PD (and other things) since PhD

—

Work on speech timing and methodology

Pertti Palo

9 May 2024

Outline

- Introduction
- Ultrasound tongue imaging
- Pixel Difference (PD)
- PD and other methods applied
  - PD on de-interlaced lip videos
  - Metrics on tongue splines
  - PD on 3D/4D ultrasound
  - PD on Raw vs Interpolated 2D data
  - Choosing the norm or metric for PD
- MRI

Introduction

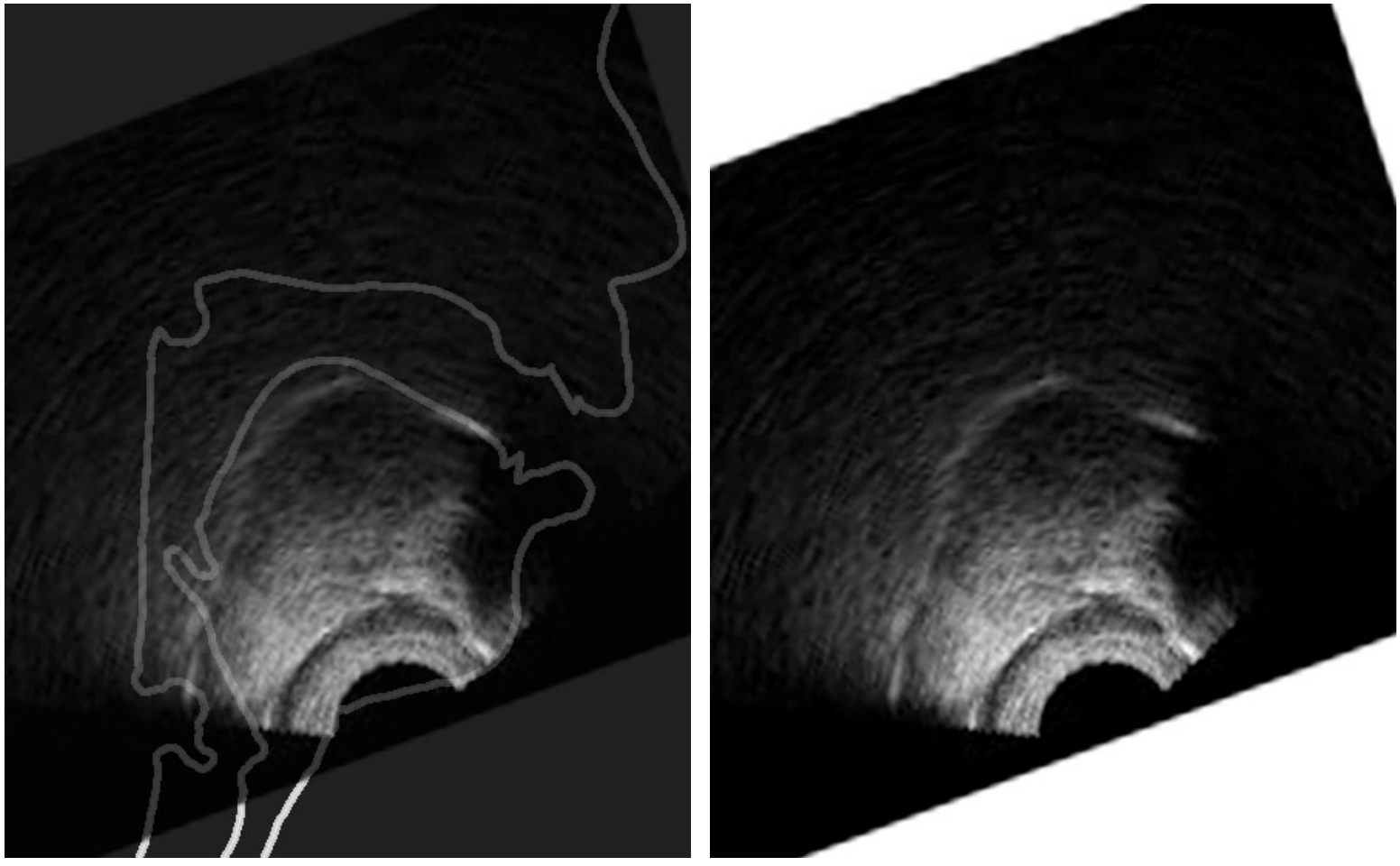
- Pre-speech articulation is interesting from several points of view, but analysing ultrasound videos manually is not great.
- In my thesis I concentrated on timing of utterance onset in both acoustics and articulation (Palo 2019).
- The data was high-speed tongue ultrasound from a delayed naming experiment – specifically one using the Rastle instructions (Rastle et al. 2005).

Classical	Stimulus (word) perception	Lexical etc processing	Movement initiation	Movement	Acoustic speech
Delayed	Lexical etc processing	Stimulus (beep) perception	Movement initiation	Movement	Acoustic speech

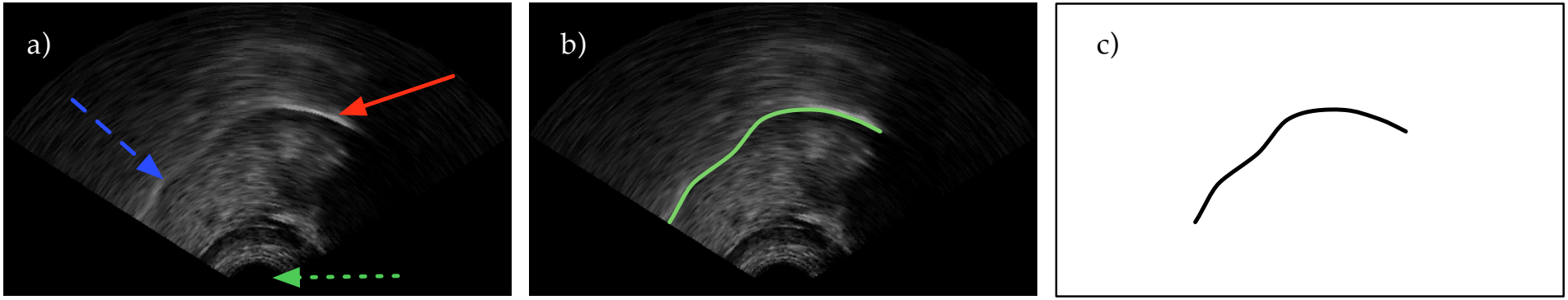
Introduction

- When trying to identify movement onset in greyscale videos with a lot of speckle 'noise', it doesn't take long to grow a desire for an easier way.
- The speckle 'noise' maybe caused by a number of factors including bubbles in the acoustic gel between the chin and the probe, and more interestingly changes in internal structures of tissues – such as muscle fibres tensing and relaxing.

Ultrasound tongue imaging: What is imaged?



