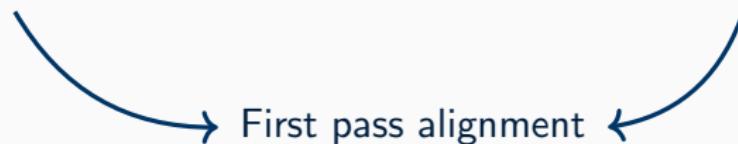
MFA alignment pipeline - First pass

Speaker-independent features

- Audio 
- Acoustic model

Decoding graph

- Transcript
- Pronunciation dictionary



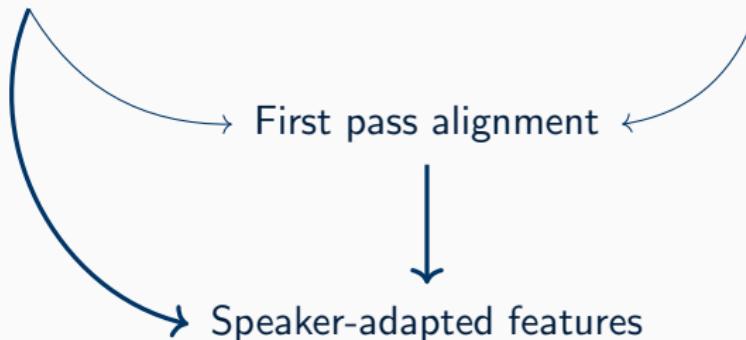
MFA alignment pipeline - Speaker adaptation

