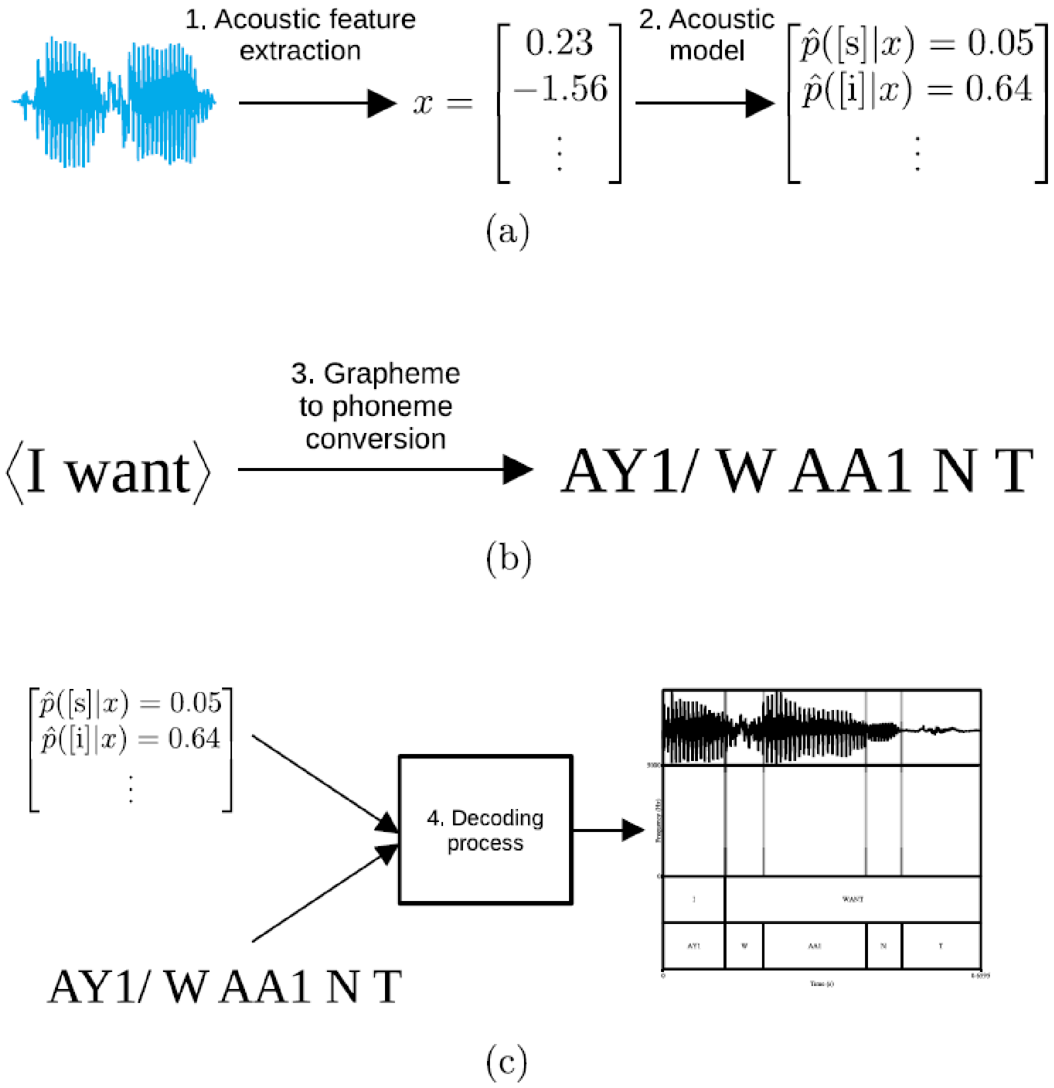
Using confidence intervals in forced alignment

