Analysing Articulatory Data with Vector Norms and Related Methods

Pertti Palo

9 Oct 2023

Outline

- I will then talk about what we can and cannot do with these methods in time domain analysis of articulatory data these days. I will take short side trips to look at similar methods applied to other articulatory data. We have looked at tongue splines and lip videos and I will discuss what kind of challenges and understanding has resulted from those attempts. Finally, analysing 3D/4D ultrasound has been a recent major focus, but unfortunately frame rate issues are not so easy to solve.
- I will finish the talk by discussing why MRI would be very interesting to analyse with these methods – or adapted versions of them – and which definite and possible challenges one might come across.

Introduction: The why

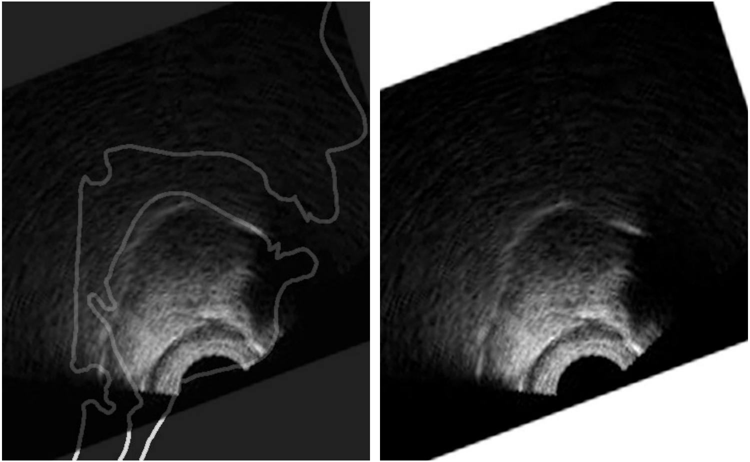
- Pre-speech articulation is interesting from several points of view, but analysing ultrasound videos manually is not great.
- In my thesis I concentrated on timing of utterance onset in both acoustics and articulation (Palo 2019).
- The data was high-speed tongue ultrasound from a delayed naming experiment – specifically one using the Rastle instructions (Rastle et al. 2005).
- 2D ultrasound has good time resolution: 80-120 fps in today's examples.

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

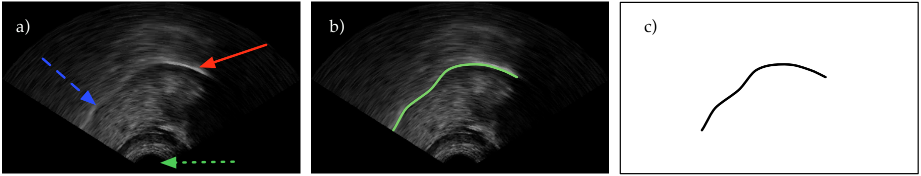
Introduction: The why

- 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.

What is being imaged by tongue ultrasound?



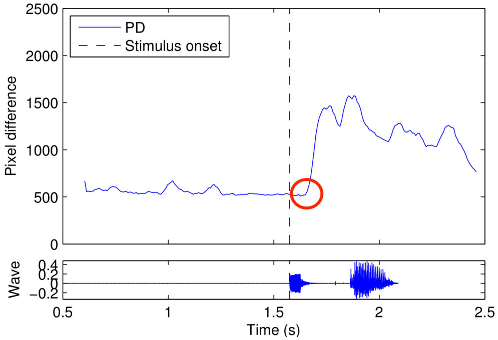
Where is the tongue?



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:

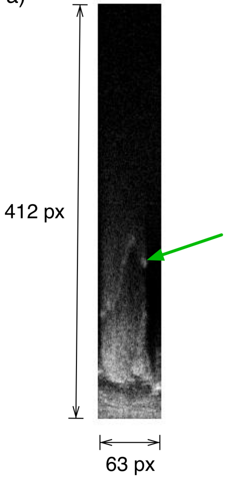
$$l2(t+0.5) = \sqrt{\sum_{i,j} (x(i,j,t+1) - x(i,j,t))}$$



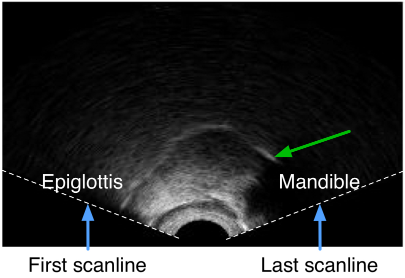
Pixel Difference (PD)

- PD is usually calculated on uninterpolated (probe-return) ultrasound data (a) as opposed to interpolated (human-readable) data (b).