Speaker-independent features

- Audio 🔊
- Acoustic model

Decoding graph

- Transcript
- Pronunciation dictionary



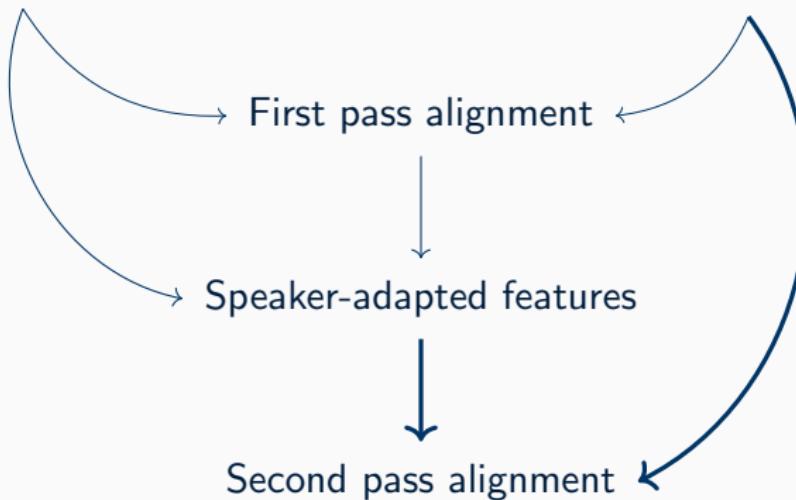
MFA alignment pipeline - Second pass

Speaker-independent features

- Audio 
- Acoustic model

Decoding graph

- Transcript
- Pronunciation dictionary



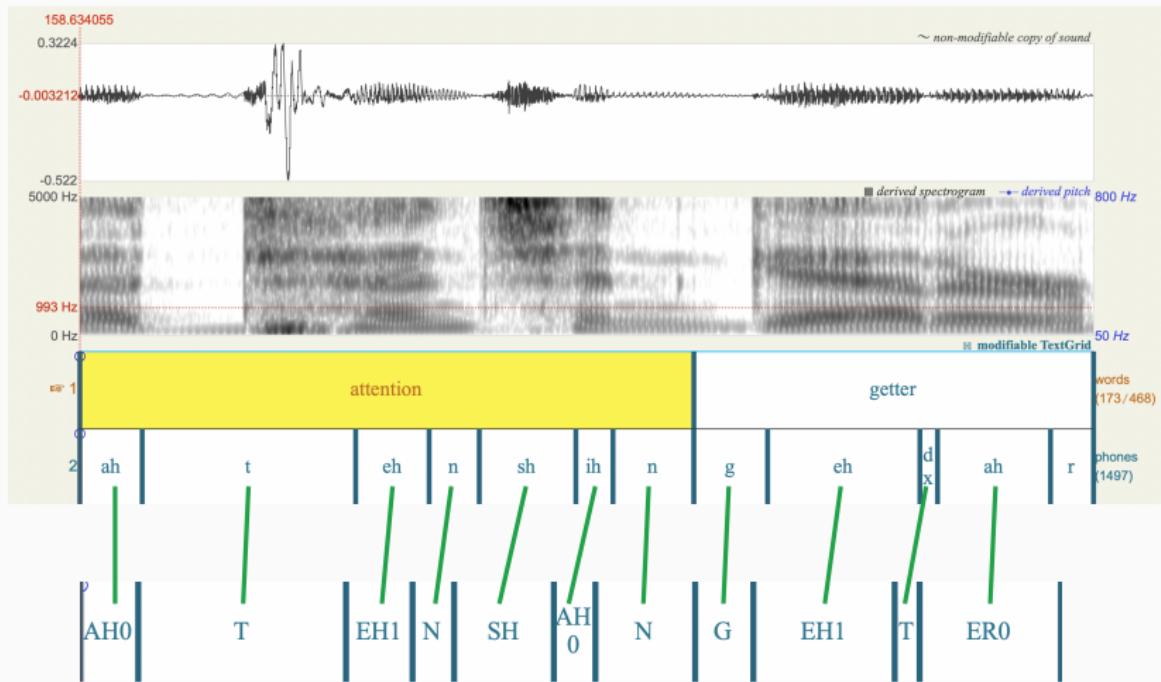
Evaluation

- Now you have alignments from the second pass
- But...how do we know the quality of alignments?
- Create an evaluation dataset with hand-corrected alignments
 - Evaluation metrics with reference alignments (`mfa align --reference_directory`)
- This can help us:
 - Select between different methods (`align`, `adapt`, etc.)
 - Select between different models (English vs. Mandarin)
 - Identify problematic utterances
- So, how do we do that?

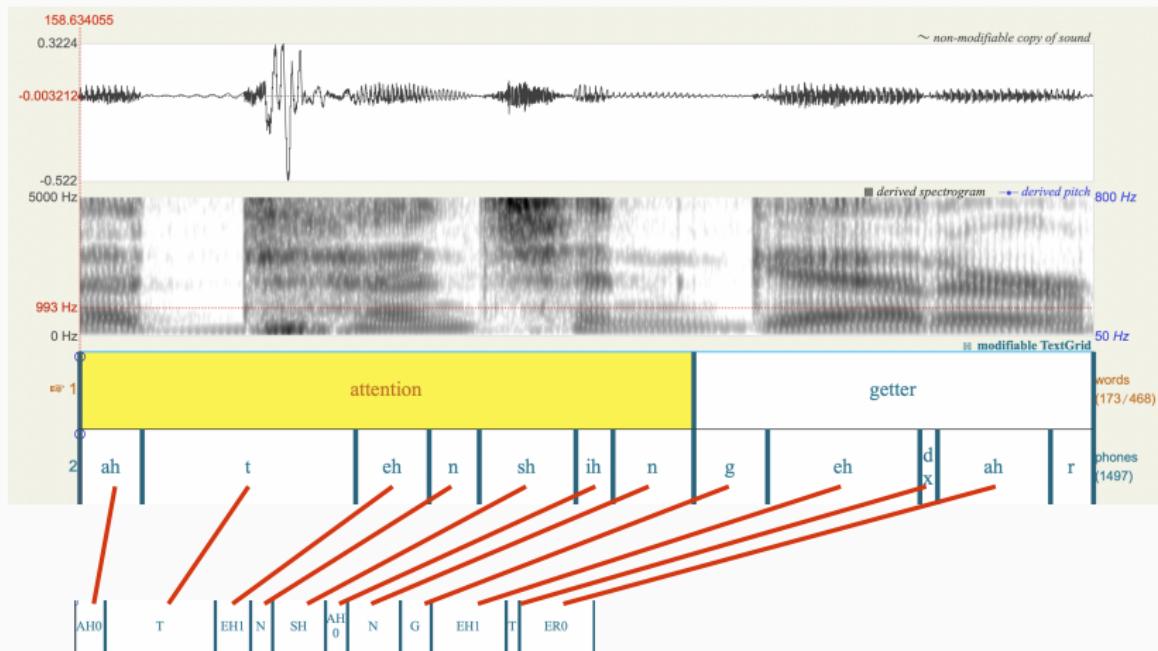
Evaluating alignments

- How can we quantify how “good” a model aligns a corpus?
- Typical metric: boundary recall within 20ms
 - For every boundary in reference alignments:
 - Is there a corresponding aligned boundary within 20ms window?
- Works ok when reference alignment is the same sequence of phones, but
 - What if there are pronunciation variants?
 - What if the evaluation data set is using a completely different phone set?
- MFA has two metrics based on an alignment of alignment intervals
 - Helps ensure that boundaries being compared are the same

Evaluating alignments - Pretty good alignment



Evaluating alignments - Pretty bad alignment



Evaluating alignments

- Levenshtein-esque alignment between reference intervals and aligned intervals

$$\text{Overlap Cost} = -1 * \left(|begin_{aligned} - begin_{ref}| + |end_{aligned} - end_{ref}| + \begin{cases} 0, & label_1 = label_2 \\ 2, & otherwise \end{cases} \right)$$

- Alignment score is the average overlap cost ignoring insertions/deletions

$$\text{Phone Error Rate} = \frac{insertions + deletions + (2 * substitutions)}{length_{ref}}$$