MaTiPS, 18 Oct 2025

Matthew C. Kelley, George Mason University



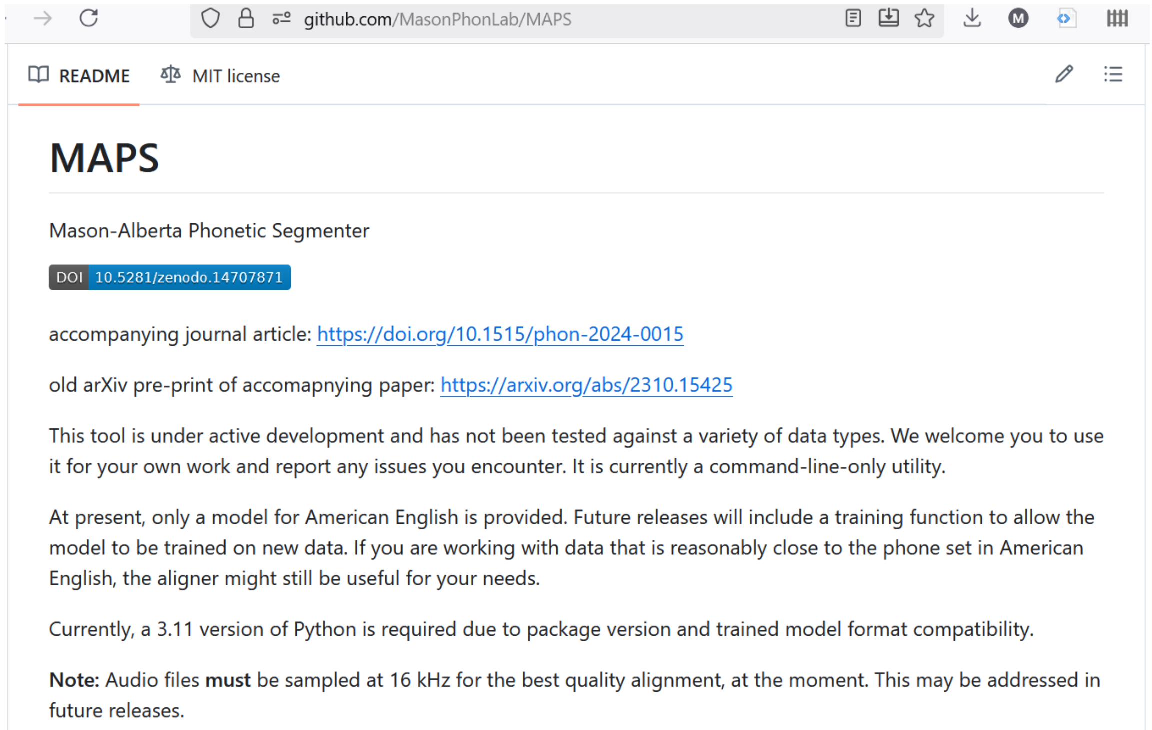
Forced alignment broadly

- Method for aligning audio and a transcription
- Various levels possible, including sentence, word, and phone(me)
- Notorious for occasional enormous errors

Forced alignment as a problem

- Conceptually, the forced alignment problem is set up as follows
- **Given** an acoustic model yielding conditional probabilities of a language's segments, an audio file, and a transcription
- **Produce** an optimally probable alignment between the transcription and audio file
- **Constrained** by every segment in the transcription being used in the correct order

The Mason-Alberta Phonetic Segmenter (MAPS)


github.com/MasonPhonLab/MAPS

README MIT license

MAPS

Mason-Alberta Phonetic Segmenter

DOI 10.5281/zenodo.14707871

accompanying journal article: https://doi.org/10.1515/phon-2024-0015

old arXiv pre-print of accomapnying paper: https://arxiv.org/abs/2310.15425

This tool is under active development and has not been tested against a variety of data types. We welcome you to use it for your own work and report any issues you encounter. It is currently a command-line-only utility.

At present, only a model for American English is provided. Future releases will include a training function to allow the model to be trained on new data. If you are working with data that is reasonably close to the phone set in American English, the aligner might still be useful for your needs.

Currently, a 3.11 version of Python is required due to package version and trained model format compatibility.

Note: Audio files **must** be sampled at 16 kHz for the best quality alignment, at the moment. This may be addressed in future releases.

- A new forced alignment system released last year (Kelley et al., 2024)
- Interpolates over alignment probabilities when placing boundaries
  - Overcomes 10 ms precision limit with boundary placement
- New feature: estimating uncertainty of boundary location
  - This is the subject for today's talk!