Ultrasound tongue imaging: Extracting a spline



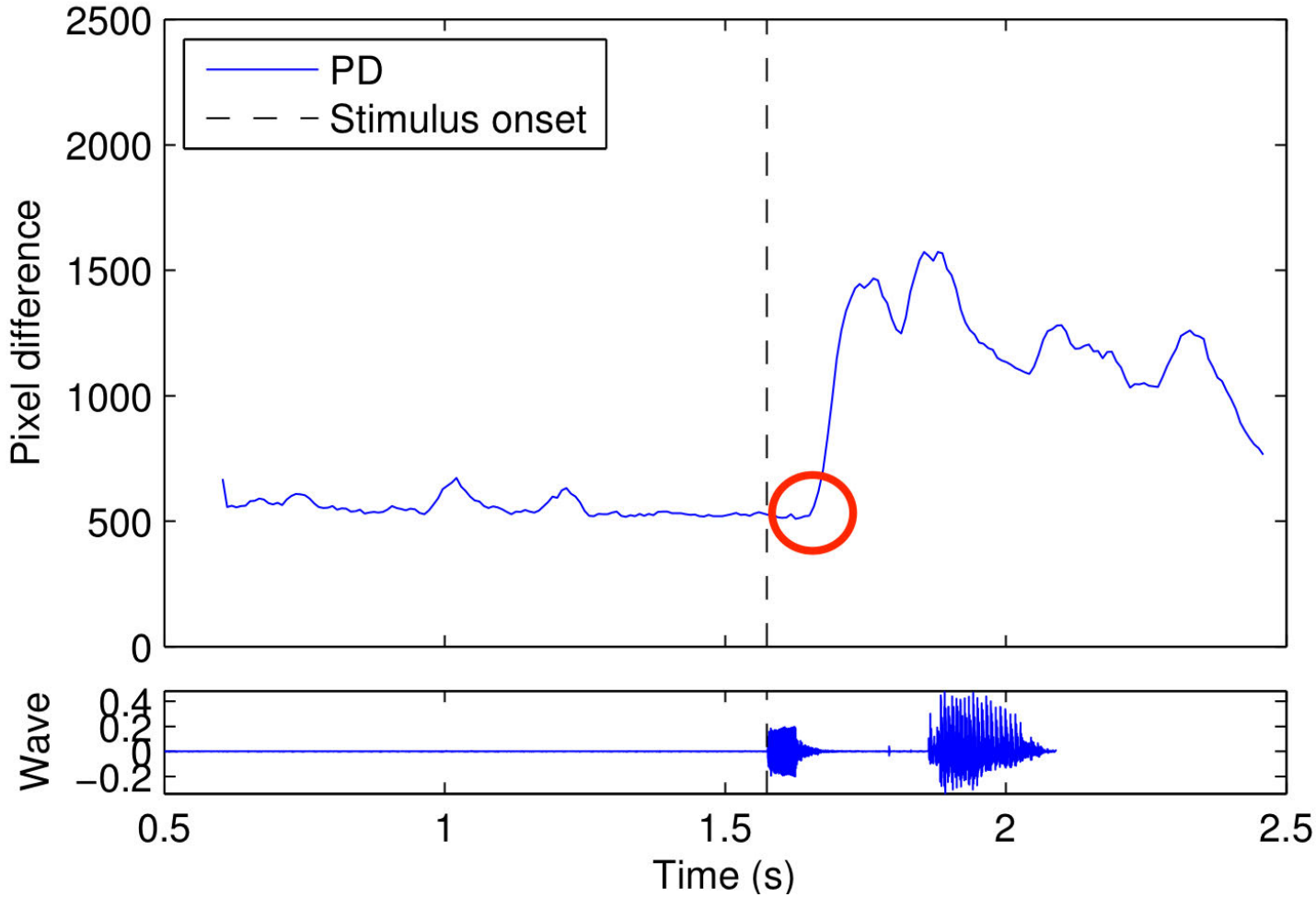
Ultrasound tongue imaging: parameters

- 2D ultrasound has good time resolution: 80-120 fps in today's examples.
- Usually, 2D ultrasound is used in the mid-sagittal plane.
- 3D/4D ultrasound can image the tongue as a volumetric object, but at the cost of lower frame rates: typically about 20 fps.
- Both have fairly good – but slightly complex – spatial resolution.

The main method

Pixel Difference (PD)

- The first tool out of the box happened to work adequately – and so for my thesis I used Euclidean distance or $l2$-norm to identify articulatory onsets.

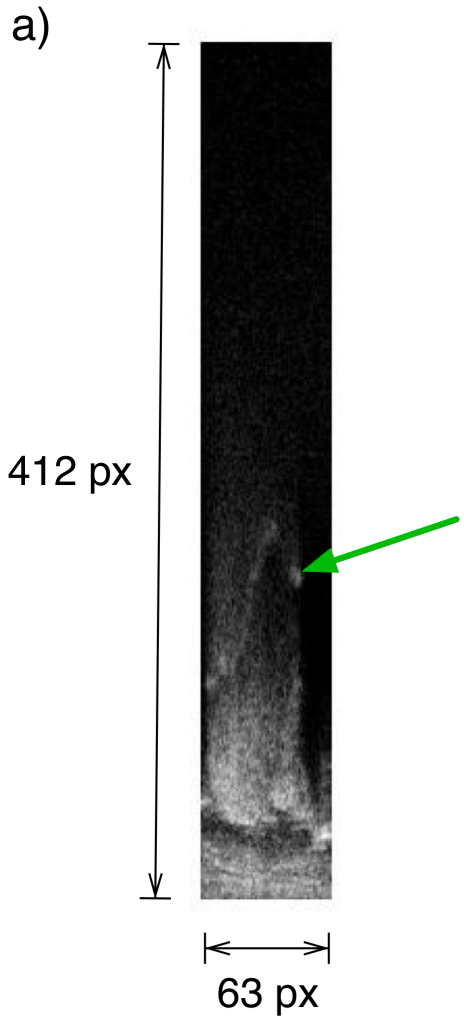


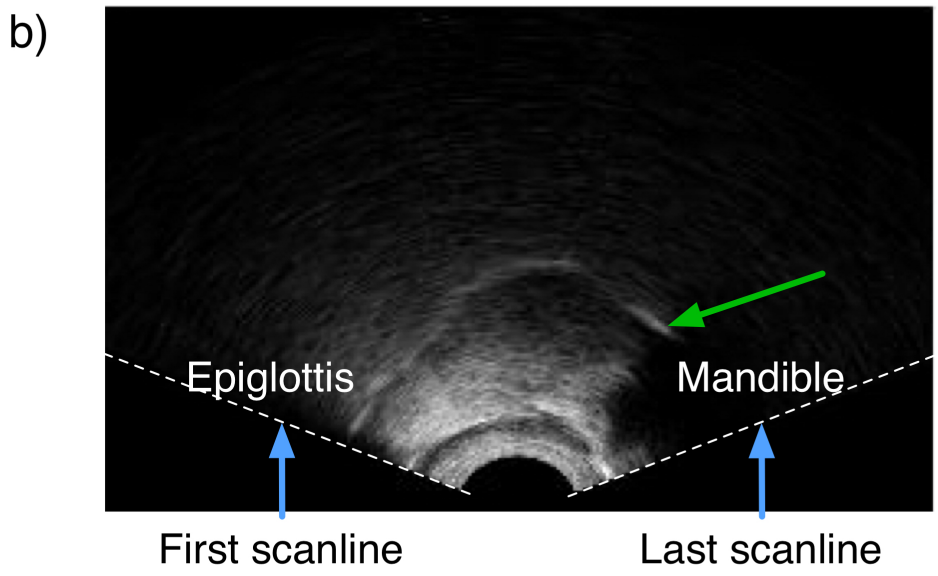
Pixel Difference (PD): Background

- The analysis methods presented here are similar to methods developed by
  - McMillan and Corley (2010) and Drake et al. (2013) who used Euclidean distance on ultrasound frames and
  - Raeesy et al. (2011) who used a similar method on MRI data.
- The way I have used it, it is actually just the Pythagorean theorem applied in a space with a lot more dimensions than 2.

Pixel Difference (PD): Raw vs. Interpolated

- PD is usually calculated on
  - (a) uninterpolated (probe-return) ultrasound data instead of
  - (b) interpolated (human-readable) data.