The essentials of adaptation

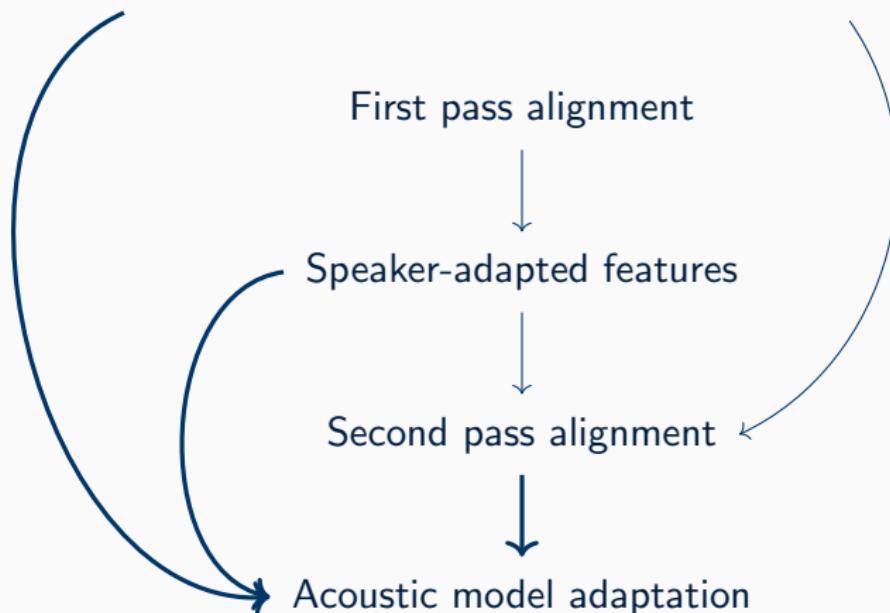
MFA adaptation pipeline

Speaker-independent features

- Audio
- Acoustic model

Decoding graph

- Transcript
- Pronunciation dictionary

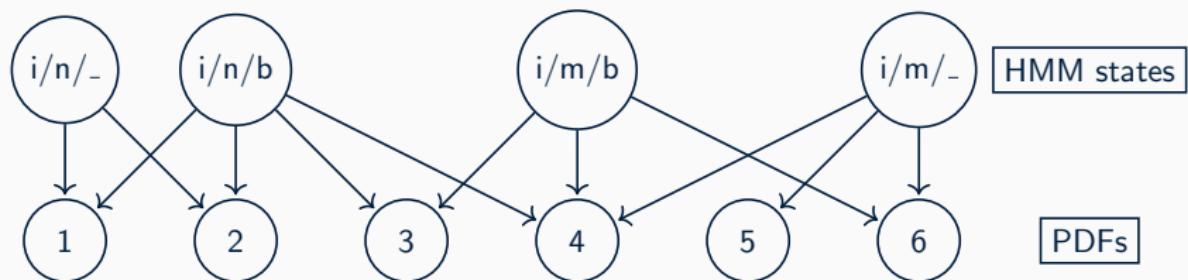


The essentials of alignment - mfa adapt

- Key point:
 - Adaptation uses automatic alignments
 - If alignments are poor quality, adaptation will be poor
- Reference alignments
 - New in 3.3!
 - Can use corrected alignments to guide all alignments
 - Implementation: penalize acoustic states that are not associated with the correct phone

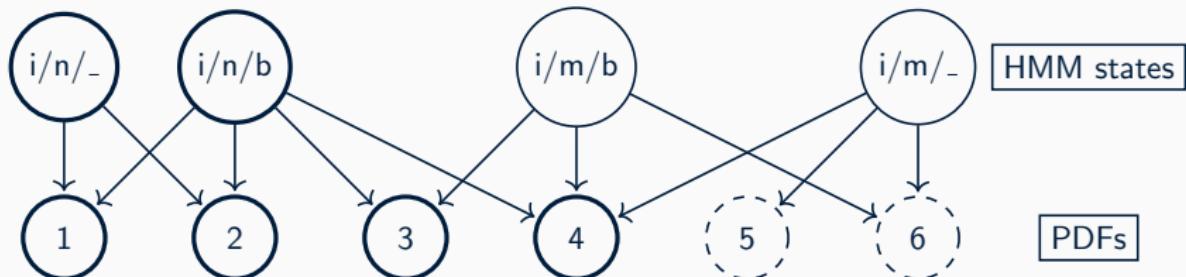
Schematic of phone clustering

- Three-phone sequence space is huge
- Cluster similar acoustic probability density functions (PDFs) across different phones
- Goal: capture allophonic/coarticulatory information
 - green beans
 - gr[i m b]eans?