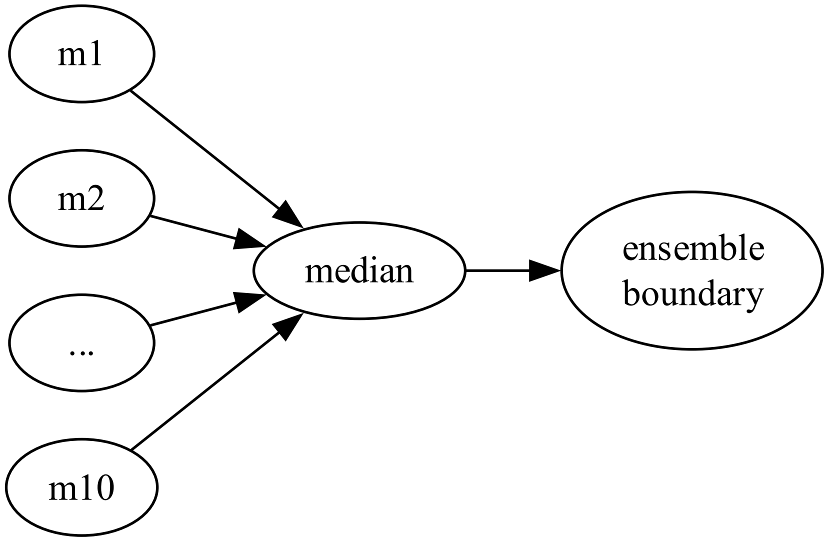
Forced alignment boundaries

- The boundaries a forced aligner places are estimates
- In the neural case (like MAPS):
  - Non-terminal boundaries are placed at the latest moment where cumulative sequence probability is greater for current segment than the following one
- The (neural) function yielding the probabilities is estimated over many learned parameters

Boundary uncertainty

- Since the boundaries are based on data-driven estimates, they have natural uncertainty surrounding their placement
- Estimating the uncertainty gives additional information about the placed boundaries
  - What timespan are we confident contains the "true" boundary?
- Pragmatically, the uncertainty can be used to flag potential boundary errors

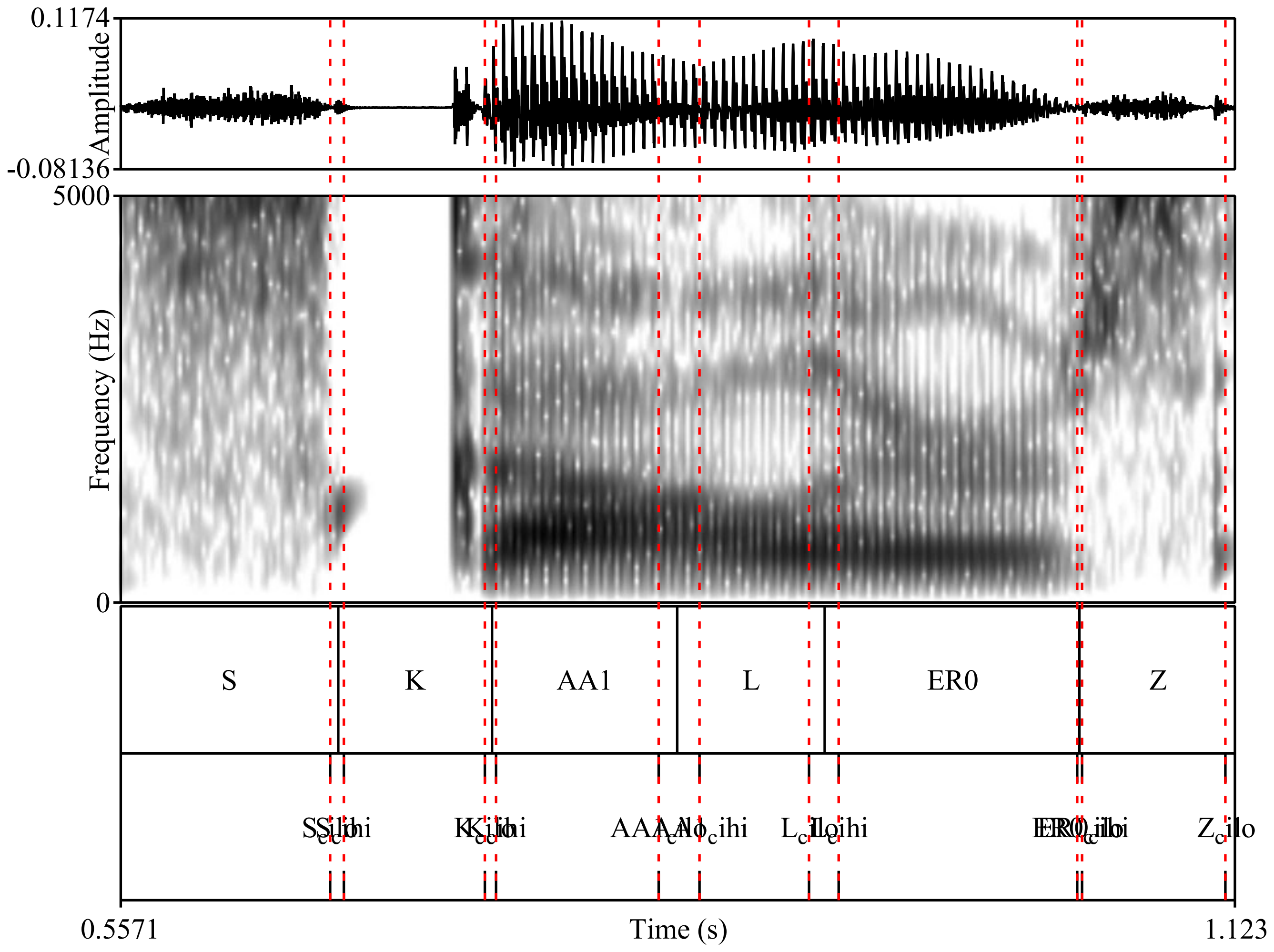
How to estimate uncertainty for neural networks