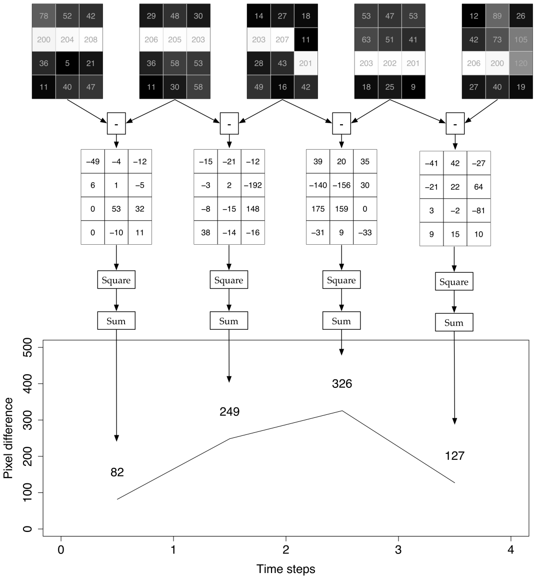
a)



b)



Pixel Difference (PD)

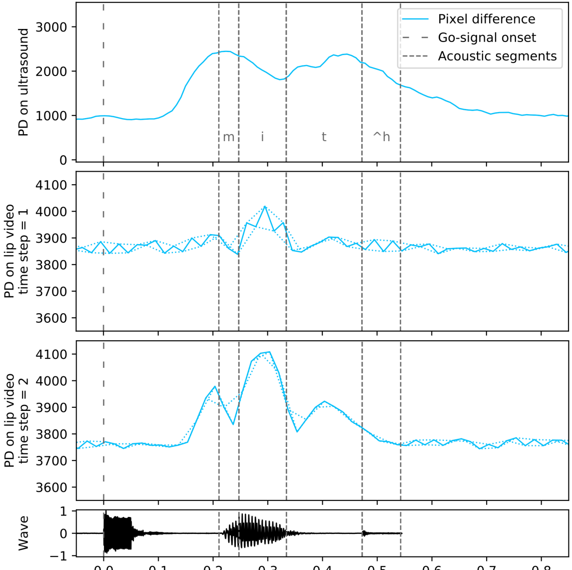


Introduction: The how

- de-interlaced video and the zigzag: time steps
- splines and the sparseness of the data
- PD on ultrasound and time resolution
- Choosing a metric

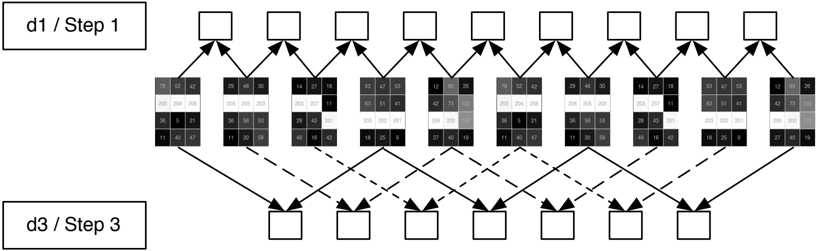
PD on de-interlaced videos

Lip video with camera attached to the ultrasound helmet.
De-interlaced at 59.94 fps.



PD on de-interlaced videos

Taking a different time step:



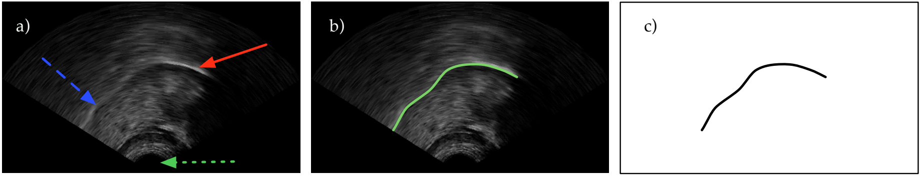
Tongue splines and problems from sparseness

Raw ultrasound:

- Typically 63x406 pixels per frame.
- Individual pixel's fluctuations get averaged out.

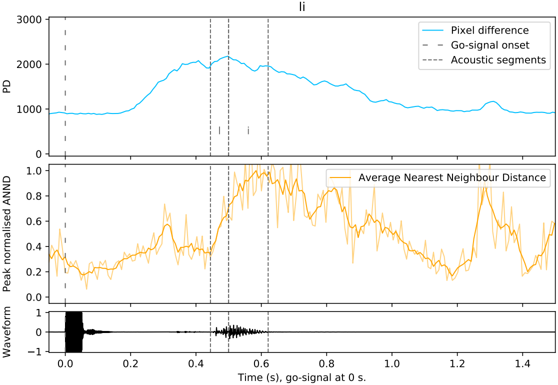
Tongue splines:

- Typically control points per frame.
- Individual point's fluctuations may end up driving the data.