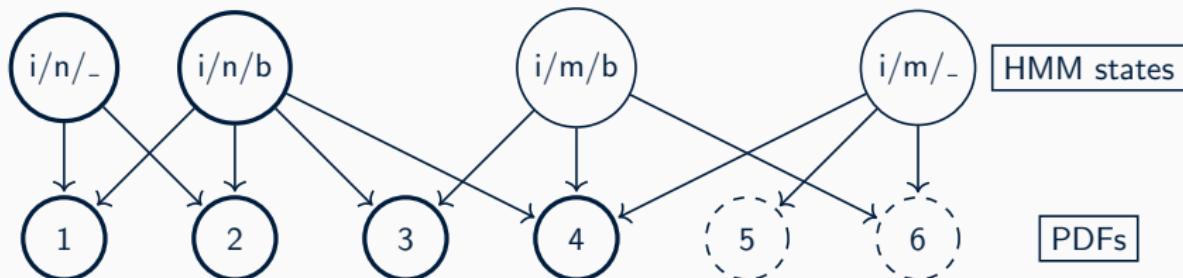
How reference phones are used in align

- PDFs give log-likelihood based on acoustic features
- If there's a reference phone
 - If the PDF is mapped to that phone, return normal log-likelihood
 - If the PDF is not mapped to that phone, return extremely unlikely log-likelihood



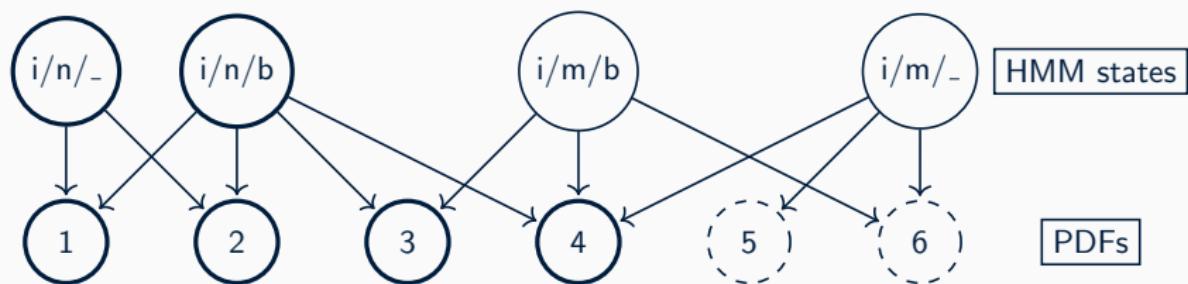
How reference phones are used in align

- PDFs give log-likelihood based on acoustic features
- If there's a reference phone
 - If the PDF is mapped to that phone, return normal log-likelihood
 - If the PDF is not mapped to that phone, return extremely unlikely log-likelihood



How reference phones are used in adapt

- Only PDFs that are used in alignments are adapted
- More data → better adaptation
- More variety → better adaptation



The basics: levers for improving alignment

Key factors

- Data Quality
 - Transcription accuracy (Description is analysis)
 - Out of Vocabulary items (OOVs)
- Data Quantity
 - More is better!
 - Variability (phones, styles, speakers)
- Target data similarity to Training data
- Experimentation - Try different things out!

Dictionary: OOVs

- Why do OOVs matter?
 - We want more instances a phone across different states
 - OOVs mean fewer examples of the phone(s)
 - Aligning OOV can cause errors downstream
- Add pronunciations
 - Target the ones with high Ns
 - Diminishing returns on single tokens

Transcripts: split long utterances

- Errors early on can be recovered from, can affect later alignments
- Split longer segments of transcripts (> 1min)
- Use breath groups to segment
- When certain utterances have poor alignment (e.g., from OOVs) splitting the utterance into smaller segments

Variation and Speakers

- Data quantity and variation matters
- More data for a single speaker
 - Elicitation style - word lists and conversations
 - More data - phone environments, tokens, etc.
- More data across speakers
- Give separate speaker IDs for stylized speech (e.g., whispering, falsetto, story telling, etc.)

**Ok that was a lot of information,
let's take a 15 minute break before
case studies!**

Part 1 - MFA Supported Language

Overview

- Easiest use-case: Target data is similar to training data
- Inputs:
 - MFA supported language (English) acoustic model(s)
 - Speech corpus MFA format (audio, transcripts)
 - Dictionary
 - Manual reference alignments

Target data: Buckeye Corpus

- Buckeye Corpus (Pitt et al. 2007) with corrections, see data prep scripts¹
- Conversational speech w/ ground truth for benchmarking alignment
- 40 speakers
 - all white, 20M/20F, distributed across age, Columbus OH

¹https://github.com/MontrealCorpusTools/mfa-models/tree/main/scripts/alignment_benchmarks/data_prep

