Articulation and Acoustics at Onset of Speech

Results from delayed naming

Pertti Palo

4 Apr 2022

Foreshadowing

- Introduction, backgroung and motivation
- Materials: Copying the previous – with a twist
- Methods: Digging deeper into ultrasound
- Results: Revisiting the beginning
- Tentative theory
- Conclusion
- References and some extra material

Introduction I

- Our story starts with speech reaction times (RTs), which are an important tool in psycholinguistics.
- Speech RTs are usually measured based on acoustics even though this is known to be problematic (Rastle and Davis 2002, Rastle et al. 2005).
- Direct measurement of movement onset gives a more detailed and a more appropriate method of evaluating speech reaction times, while at the same time making it possible to study the motor control and phonetical aspects of speech initiation (Kawamoto et al. 2008, Palo 2019).
- However, processing large amounts of articulatory data can be too time consuming to be feasible (Roon 2013).

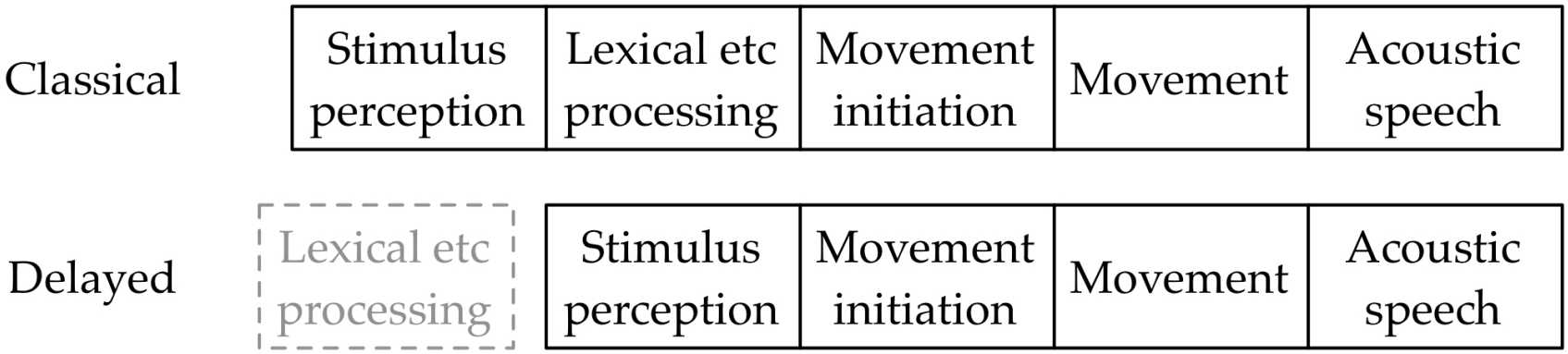
Introduction II

- In my PhD project I studied speech initiation with articulatory methods.
- I also developed methods for fast and accurate annotation of ultrasound data.
- The methods (and some additions) are now available in Python on Github (Palo, P. and Moisik, S. R. and Faytak, M. 2020, Faytak et al. 2020).

Basic reaction time tasks

Task instructions make a great difference in what happens when speech is initiated. For example:

- Classical naming: As a word appears on the screen read it out loud as fast as possible.
- Delayed naming: Read the word out loud as fast as possible after the beep.



Delayed naming, Rastle version

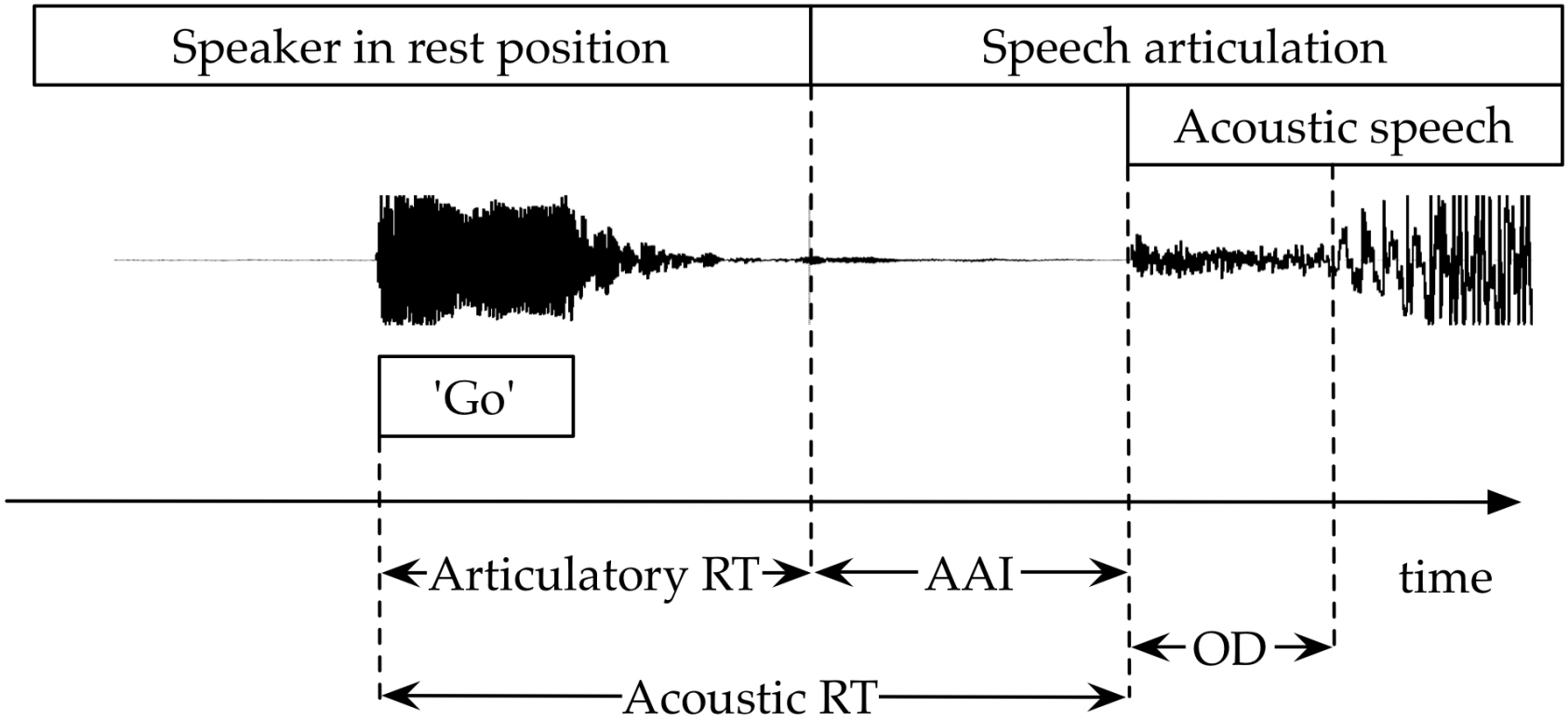
Instructions:

- First rehearse reading the word out loud (or internally).
- Remain at rest until you hear the go signal – finger click – and then produce the target word as soon and as accurately as possible.

Let's try it:

- Practice run: say 'caught'.
- Speeded trial: say 'caught' as soon as possible after you here the click.

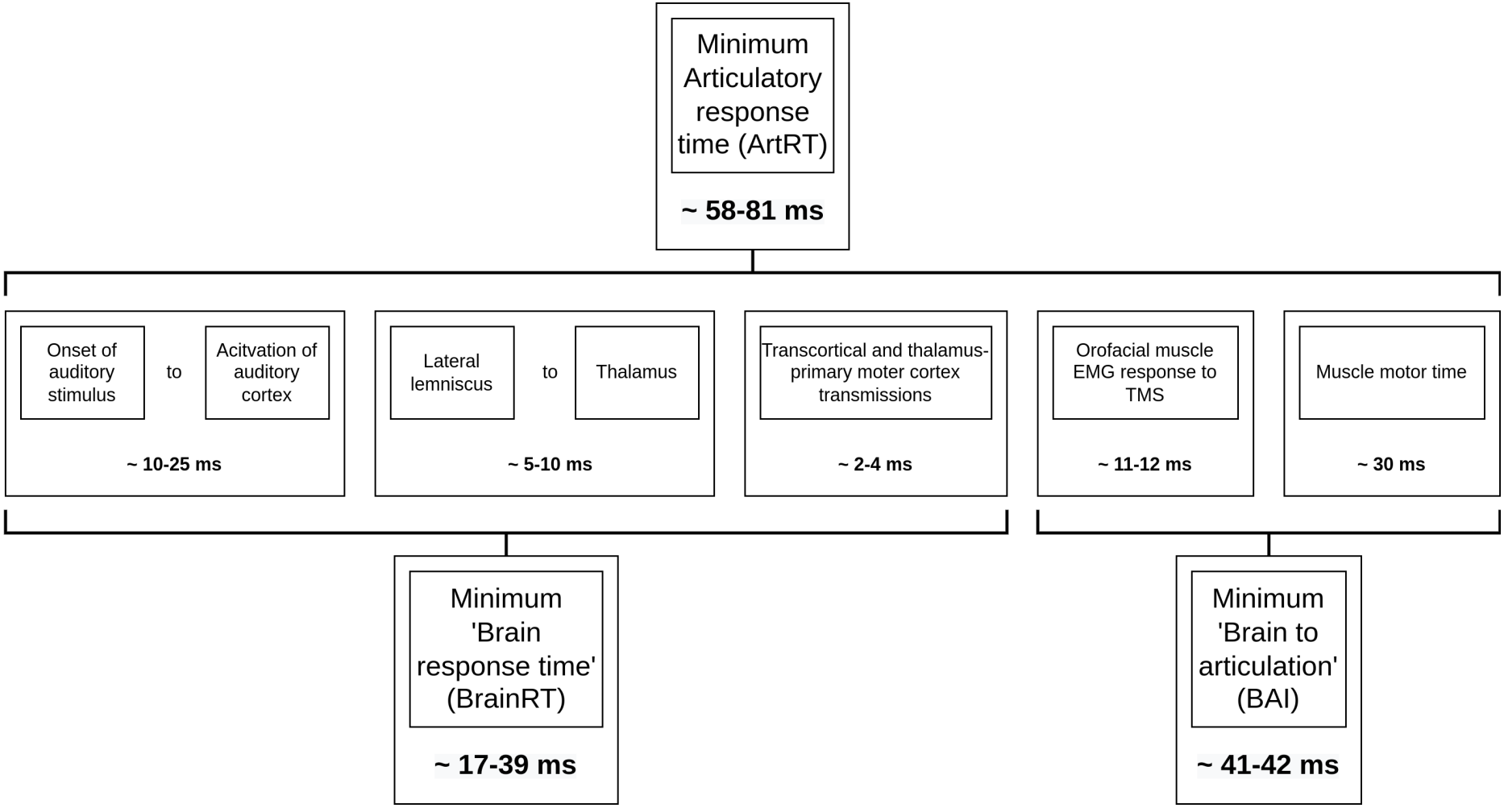
What happened?



- AAI = Articulatory onset to Acoustic onset Interval
- OD = Onset (or obstruent) Duration