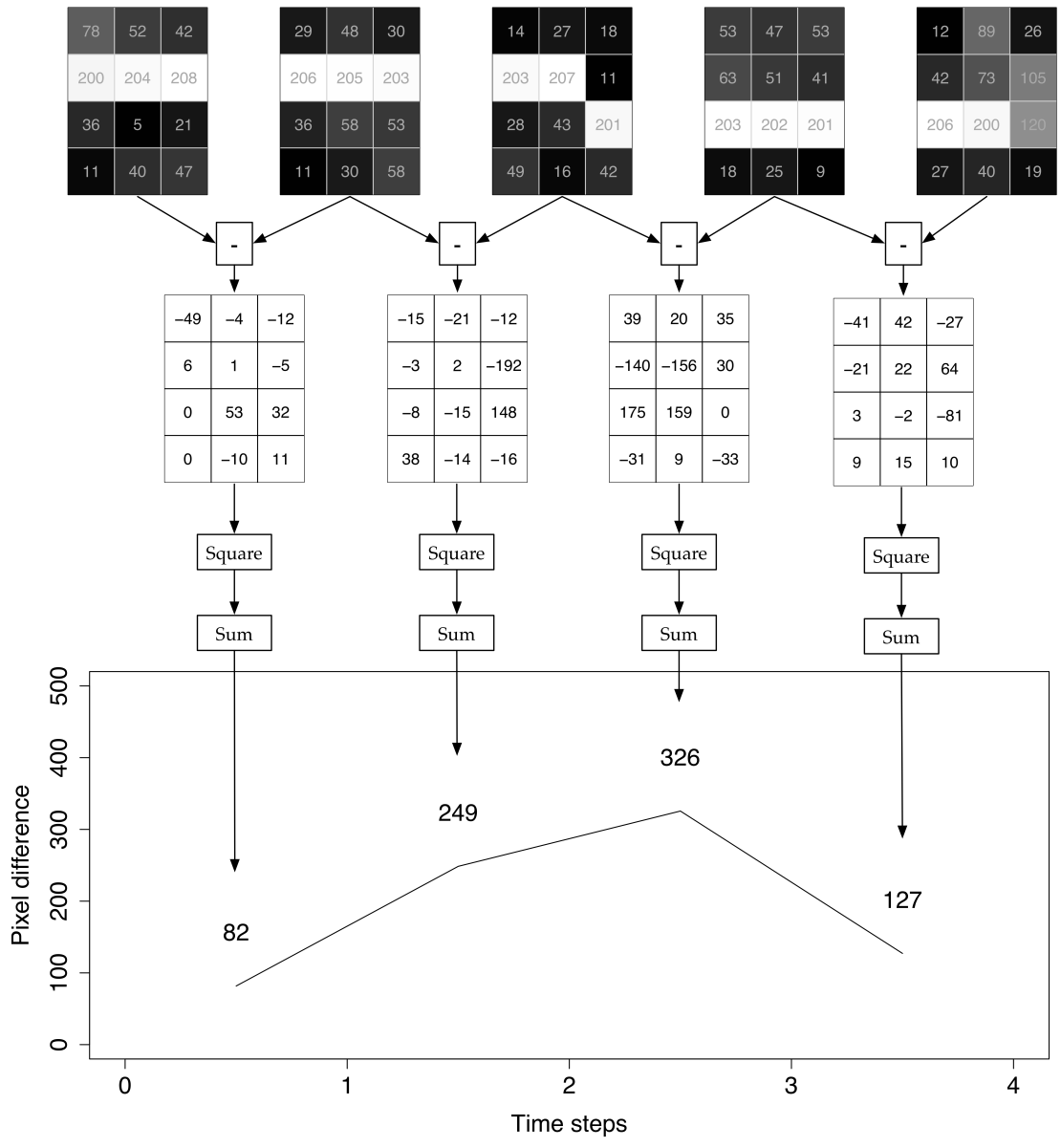


Pixel Difference (PD): The maths

$$l2(t+0.5) = \sqrt{\sum_{i,j}(x(i,j,t+1) - x(i,j,t))^2}$$

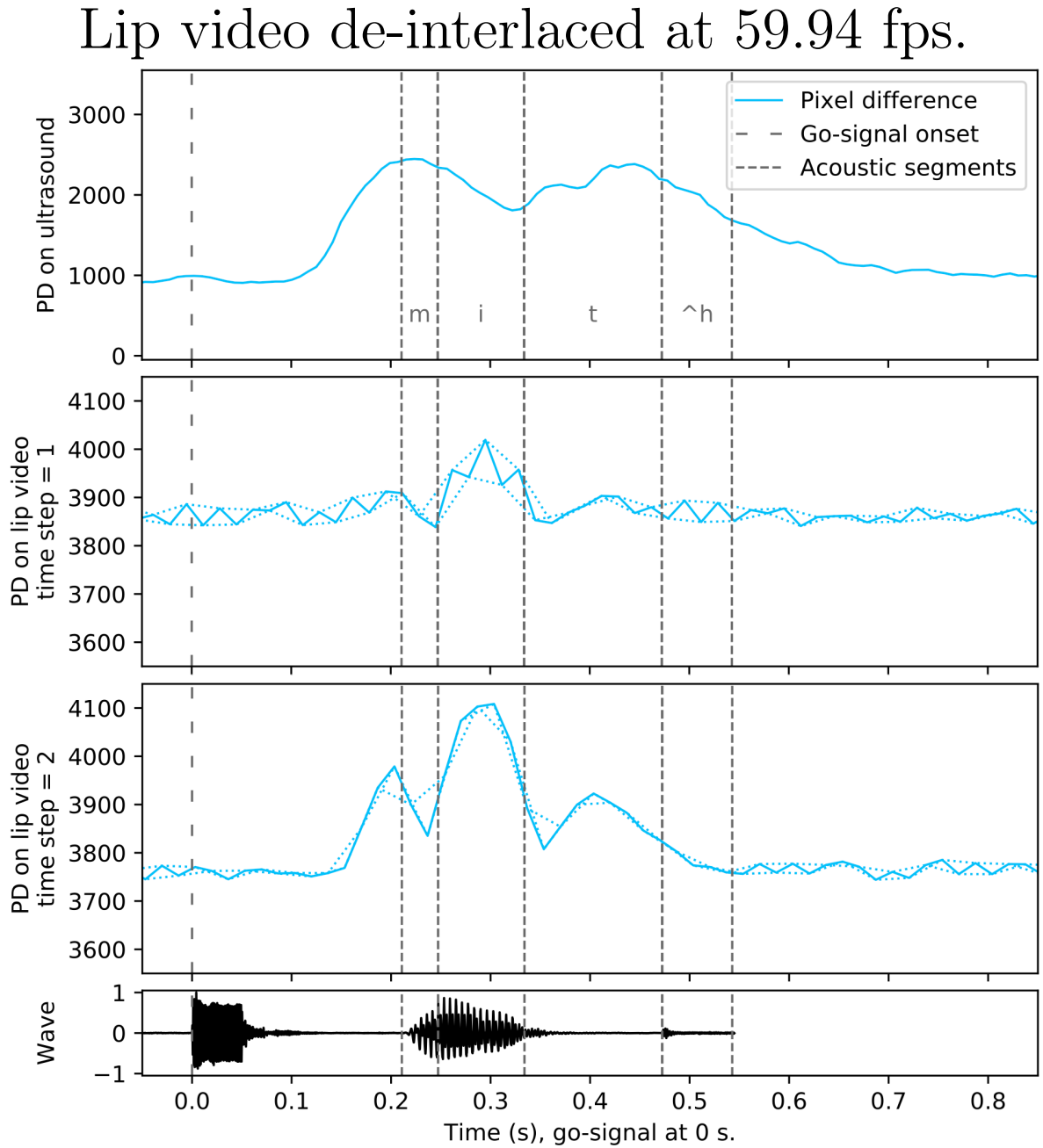
- i and j are indices that span the width and height of the image, t is the time index.
- Like said, this is actually just the Pythagorean theorem applied in a space with a lot more dimensions than 2.

Pixel Difference (PD): The maths visually



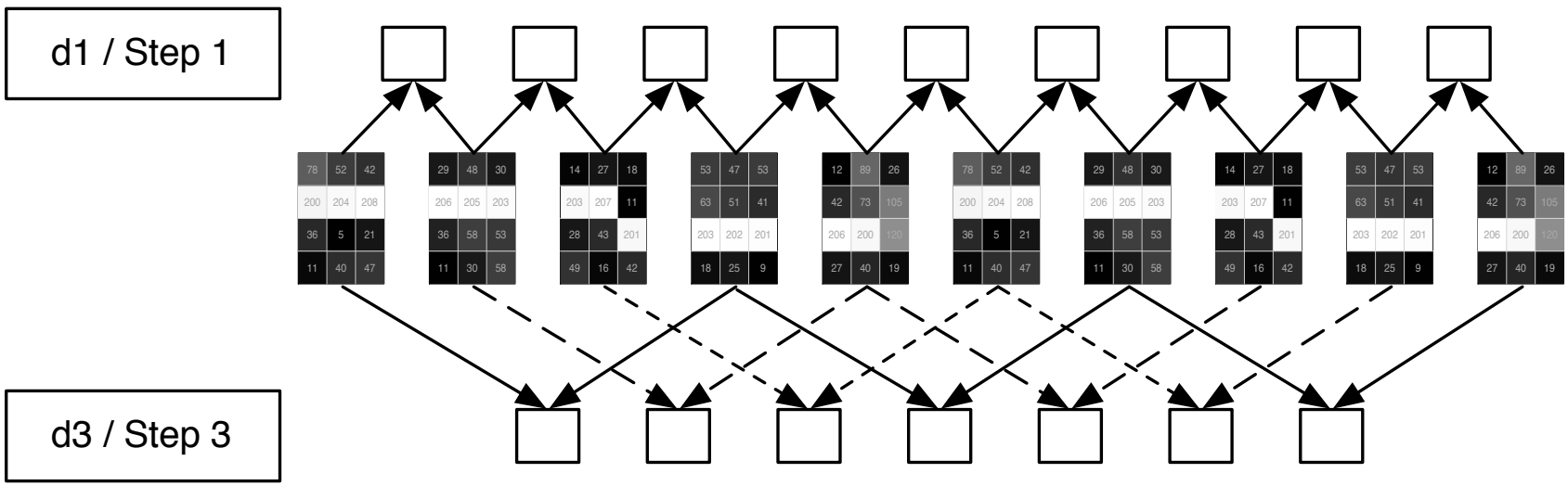
PD and other metrics applied to articulatory data

PD on de-interlaced videos



PD on de-interlaced videos

- The graphic below demonstrates taking a time step of 1 vs 3 on ultrasound.
- I tried that for my PhD thesis (Palo 2019), but found that for ultrasound a time step of 1 is preferable.
- For de-interlaced videos the best time step is 2 (Palo 2021).