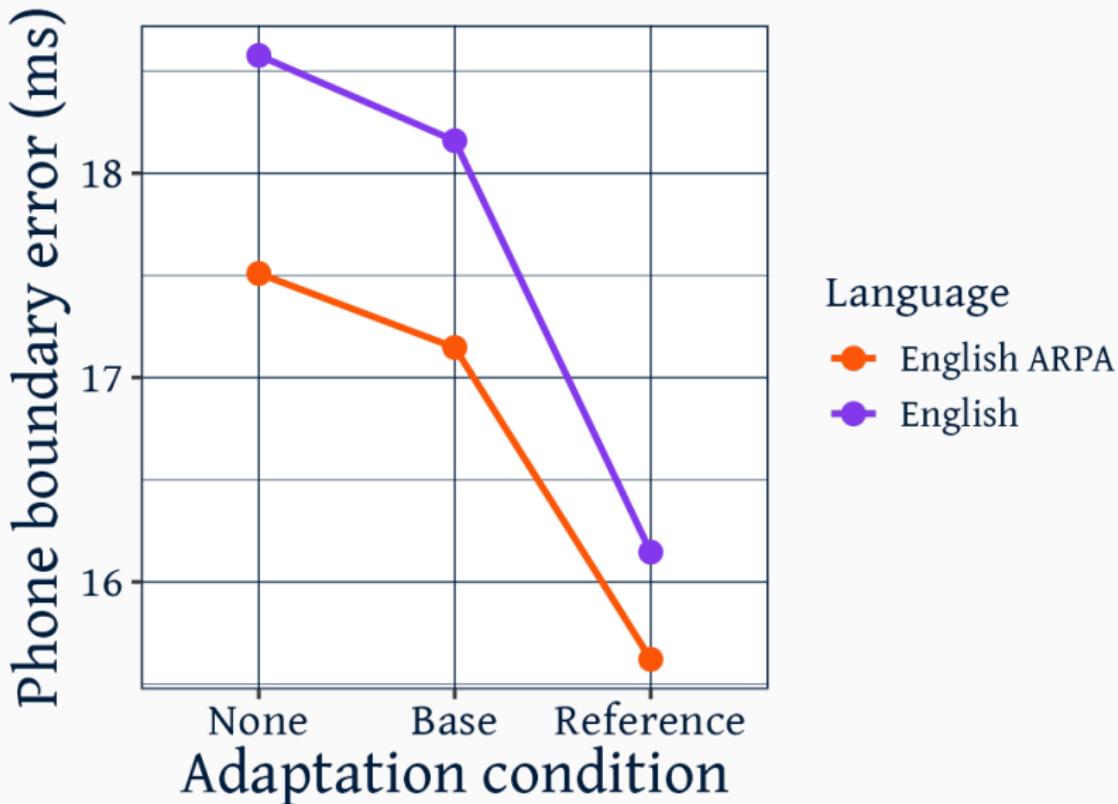
Training data

Model	Variety	Style	Hours	Speakers
English ARPA ^a	General US English	Read Speech	982	2,484
English MFA ^b	Multi-dialectal, Global English, L2 Speech	Conversational, Read Speech	3,771	78,975

^aEnglish (US) ARPA docs

^bEnglish MFA docs

Quality of alignments



Takeaways

- Adaptation
 - Alignments with English models already pretty good (18ms)
 - Running adapt still improves alignments *a little* (0.4ms)
 - Having oracle adaptation alignments improves them more (2ms)
- English ARPA outperforms English MFA
 - English ARPA source data is the same variety as target
 - English MFA source data is more, but global varieties

Part 2 - Novel Speech Variety

Overview

- Use case: Language is supported in MFA, but the variety of the target data is not represented in the model
- Inputs:
 - MFA supported language (English) acoustic model
 - Speech corpus MFA format (audio, transcripts)
 - Dictionary
 - Reference alignments (WhisperX)

Target data: VariCS Corpus

- Variation in Child Speech (VariCS)²
 - Recordings from 275 children speech
 - Age 5-12 years from Scotland, UK
 - Recorded in the primary schools, noisy environments etc.
 - Includes several tasks
 - For today, looking at the picture naming
 - because children, not entirely controlled but not fully spontaneous either
 - Transcribed by WhisperX - no adjustments made to transcript