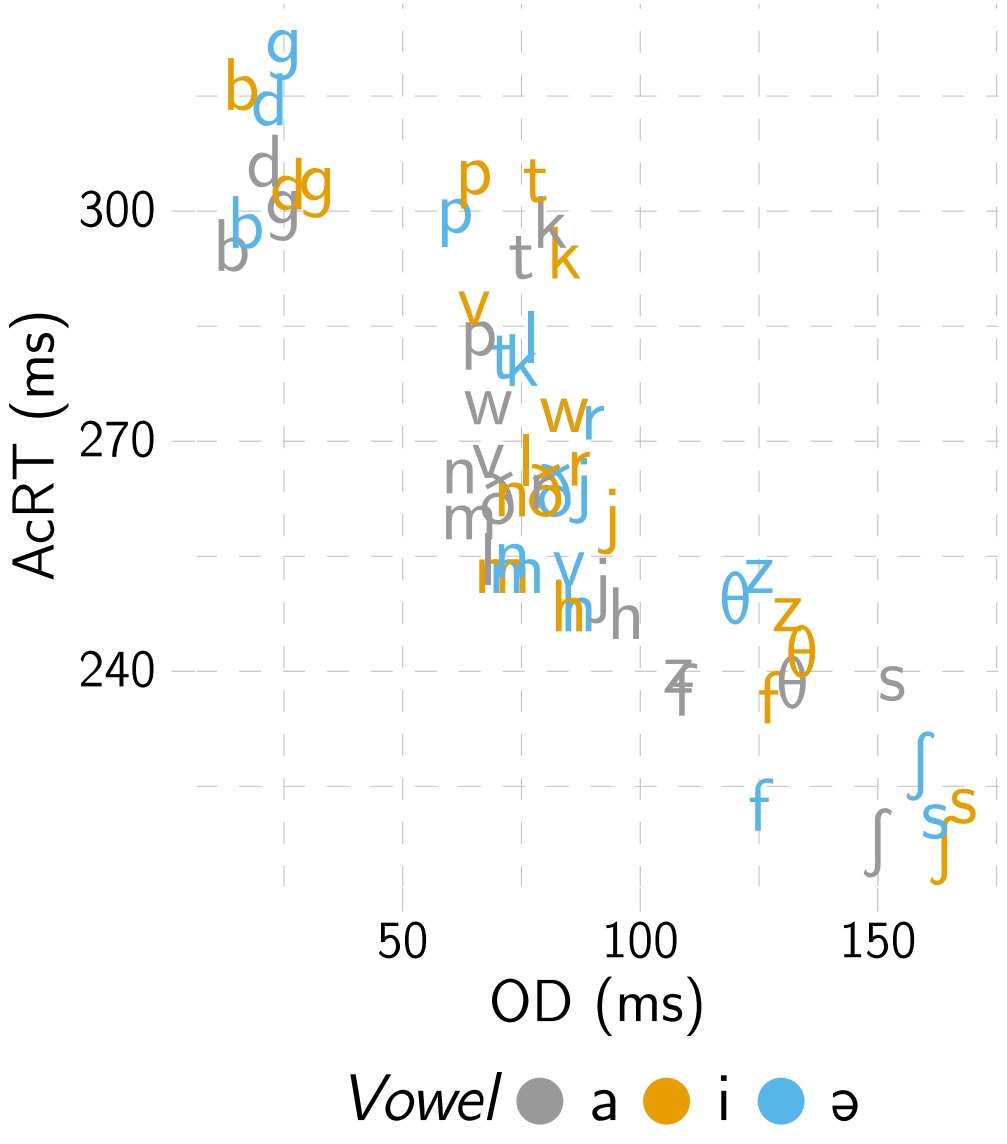
What do we know about lower limits of RTs?

- Chiu and Gick (2014) provide a conservative lower bound estimate for a vocal responses to auditory stimuli based on the STARTLE paradigm.



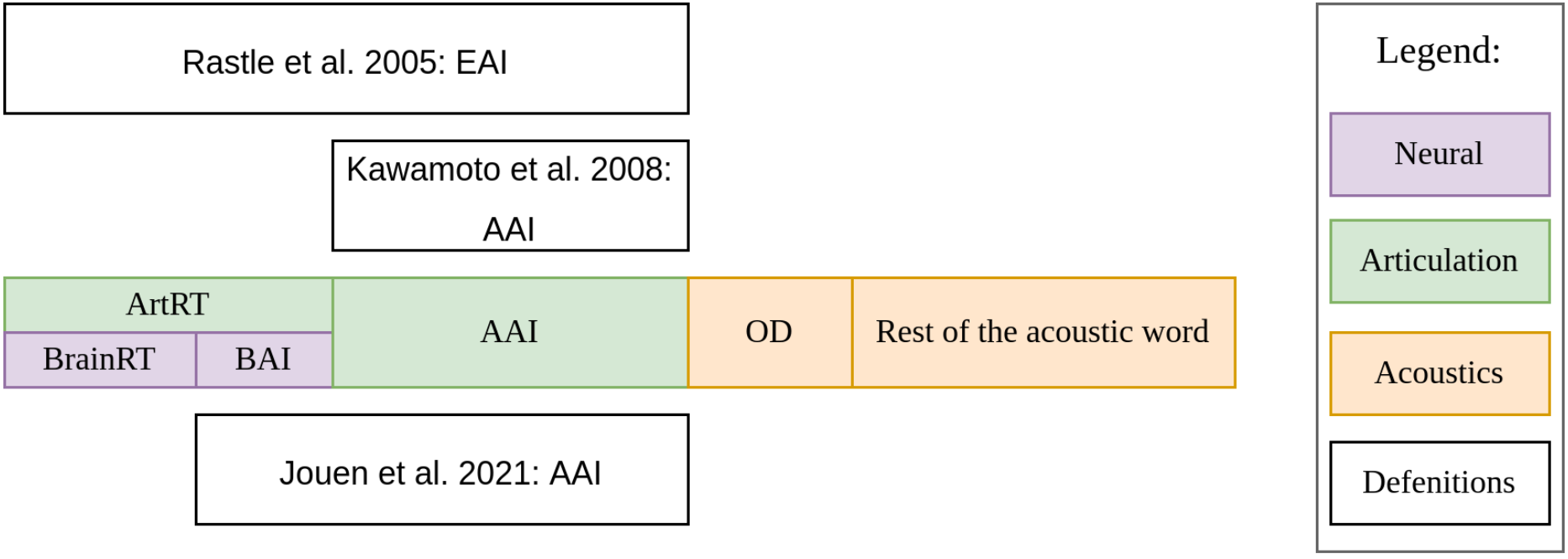
What do we know about acoustic reaction times?

- Rastle et al. (2005) measured delayed naming reaction times.
- They report acoustic RTs (AcRT) and onset durations (OD) for all phonotactically legal English consonant onsets.
- In the data acoustic RT is very neatly inversely correlated with OD.