- Well… you can fit a Bayesian neural network
  - Neural networks already take a while to train
- Instead, I used a 10-network ensemble to produce multiple estimates
- Place boundary at median for robustness to outliers
  - Use 2nd lowest and 9th highest boundaries as confidence interval edges (yields 97.85% confidence)

Sample alignment

- Sample CIs for scholars [ˈskɑlɚz] from *...blical scholars arg...* In "biblical scholars argue".
- High and low values for the CIs are drawn in dashed red lines.



Data output

- Praat TextGrid data format hard to work with for CI data
- As such, a table is also output from the alignment with the CIs for statistical or programmatic analysis

A	B	C	D	E	F	G	H	I
file	word	word_mintime	word_maxtime	segment	segment_mintime	segment_maxtime	segment_lo_ci	segment_hi_ci
dark_suit_sentence_16khz_ensemble.TextGrid	sil	0	0.025	H#	0	0.025	0.02499999	0.0250000
dark_suit_sentence_16khz_ensemble.TextGrid	SHE	0.025	0.197852231899796	SH	0.025	0.140936898861931	0.13793531110899	0.14457111390484
dark_suit_sentence_16khz_ensemble.TextGrid	SHE	0.025	0.197852231899796	IY1	0.140936898861931	0.197852231899796	0.189997872467614	0.21440569893811
dark_suit_sentence_16khz_ensemble.TextGrid	HAD	0.197852231899796	0.467943509514459	HH	0.197852231899796	0.276315000557992	0.269984330880739	0.28010692708951
dark_suit_sentence_16khz_ensemble.TextGrid	HAD	0.197852231899796	0.467943509514459	AE1	0.276315000557992	0.401911378842295	0.398115363436123	0.40541860332451
dark_suit_sentence_16khz_ensemble.TextGrid	HAD	0.197852231899796	0.467943509514459	D	0.401911378842295	0.467943509514459	0.465	0.47
dark_suit_sentence_16khz_ensemble.TextGrid	YOUR	0.467943509514459	0.559147188283383	Y	0.467943509514459	0.485	0.475	0.49

Identifying segmentations that need further investigation

File	Word	Segment	Width (ms)
FAEM0SA1	YEAR	[ɹ]	24.3
FAEM0SI1392	EXAMPLE	[ɪ]	57.2
FAEM0SI1392	EXAMPLE	[g]	28.1

- A threshold can be used to identify segmentations that may be problematic
  - Here, using a 20 ms threshold
- Compared with duration, CI width can be specified absolutely
  - However, it is acoustic model dependent

Identifying difficult-to-segment bigrams

Bigram	Median width (ms)
[ij]	47.75
[ʃs]	42.79
[æa͡ɪ]	40.39
[ʊl]	36.54
[a͡ɪe͡ɪ]	36.31
[o͡ʊl]	35.81
[a͡ɪæ]	33.39
[ʌa͡ɪ]	33.15
[pθ]	32.87
[uɑ]	30.94

- Difficult-to-segment bigrams can be identified using median width
- These are 10 greatest CI widths for bigrams that appeared at least 25 times in TIMIT+Buckeye

Conclusion

- CI estimation for force-aligned boundaries now possible in MAPS
- Wide CIs can be used diagnostically to highlight alignment errors
- Perhaps we should push past non-overlapping, strictly ordered ways of representing acoustic events
  - I.e., neither interval nor point tiers are adequate for CIs in Praat
- Scientifically, it feels immoral to ignore uncertainty

Thank you!

References

- Kelley, M. C., Perry, S. J., & Tucker, B. V. (2024). The Mason-Alberta Phonetic Segmenter: A forced alignment system based on deep neural networks and interpolation. *Phonetica*, *81*(5), 451–508.