²Christodoulidou et al., submitted

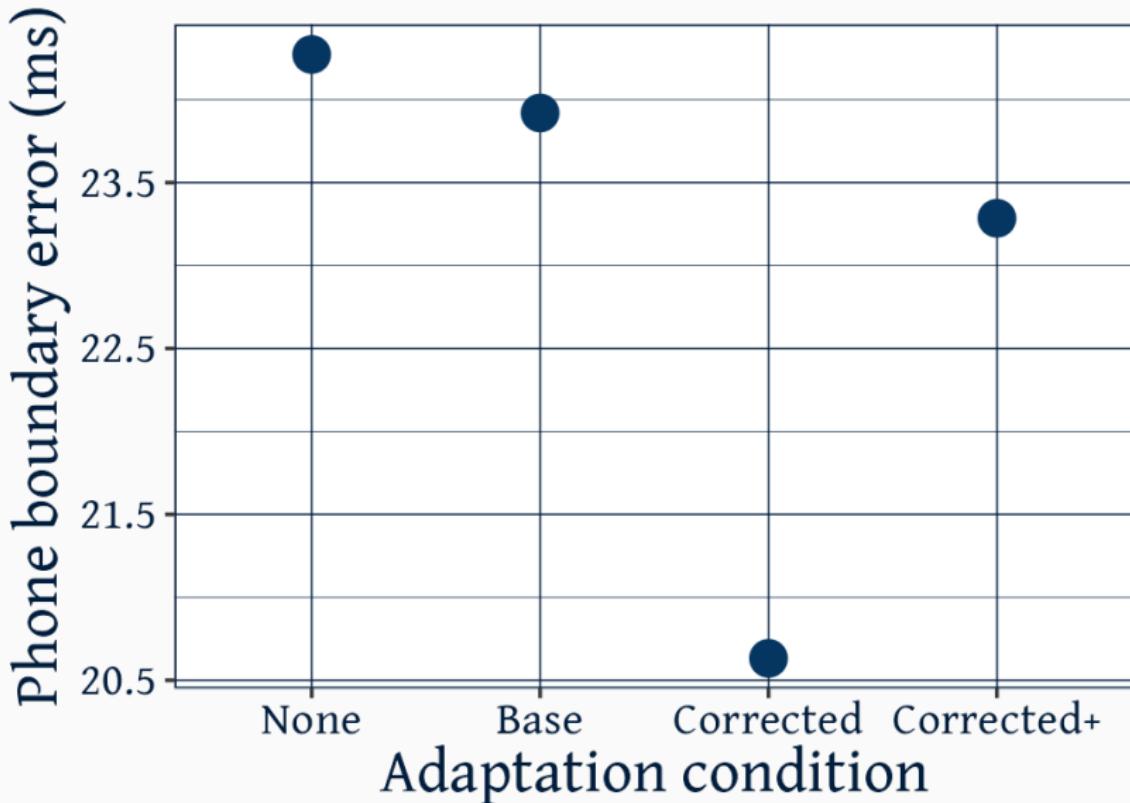
Target data: Evaluation subset

- Similar to Christodoulidou et al. (submitted)
- Eval set is 9 manually corrected files from 5 children
 - 9 manually corrected files from 5 children
 - 48 utterances per child
- Adaptation data sets
 - Base: 9 files from eval set without using manual alignments
 - Corrected: 50 manually corrected files
 - Corrected+: Corrected files plus all other uncorrected files

Training Data

- English US ARPA model
 - Different Target and Training Data
 - US → Scottish English
 - Adult Read Speech → Spontaneous-ish Children's Speech

Quality of alignments



Takeaways

- Generally larger error than for Buckeye (24ms vs 18ms)
- Adaptation
 - Base adaptation doesn't help that much (0.4ms)
 - Corrected helps the most (3.7ms)
 - Adding noisy adaptation data reduces gains (1ms)
 - More data isn't always better

Part 3 - Non-supported Language

Overview

- Use Case: There is no current language support for the target data
- Inputs
 - Speech corpus MFA format (audio, transcripts)
 - Dictionary (in Target language)
 - MFA Acoustic models of varying similarity to target data

Target data: Buckeye Corpus

- Suspend disbelief, Buckeye represents the under-resourced variety
- Allows for comparisons across adapting with source languages that vary in degree of phone similarity
- English ARPA → English MFA → German → Czech → Mandarin

Training Data

Model	Variety	Style	Hours	Speakers*
English ARPA ^a	General US English	Read Speech	982	2,484
English MFA ^b	Multi-dialectal, Global English, L2 Speech	Conversational, Read Speech	3,771	78,975
German MFA ^c	German	Read Speech	1,315	18,012
Czech MFA ^d	Czech	Spontaneous, Read Speech	643	1,261
Mandarin ^e	Multi-dialectal	Read Speech	612	7,316