Revisiting the definitions and some differences

Just so that we know what we mean by these acronyms. As far as I can tell this is how the different intervals in different articles – including Jouen et al. (2021) – relate to each other:



Research question

The inverse correlation pattern was first reported by Fowler (1979) with the regression coefficient also roughly -0.5. So, this begs the question:

Where does that inverse correlation pattern come from?

Materials

Experiment

- Partially replicates, but also expands the materials of the experiment by Rastle et al. (2005).
- Expands methods by adding ultrasound recording.
- /CCCVC/, /CCVC/, /CVC/, and /VC/ English lexical words were produced by two female speakers and one male speaker of Standard Scottish English.
- Measured with 2D ultrasound tongue imaging (at 120 fps) and synchronised sound recording.

Problems

- Segmentation is time and sanity consuming.
- Video segmentation might be faster than one would first think, but it is definetely at least as sanity consuming as one would think.
- Human annotators are inconsistent.
- Computer assisted segmentation can ease all of these problems.

Method

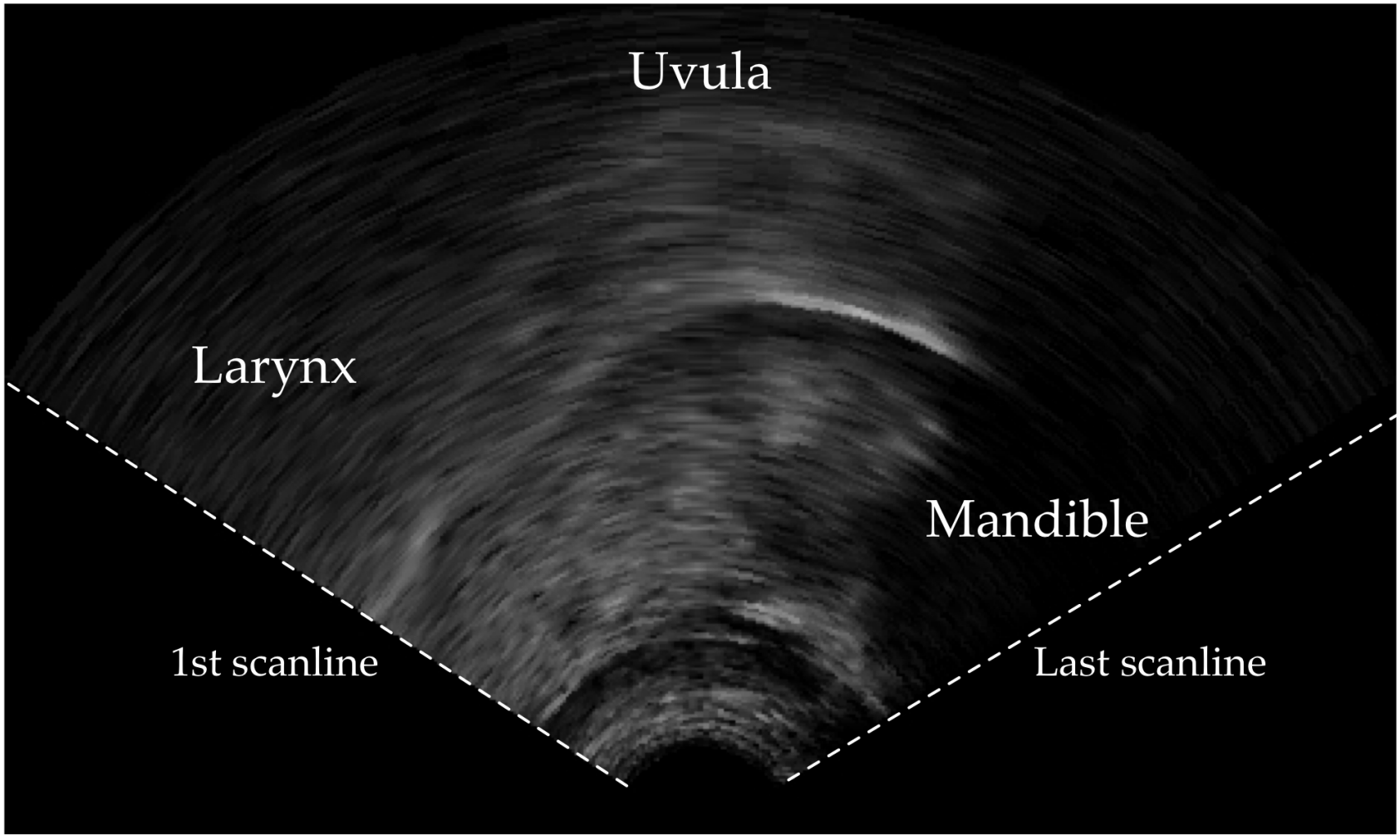
Background of the method

- The analysis methods presented here are similar to methods developed by
  - McMillan and Corley (2010), Drake et al. (2013) who used Euclidean distance on ultrasound frames and
  - Raeesy et al. (2011) who used a similar method on MRI data.
- However, we exploit knowledge of how ultrasound scanning works to produce more fine grained analysis.