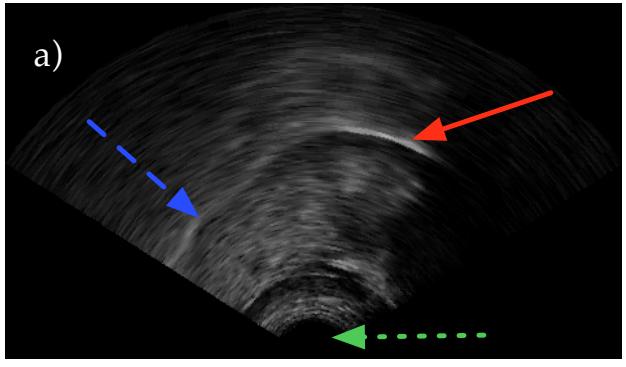
Tongue splines: Problems from spatial sparseness

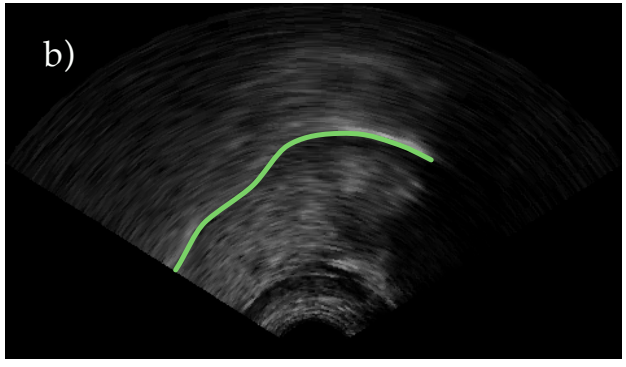
Raw ultrasound:

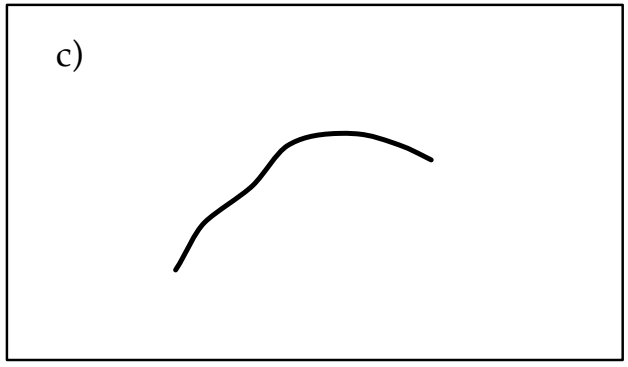
- Typically on the order of 10k pixels per frame, today 63x412 pixels per frame.
- Individual pixel's fluctuations get averaged out.

Tongue splines:

- Typically on the order of 30-50 control points per frame, today 42 control points per frame.
- Individual point's fluctuations may end up driving the data.