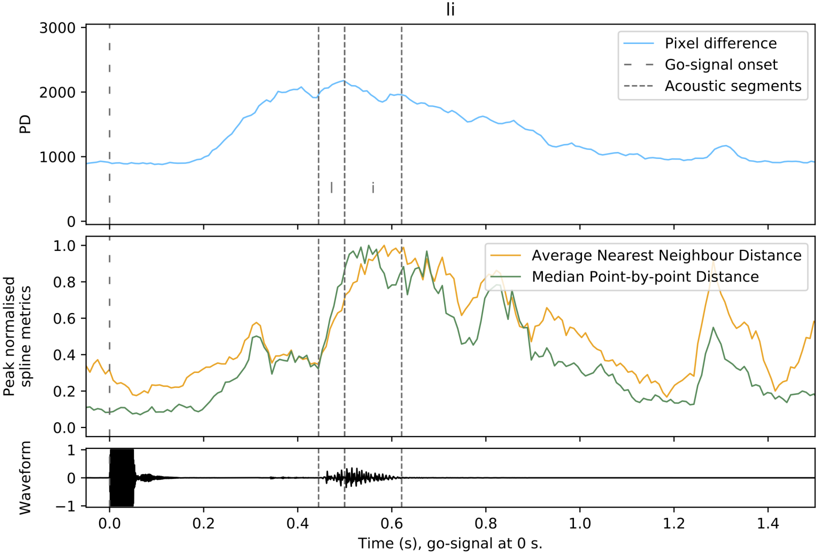
Tongue splines and problems from sparseness

- Longer time step and averaging produce better results, but still with limits.
- Here and in the next slide ANND and MPBPD have been calculated with time step 3 and smoothed with a moving average filter with a 5 frame window.



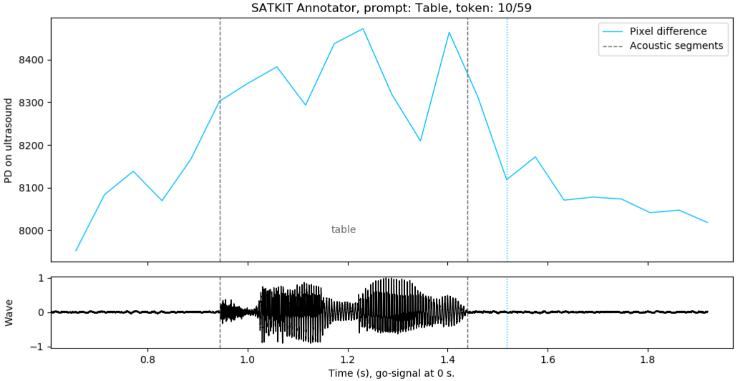
Tongue splines and problems from sparseness