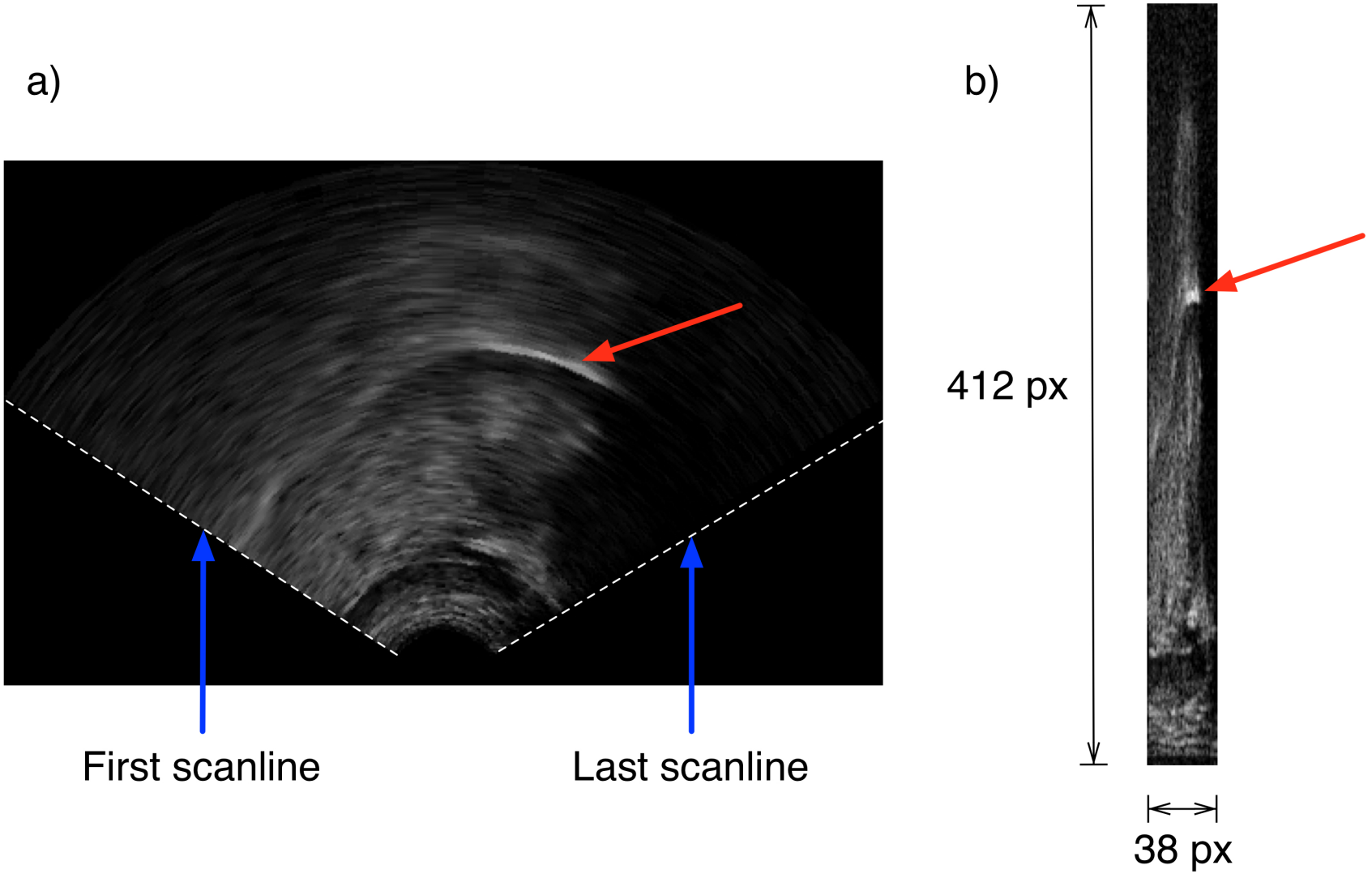
How is ultrasound data produced?

- An ultrasound pulse is sent out from the probe in one direction at a time.
- Echoes from that direction are listened to for a certain time (which depends on the probe and scanner settings).
- This produces one scanline.
- By repeating the process over a number of directions and interpolating the gathered data the system produces human readable images.
- The data is usually presented to humans in interpolated form.

Regular, interpolated ultrasound and anatomy



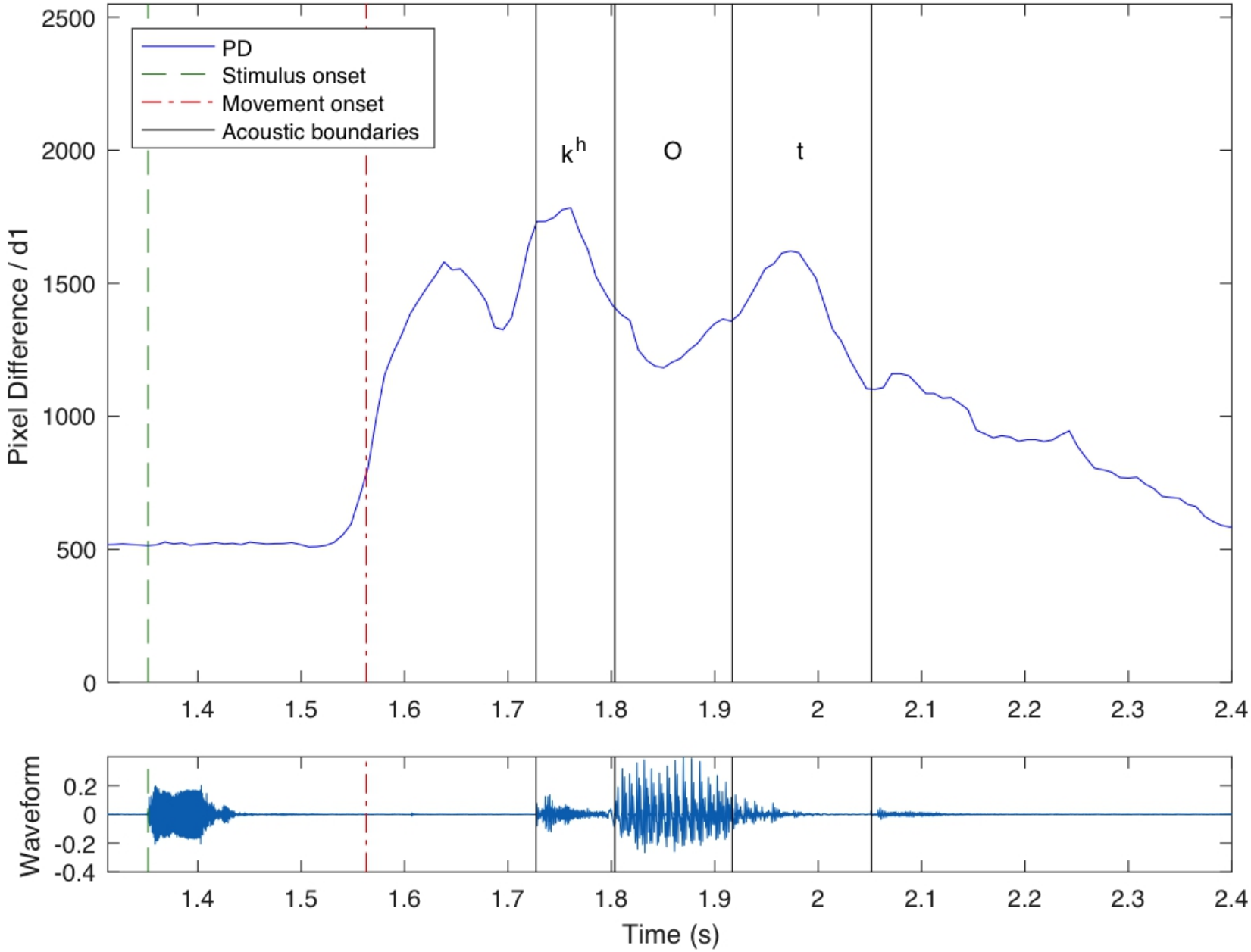
What is raw ultrasound data / probe return data?