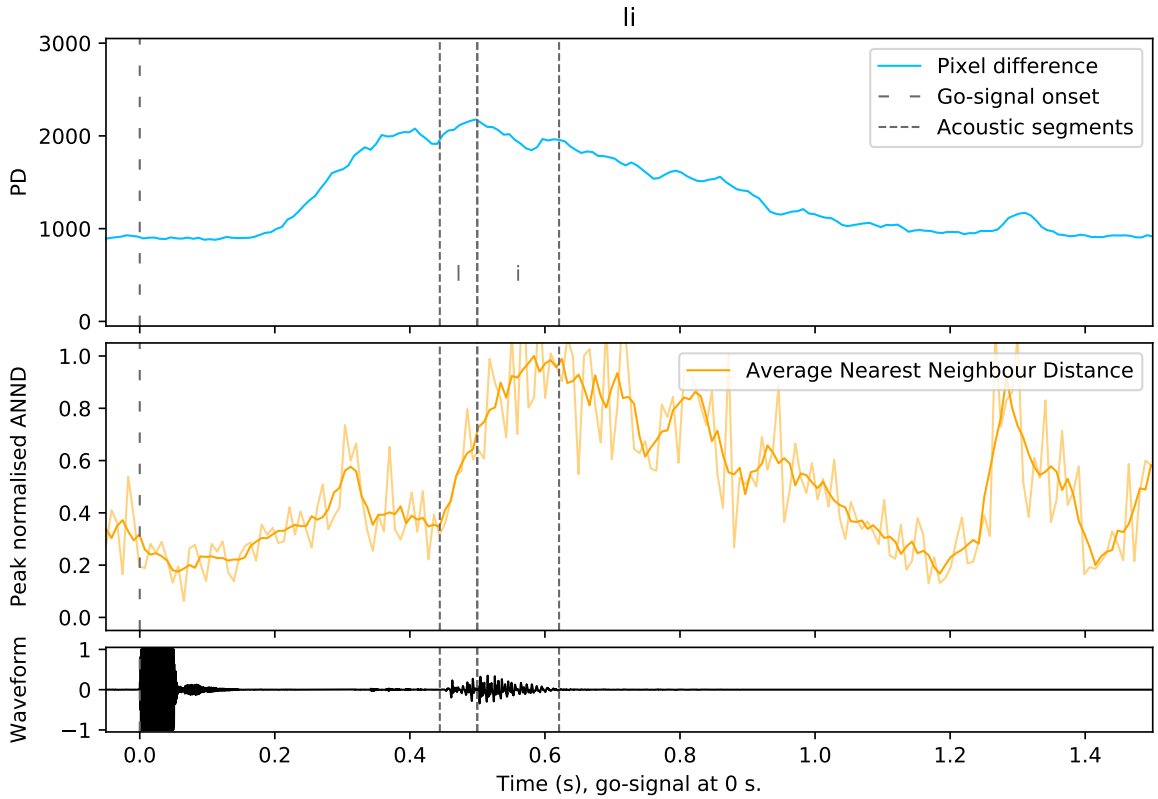






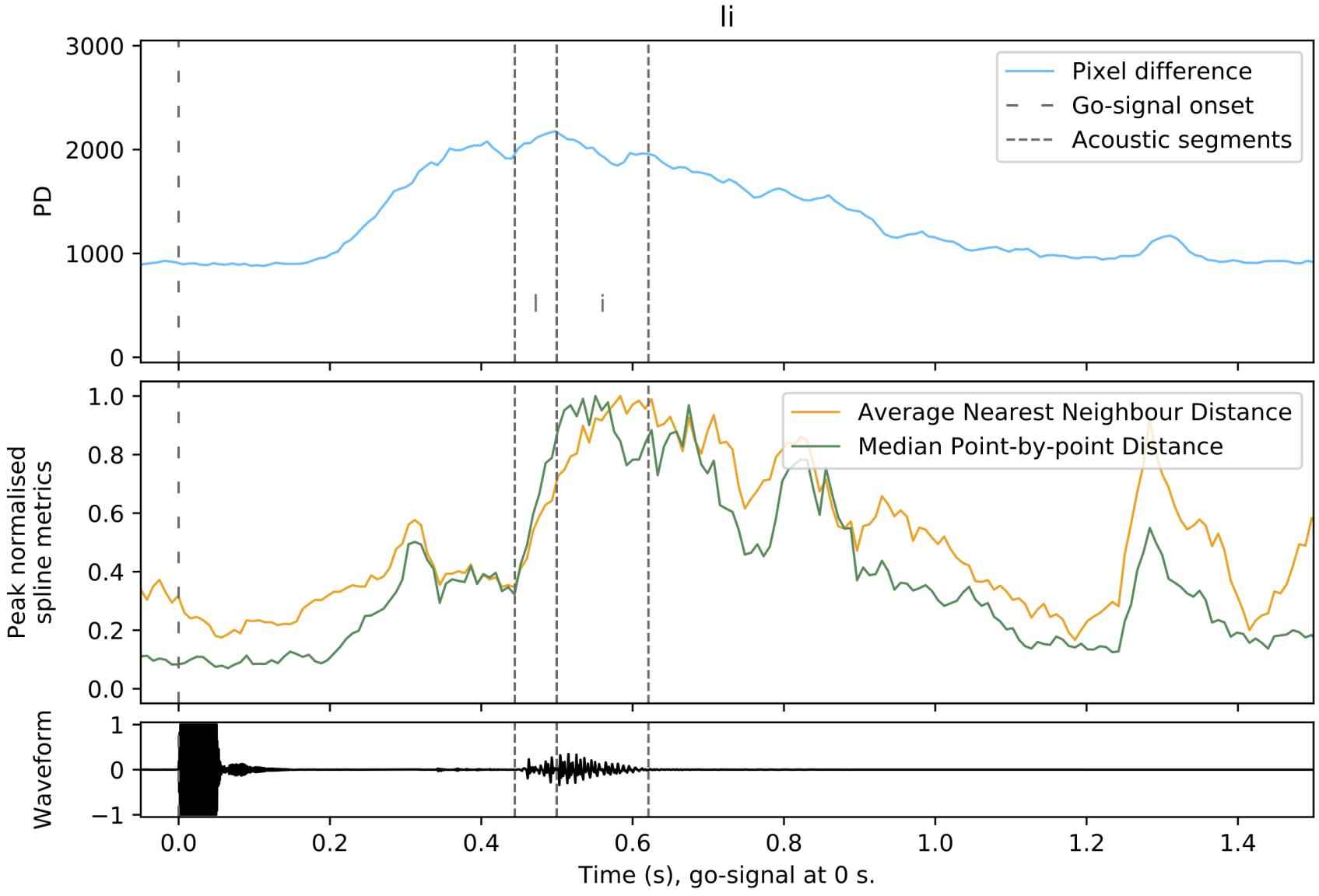
Tongue splines: Problems from spatial sparseness

- Longer time step and averaging improve the results.
- Here and in the next slide ANND (Zharkova and Hewlett 2009) and MPBPD (Palo 2020) have been calculated with time step 3 and smoothed with a moving average filter with a 5 frame window.



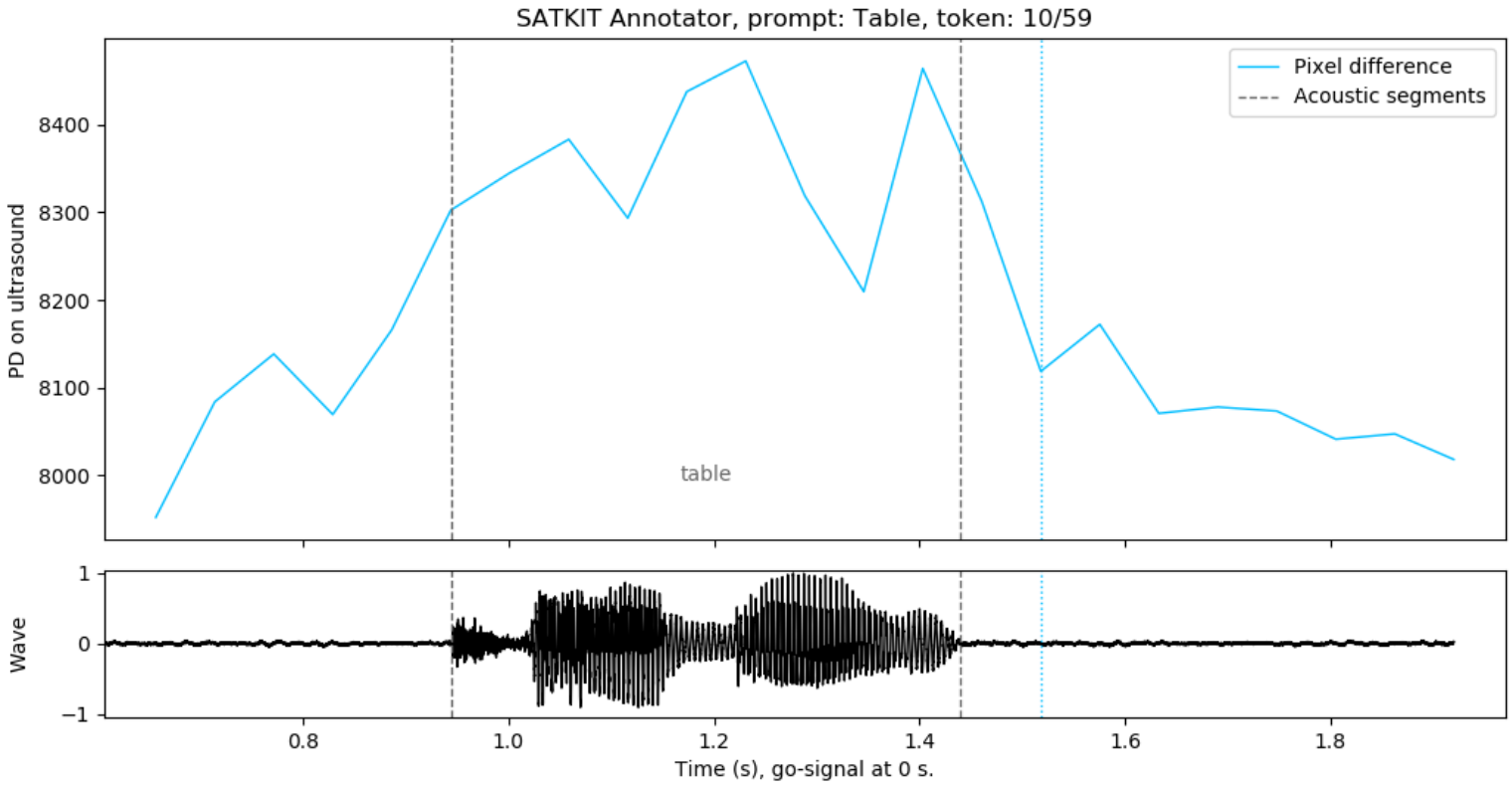
Tongue splines: Problems from spatial sparseness

- Choice of metric can help, but not with everything.