^aEnglish (US) ARPA docs

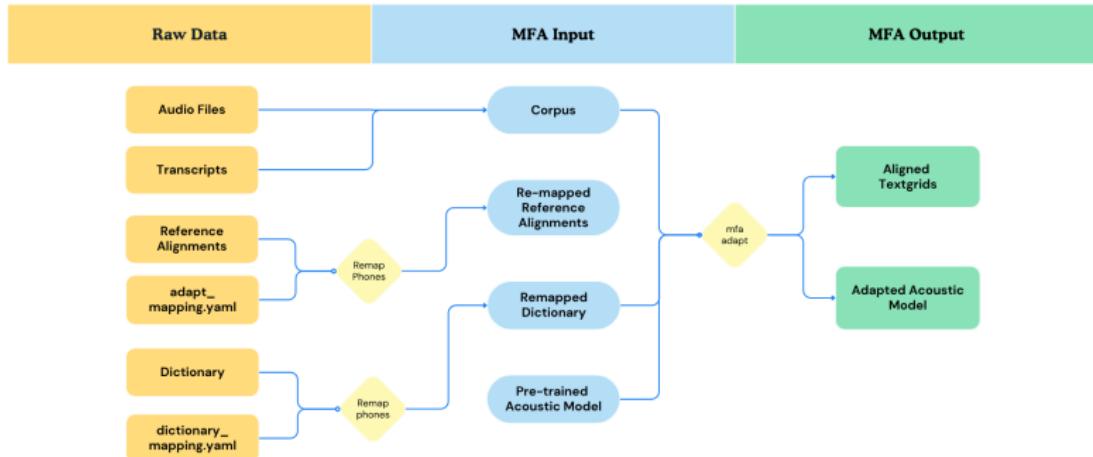
^bEnglish MFA docs

^cGerman MFA docs

^dCzech MFA docs

^eMandarin MFA docs

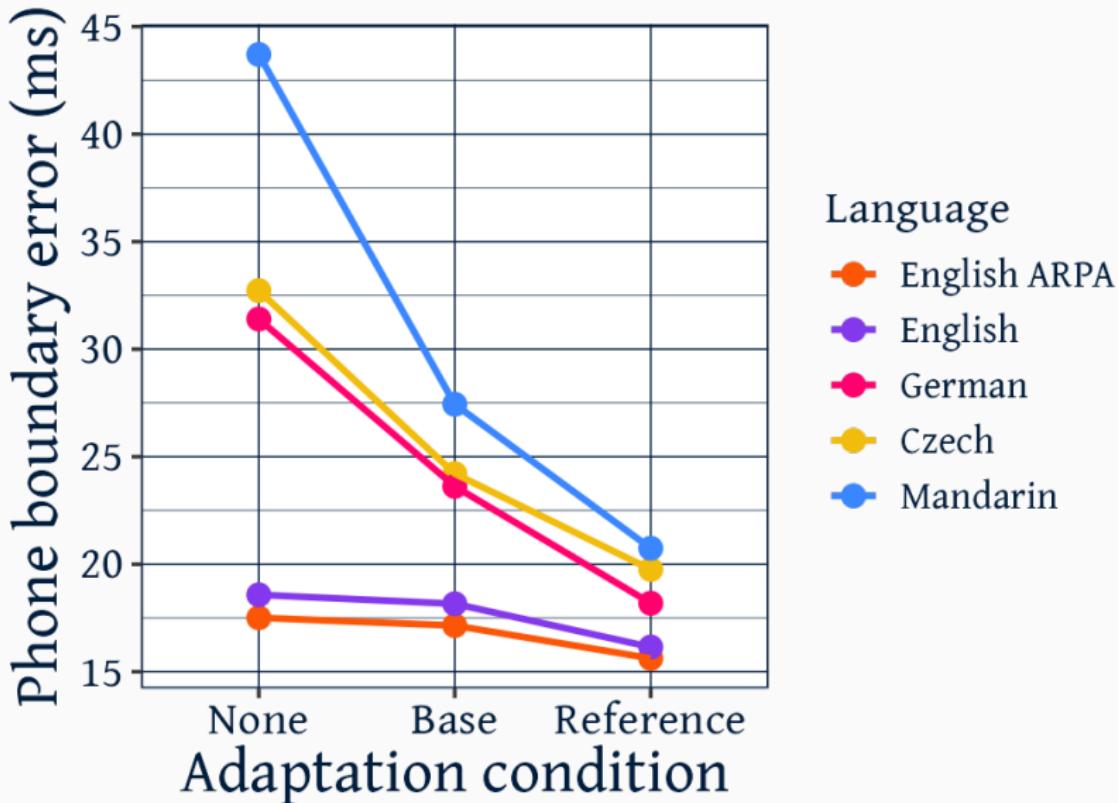
Methods



Target vs. Training Phones

Model	English → Training Phones
English ARPA	$[p^h, p^j, p^w] \rightarrow [P]$ $[m] \rightarrow [AH0 M]$
German MFA	$[ð] \rightarrow [d, v]$ $[θ] \rightarrow [t, f]$ $[ɔj] \rightarrow [ɔY, ɔ j]$ $[r] \rightarrow [R, e]$
Czech MFA	$[ej] \rightarrow [ɛ:, ɛ j]$ $[aj] \rightarrow [a j]$ $[ð] \rightarrow [d, v]$ $[θ] \rightarrow [t, f]$ $[r] \rightarrow [r, ə]$
Mandarin MFA	$[ʈʂ] \rightarrow [tʂ, tʂ, ts]$ $[ɳ] \rightarrow [ɳɿ]$ $[ʂ] \rightarrow [ʂ, ʂ, s]$ $[ɿ] \rightarrow [ɿ]$