Algorithms: Pixel Difference (PD)

- Interpret each raw frame as a huge N-dimensional vector ($n = 63 \times 256$ in our case).
- Calculate the Euclidean distances between these vectors.
- Result gives a holistic measure of change between two frames.
- The next slide will show that the resulting curves are easy to annotate manually – and illustrates a typical case where manual video annotation gives a later movement onset time.
- Automatic annotation required an extra step (we'll get back to that in a moment).

Algorithms: Pixel Difference (PD)

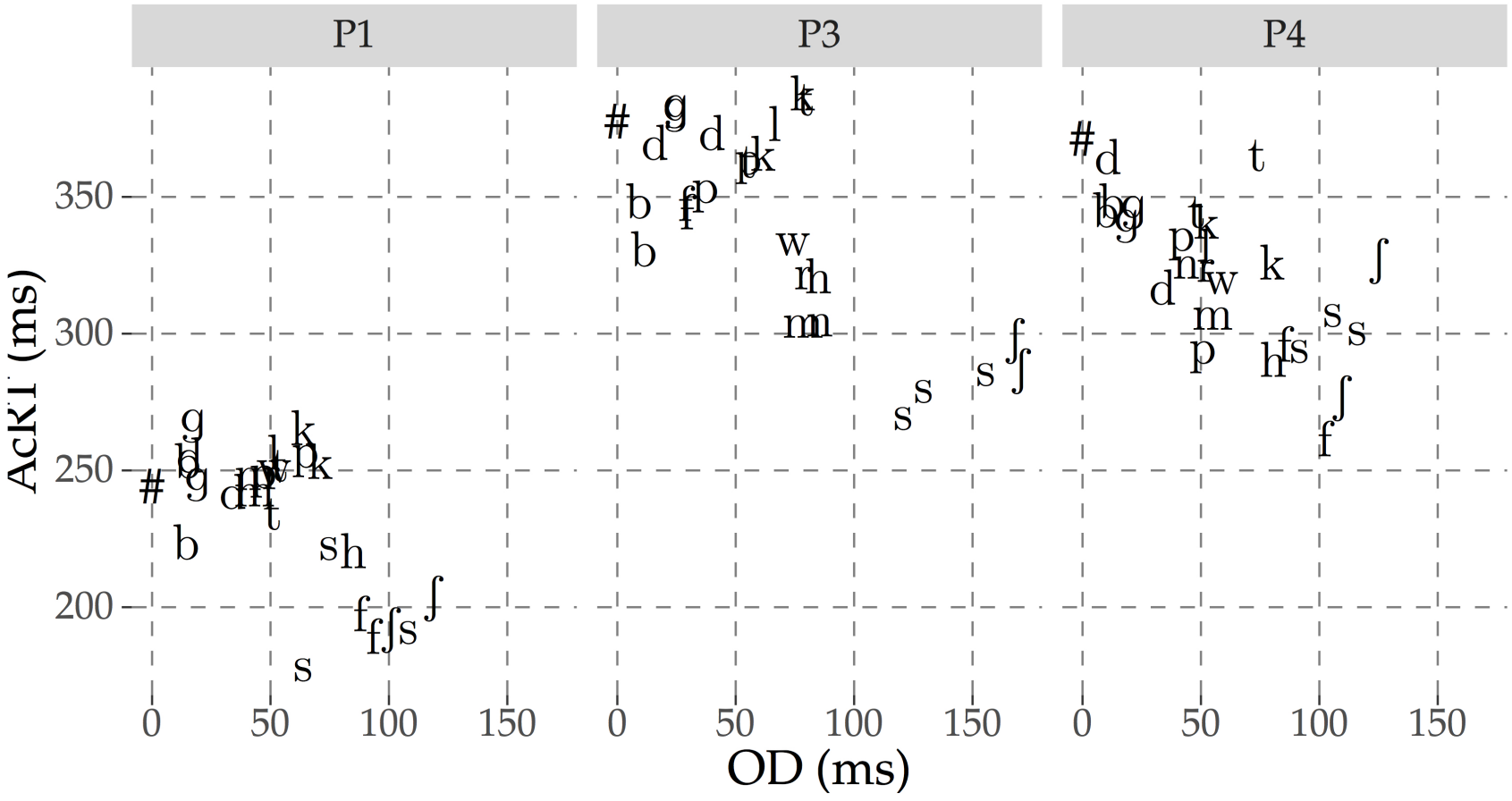


Algorithms 2: Scanline Based Pixel Difference (SBPD)

- Take each of the scanlines (63 in our case) as a smaller M-dimensional vector ($m = 256$ in our case).
- Calculate the Euclidean distances between corresponding scanlines between frames.
- Result gives a measure of change between two frames with location (in sagittal plane, back to front) presented by the scanline number.
- By applying dynamic time warping with a test function to each scanline's PD curve we get a vector of 63 onset times for each recording.
- Taking the median of the individual onset candidates has proven to be a reliable way of automating onset detection (Palo 2019).