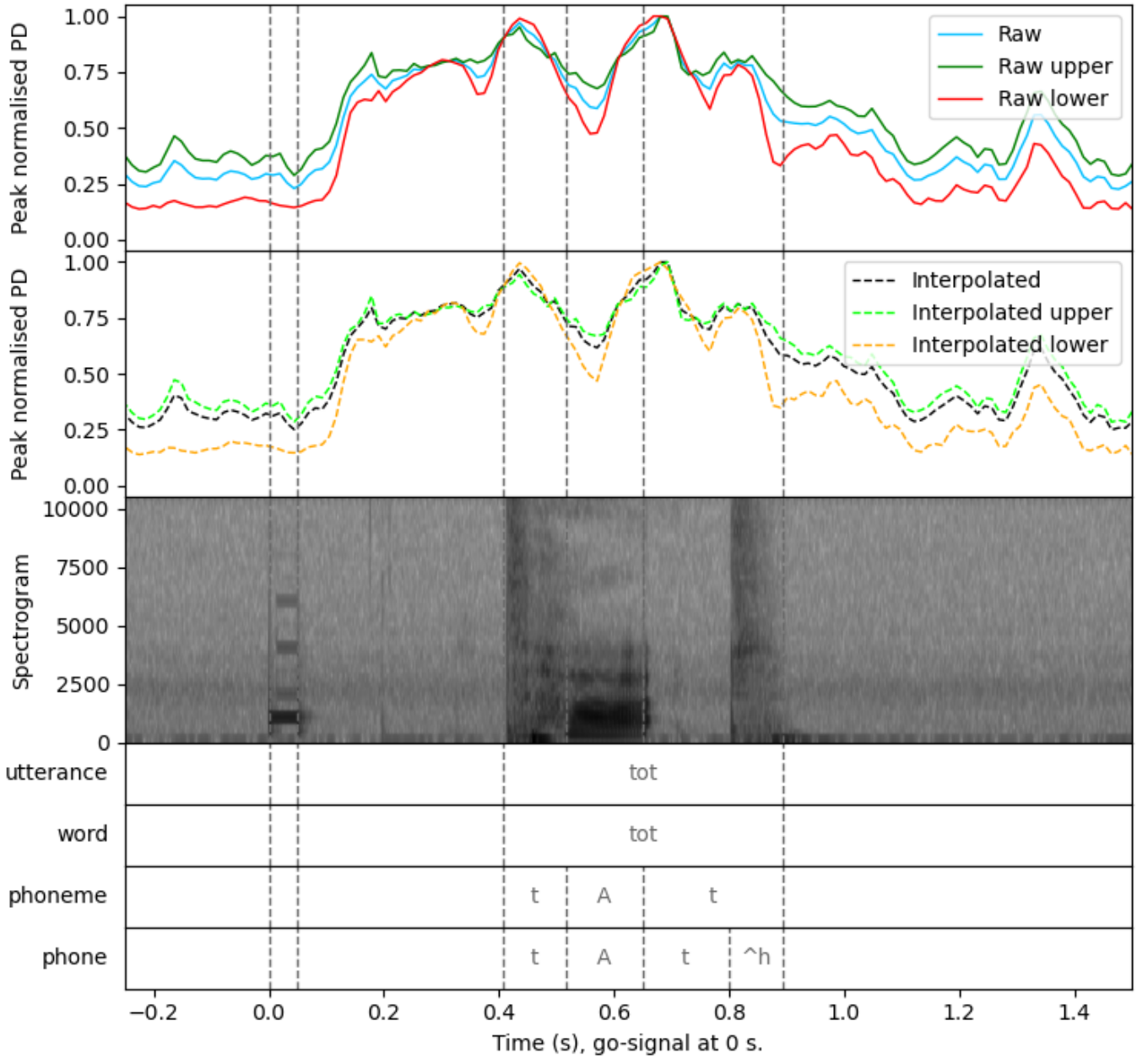


3D/4D ultrasound

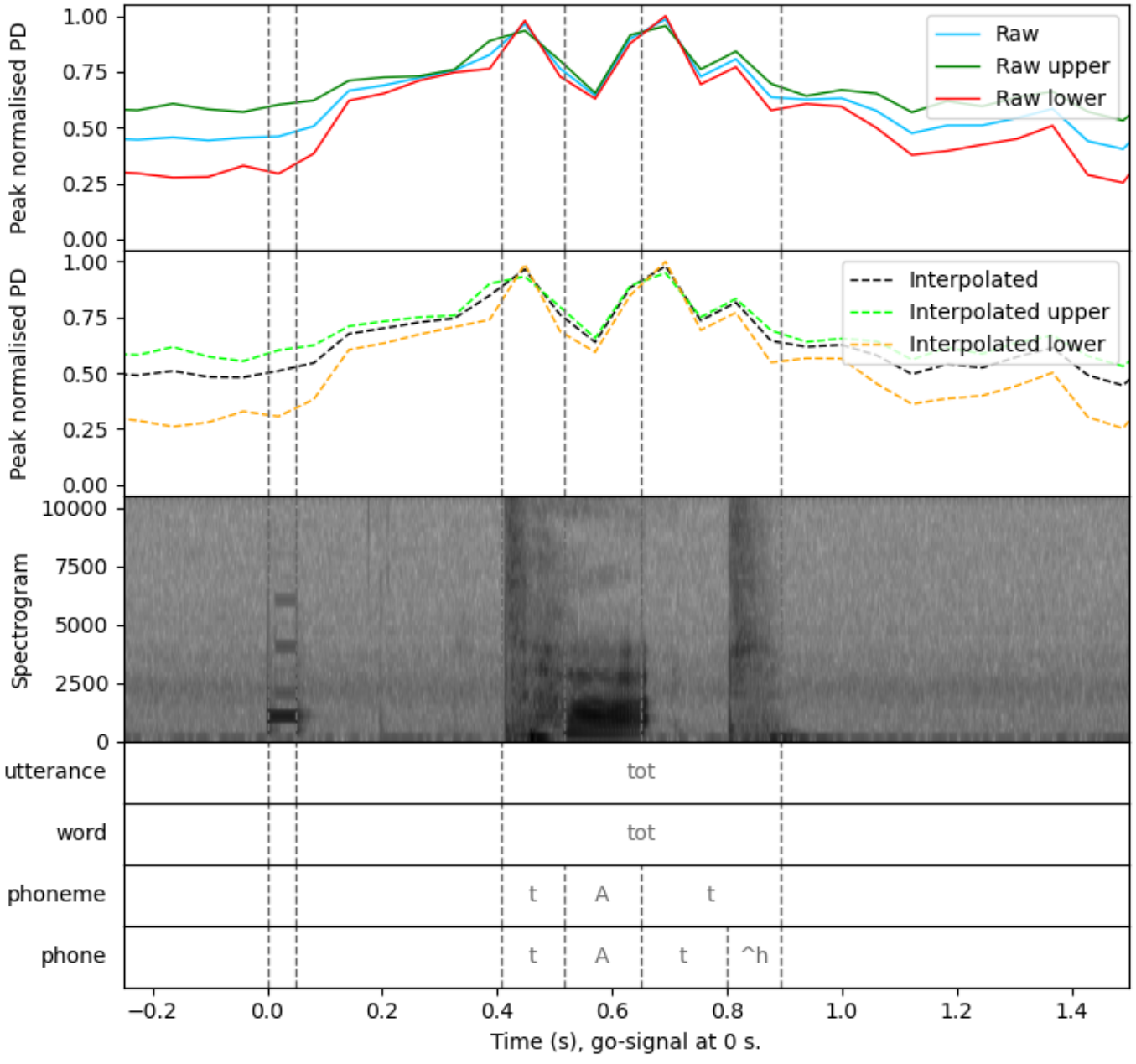
- Capturing a 3D frame takes a lot longer.
- The images are always interpolated.
- In analysis even on good (lucky) samples onset and gesture recognition becomes difficult.



PD on Raw vs Interpolated 2D data



PD on data with artificially lowered frame rate



In the works: Choosing the metric for PD

- PD has so far usually been calculated as the Euclidean distance or $l2$-norm.
- We've recently been looking at principled ways of selecting the norm for a given data source – such as 2D ultrasound – from the different lp-norms where $p \in]0, \inf[$.
- It looks like the optimal norm for 2D ultrasound is $l1$ (or close to it):

$$l1(t + 0.5) = \sum_{i,j} |x(i, j, t + 1) - x(i, j, t)|$$

MRI: Challenges and how to deal with them

- Frame rate can be a problem.
- If there are systematic changes frame-to-frame caused by the imaging and reconstruction these may show up in PD analysis.
- Best way to get ahead with using PD for analysis would be to get a small pilot sample and run the basic version on it: $l2$ or $l1$-norm, time step $= 1$, no smoothing.
- Apply larger time steps and smoothing if needed.
- Test different norms and/or look to different metrics all together.

MRI: What makes it exciting?

- Triangulation is always good: One view is no view.
- MRI can observe structures that are otherwise really difficult to image.
- Fast imaging sequences should be able to deal with the frame rate issues.
- Metrics like PD work well on data with fairly dense spatial resolution.

Thank you! Do you have questions?

References

Drake, E., Schaeffler, S., and Corley, M. (2013). ARTICULATORY EVIDENCE FOR THE INVOLVEMENT OF THE SPEECH PRODUCTION SYSTEM IN THE GENERATION OF PREDICTIONS DURING COMPREHENSION. In *Architectures and Mechanisms for Language Processing (AMLaP)*, Marseille.

McMillan, C. T. and Corley, M. (2010). Cascading influences on the production of speech: Evidence from articulation. *Cognition*, 117(3):243–260.

Palo, P. (2019). *Measuring Pre-Speech Articulation*. PhD thesis, Queen Margaret University, Edinburgh.

Palo, P. (2020). Can we detect initiation of tongue internal changes before overt movement onset in ultrasound? In *Proceedings of the 12th International Seminar on Speech Production (ISSP 2020)*, pages 242–245, Online / New Haven, CT.

Palo, P. (2021). Computer assisted segmentation of tongue ultrasound and lip videos. *Journal of the Canadian Acoustical Association*, 49(3):44–45.