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

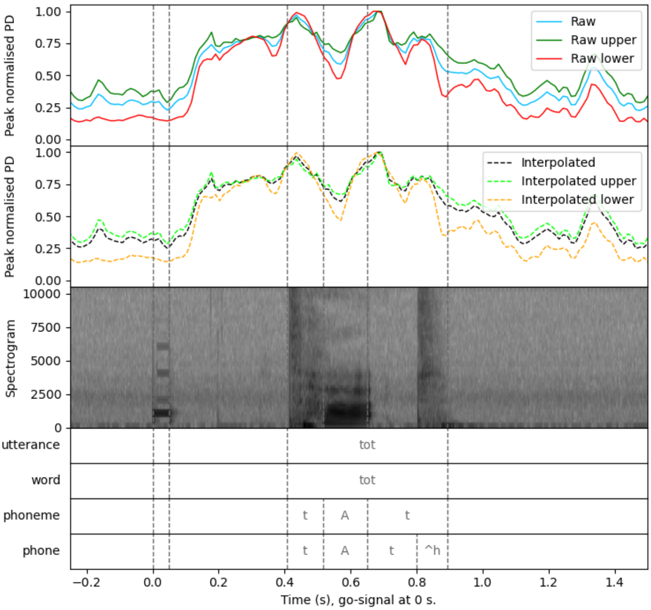


3D/4D ultrasound

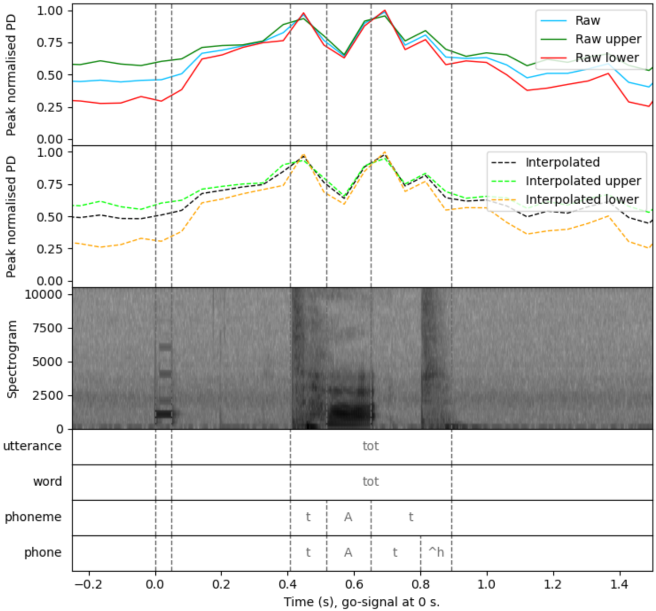
- Capturing a 3D frame takes a lot longer and we only have access to interpolated data.
- The images are always interpolated.
- In analysis even on good (lucky) samples onset and gesture recognition becomes difficult.



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):

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

So how about MRI then?

- 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.

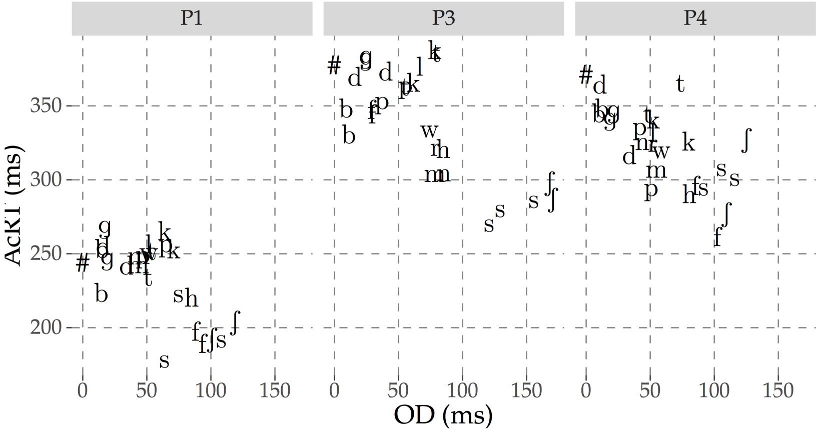
References

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

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.

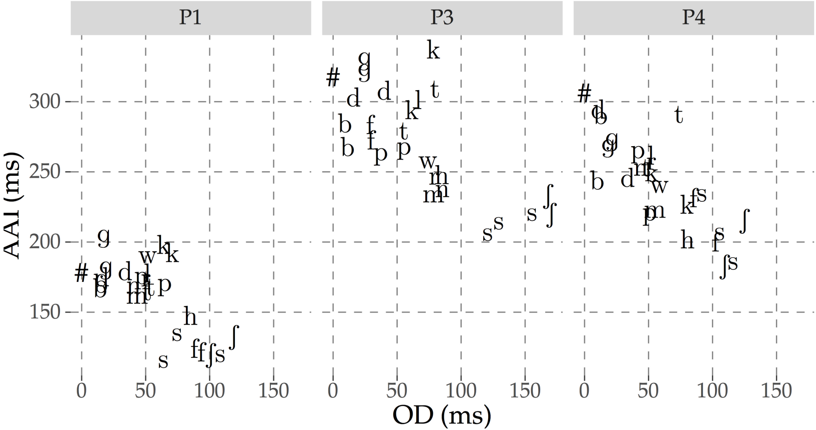
Extra material

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.

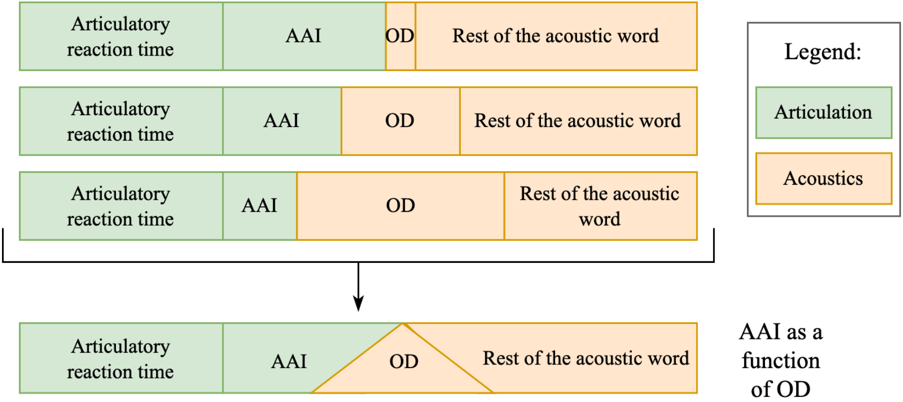
Delayed naming results: Articulatory to Acoustic Interval



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.