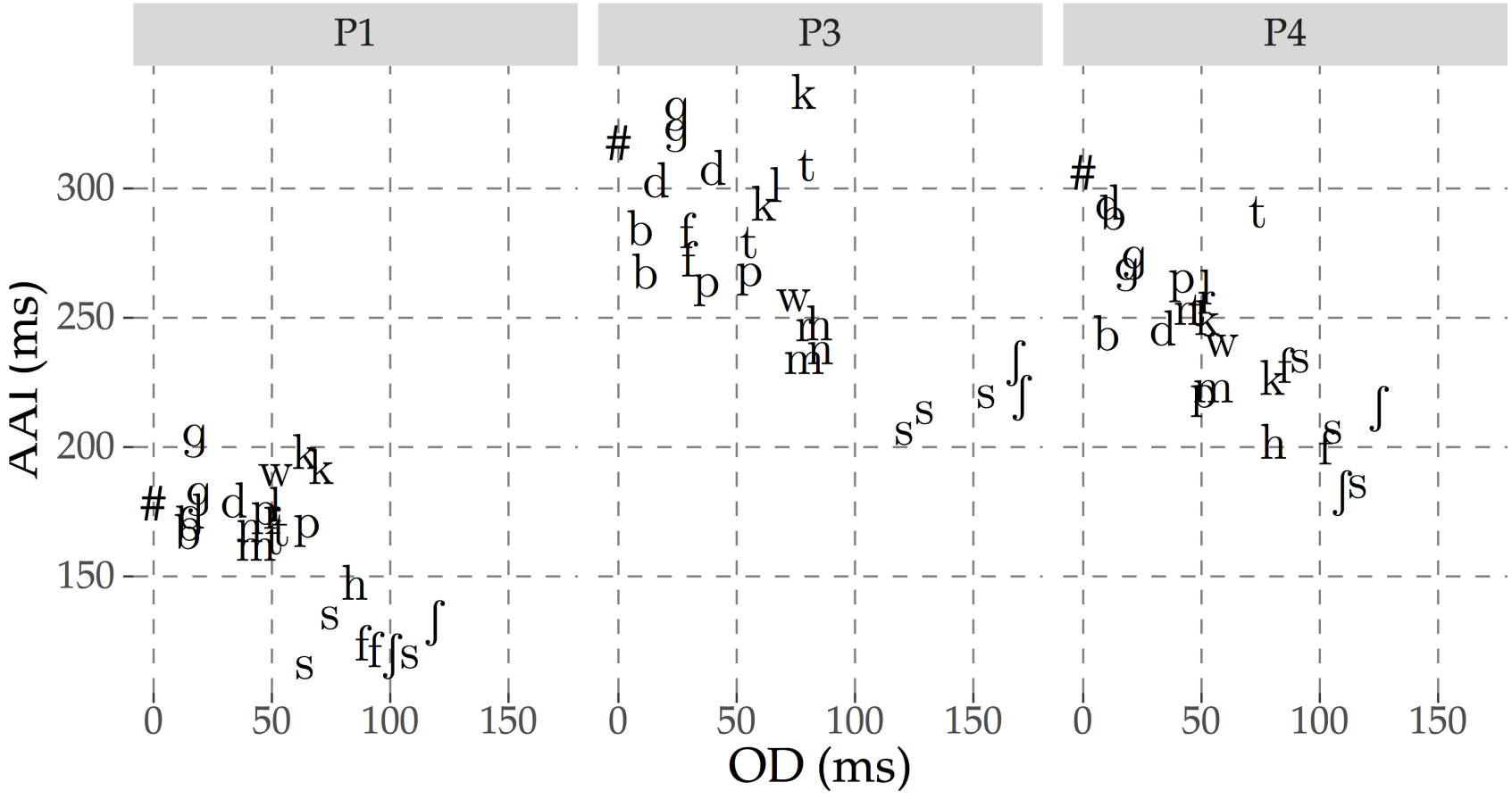
Results

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.

Results: AAI



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.

Results: Statistical modelling

Fitting linear mixed models to the data with a step-up model selection procedure, we have for Acoustic Reaction Time[1]:

$$AcRT \sim OD + RhymeDur + trial + (1|id) + (1|word), \quad (1)$$

or as a (general) prediction model:

$$AcRT = -0.42 \times OD + 0.20 \times RhymeDur + 0.06 \times trial. \quad (2)$$

And for Articulatory to Acoustic Interval (AAI)[2]:

$$AAI \sim OD + RhymeDur + trial + (1|id) + (1|word), \quad (3)$$

or as a (general) prediction model:

$$AAI = -0.40 \times OD + 0.31 \times RhymeDur + 0.12 \times trial. \quad (4)$$

---

[1] P-value ≈ 0.04548 when comparing against a model without a trial term. Previous step – adding RhymeDur – had a P-value $\approx 7.616 \times 10^{-9}$.