Raeesy, Z., Baghai-Ravary, L., and Coleman, J. (2011). Parametrising Degree of Articulator Movement from Dynamic MRI Data. In *12th Interspeech*, pages 2853–2856.

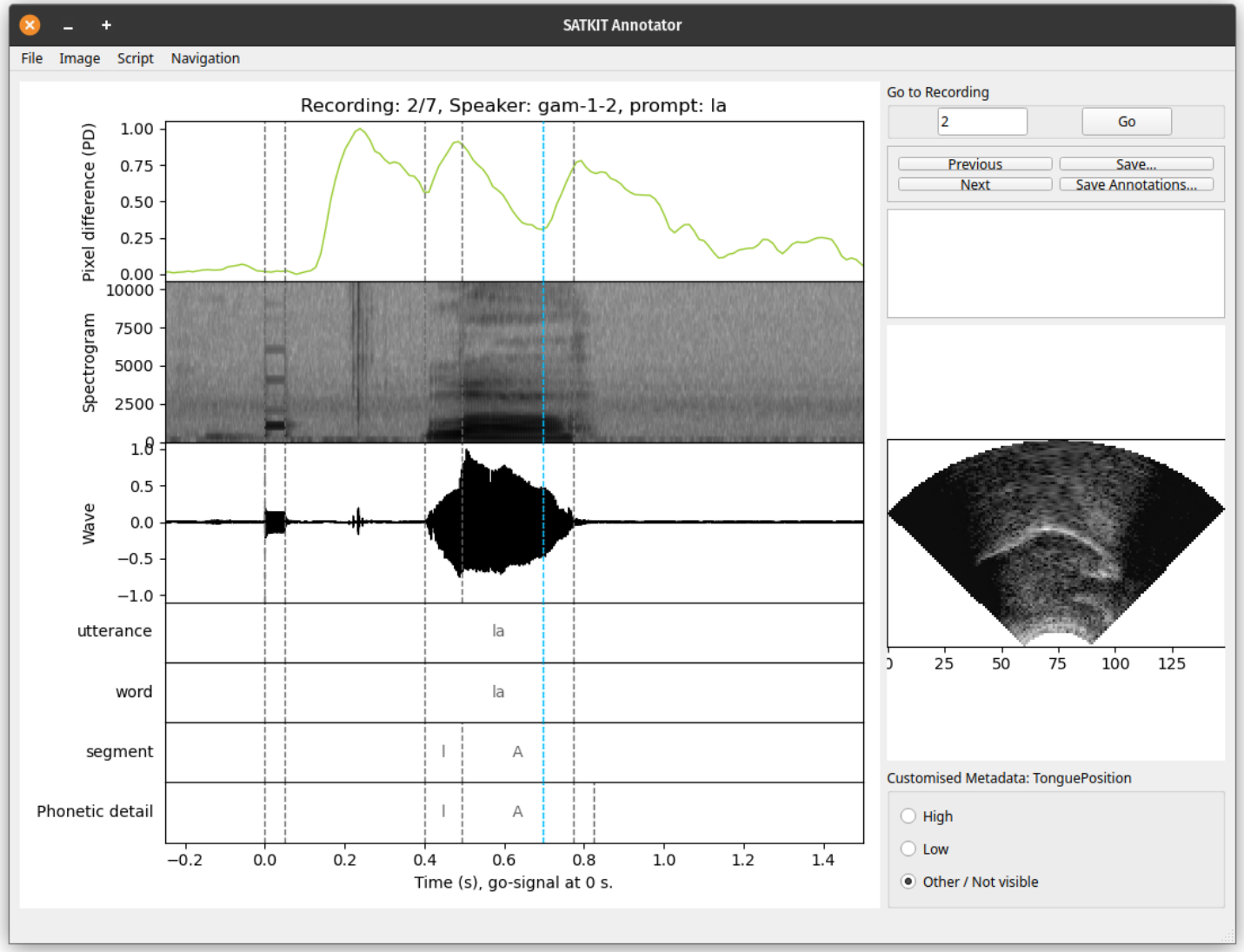
Rastle, K., Harrington, J. M., Croot, K. P., and Coltheart, M. (2005). Characterizing the Motor Execution Stage of Speech Production: Consonantal Effects on Delayed Naming Latency and Onset Duration. *Journal of Experimental Psychology: Human Perception and Performance*, 31(5):1083–1095.

Zharkova, N. and Hewlett, N. (2009). Measuring lingual coarticulation from midsagittal tongue contours: Description and example calculations using English /t/ and /ɑ/. *Journal of Phonetics*, 37:248–256.

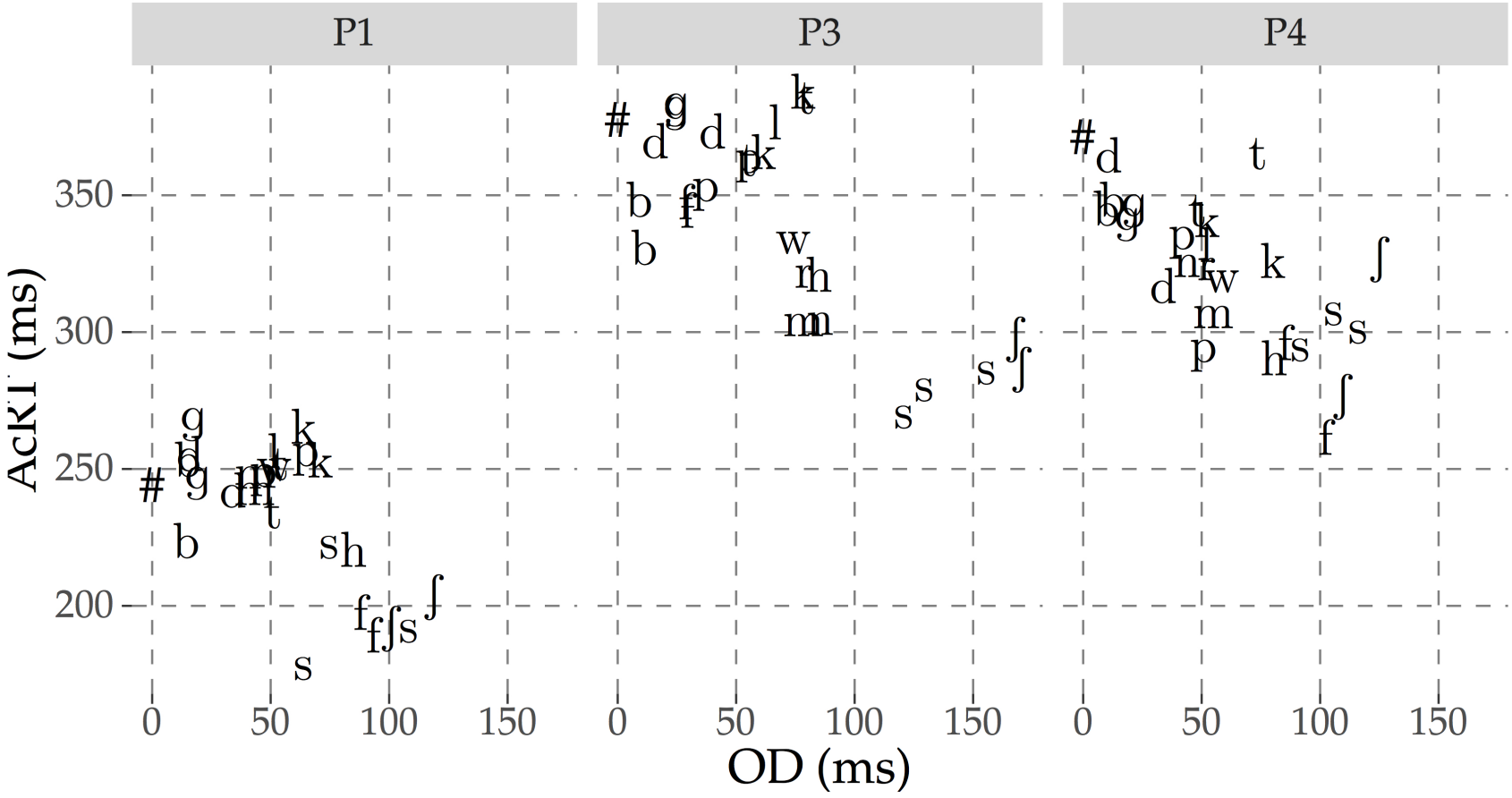
Extra material

Speech Articulation ToolKIT

- `https://github.com/giuthas/satkit`.
- Written in Python and in active development.