Quality of alignments



Quality of alignments

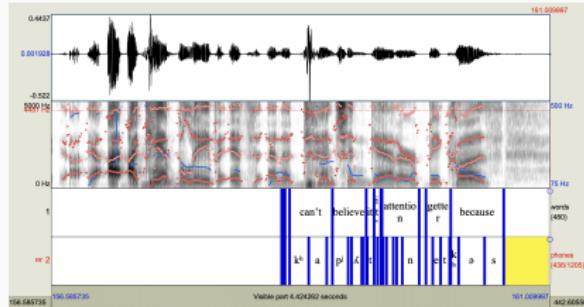


Figure 1: Mandarin - Align

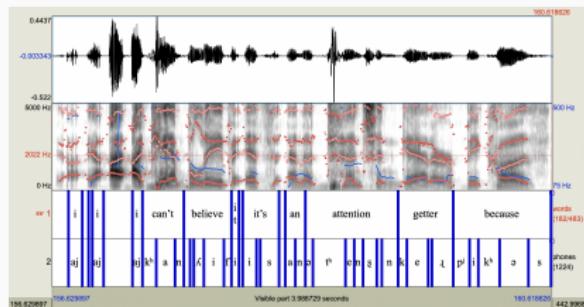


Figure 2: Mandarin - Adapt with reference

Quality of alignments

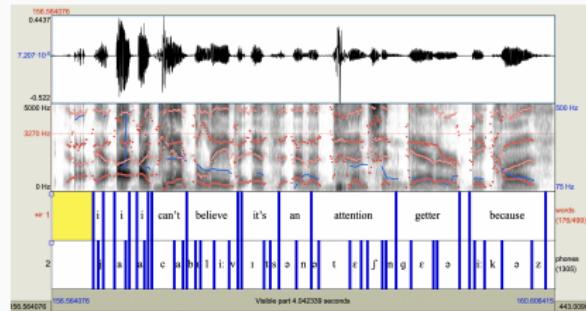


Figure 3: Czech - Align

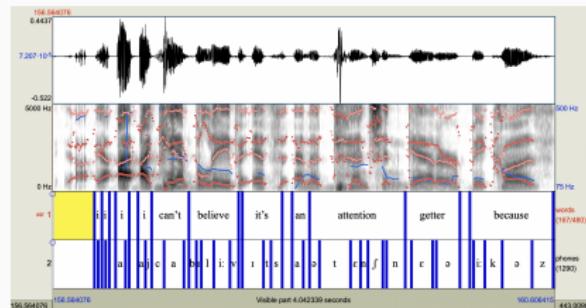


Figure 4: Czech - Adapt with reference

Quality of alignments

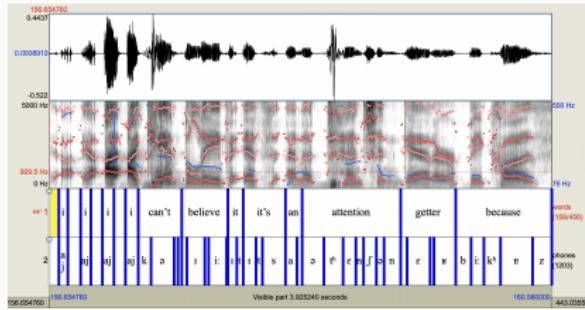


Figure 5: German - Align

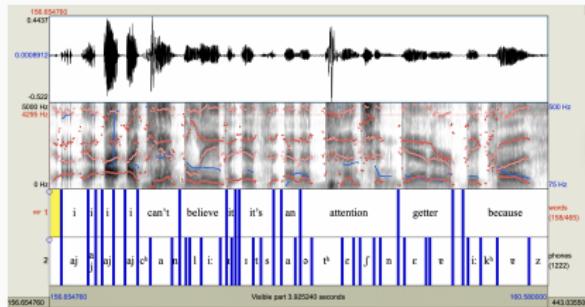


Figure 6: German - Adapt with reference

Takeaways

- Alignment performance follows similarity
 - English < German <= Czech < Mandarin
- Adaptation
 - Adaptation helps Mandarin the most (16ms)
 - Adaptation still shows gains for German (7.8ms) and Czech (8.5ms)
 - Base adaptation gives a majority of the gains
 - Mandarin: 70%
 - Czech: 66%
 - German: 59%

Final Thoughts

Final Thoughts

- Adaptation generally helps
 - Generally shows decreased boundary errors
 - Beware that these are mean measures
 - Still investigations to be done on errors
 - I cannot confidently say that adaptation will help ALL cases
- Aim for high quality data
- Consider how similar your target and training data is

Final thoughts

- Don't be afraid to experiment!
- Invest in creating small eval set
 - Can just be spot-checked good alignments
 - mfa align ... --include_original_text
 - Can spot-check Praat, Anchor
- Incorporate eval set into your final adaptation
 - No sense wasting high quality data
 - Lose evaluation on generalizing
 - Generalizability is not really a concern
 - You want the best alignments possible for a corpus, so it's fine to train on your eval data (eventually)

Align-a-thon at LabPhon 2026

- Ground truth alignments for English
 - Licensing issues
- Are some boundaries more important than others to phoneticians?
 - Silence to stop?
 - Stop to vowel?
 - Vowel to silence?

We'll be hosting an align-a-thon workshop at LabPhon 2026 in
Montreal next summer!

Hope to see you there!

Thank You!

This workshop is supported by



Canada
Research
Chairs

Chaires
de recherche
du Canada

Canada

Q&A