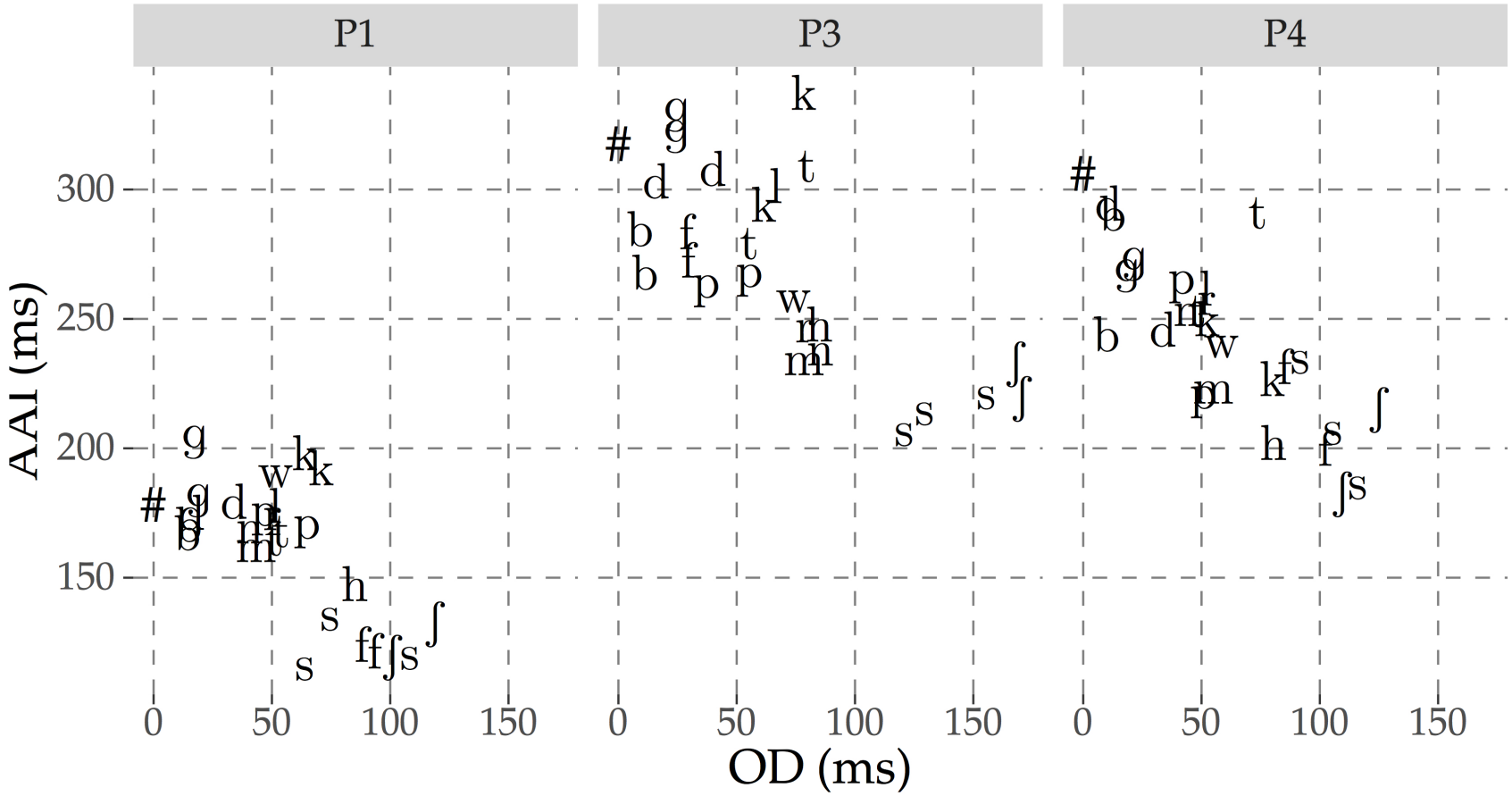


Delayed naming results: Acoustics



Medianised within participant, over several repetitions and over the vowels /a,i,ɔ/. Over all analysable n = 1386: 439 from P1, 672 from P3, and 275 from P4.

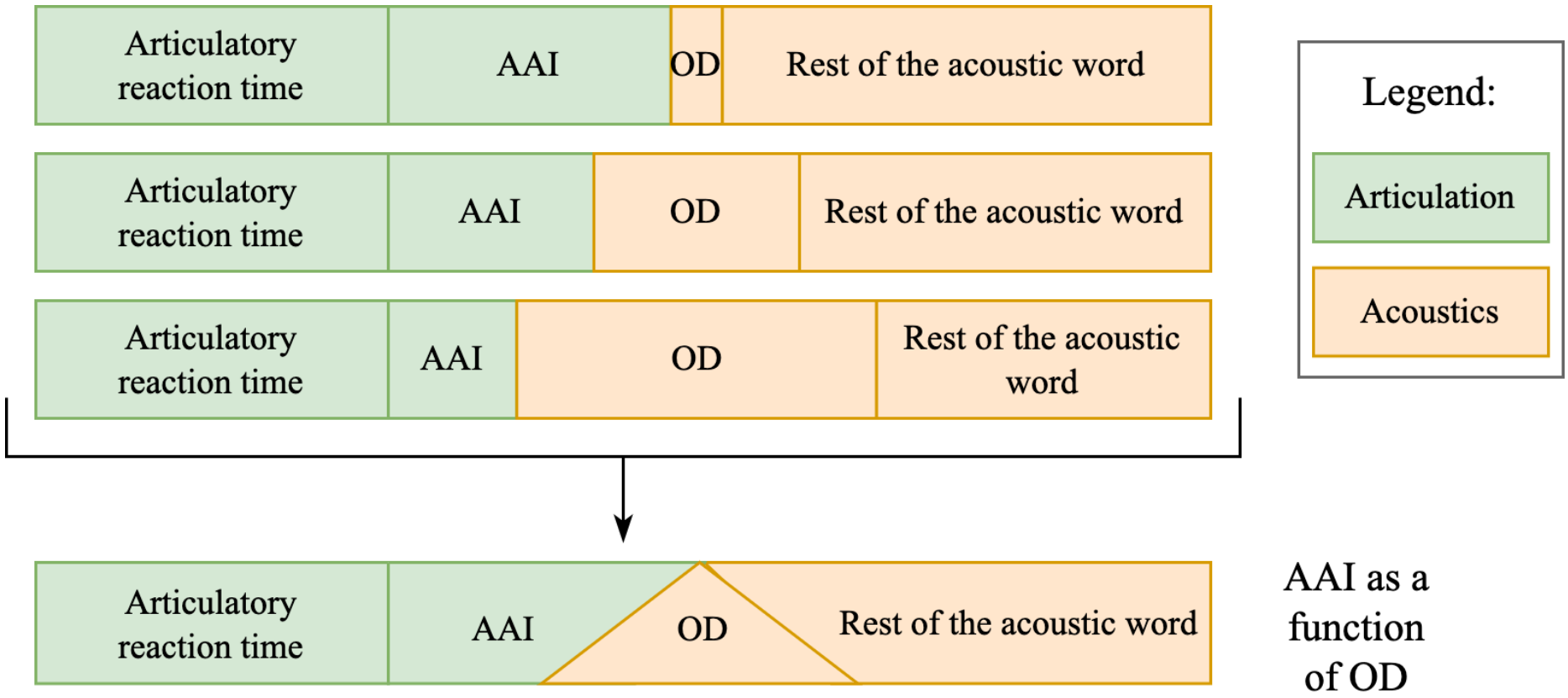
Delayed naming results: Articulatory to Acoustic Interval



Medianised within participant, over several repetitions and over the vowels /a,i,ɔ/. Over all analysable n = 1386: 439 from P1, 672 from P3, and 275 from P4.

Theory: Effect of OD on AAI

- As the Onset Duration (OD) gets longer, Articulatory to Acoustic Interval (AAI) shortens.
- First three lines represent individual utterances, final line is a conceptual model of the effect of continuously lengthening OD.6



Theory: Effect of articulatory rate on AAI

- If we keep the utterance content constant but vary articulation rate, all parts (AAI, OD, and acoustic word) get longer as articulation rate goes down.