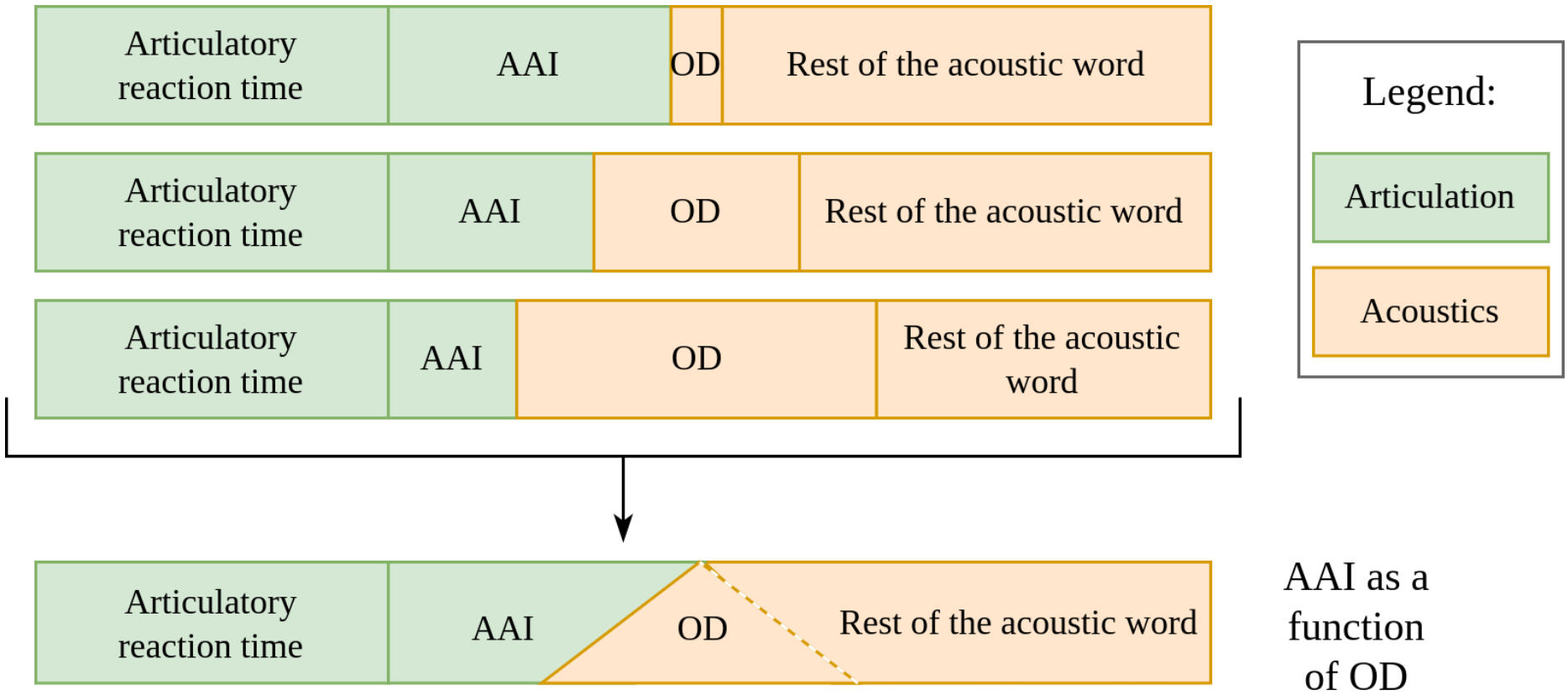
[2] P-value ≈ 0.0003768 compared against a model without a trial term.

Results: Statistical modelling (continued)

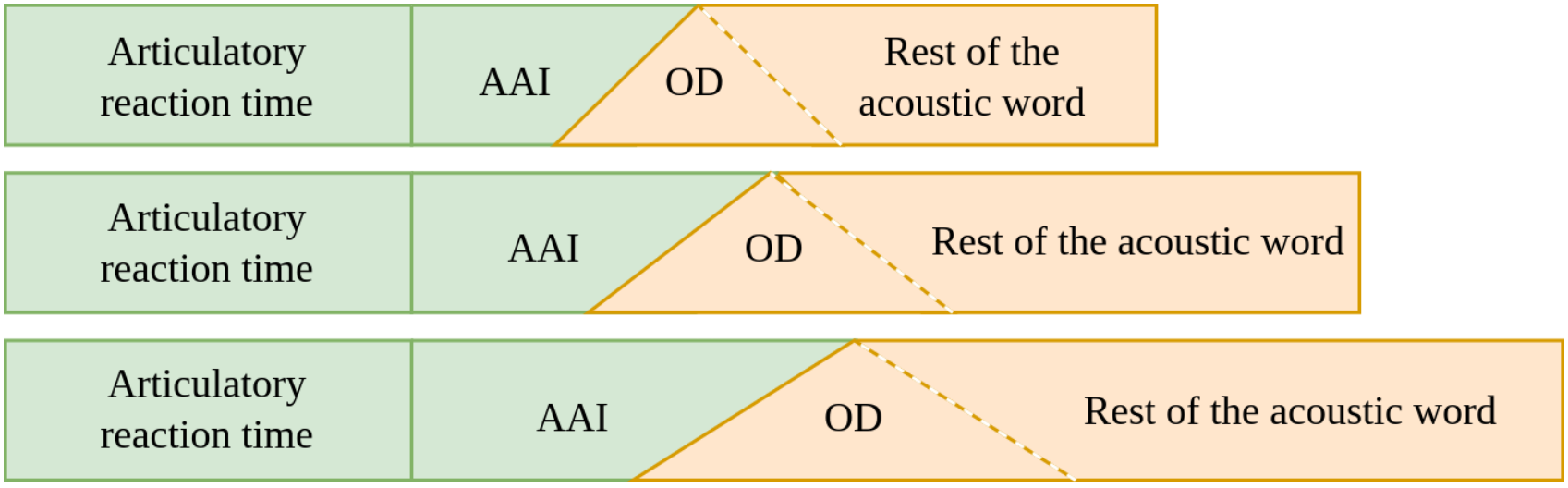
- Acoustic data repeats the general inverse correlation pattern of previous studies with some refinements.
- AAI repeats the correlation pattern and is likely its actual locus.
- No vowel quality effects were found.
- In contrast, articulatory reaction time could not be predicted with utterance duration variables and in particular did not have any statistically significant correlation with OD. Instead it seems to be a noisy constant ≈ 120 ms in this dataset.

Tentative theory

Effect of OD on AAI



Effect of articulatory rate on AAI



Conclusions

- The present results are in agreement with the literature in terms of acoustics.
- The inverse correlation pattern of Acoustic Reaction Time and Onset Duration originates int the inverse correlation of AAI and Onset Duration.
- The modelling results on AAI lead to the conclusion that, in terms of timing, the silent articulation preceding an utterance should be considered part of speech.

References I

Chiu, C. and Gick, B. (2014). Startling speech: eliciting prepared speech using startling auditory stimulus. *Frontiers in Psychology*, 5(1082).

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.

Faytak, M., Moisik, S. R., and Palo, P. (2020). The speech articulation toolkit (satkit): Ultrasound image analysis in python. In *Proceedings of the 12th International Seminar on Speech Production (ISSP 2020)*, pages 234 – 237, Online / New Haven, CT.

Fowler, C. (1979). "Perceptual centers" in speech production and perception. *Perception & Psychophysics*, 25(5):375 – 388.

Fowler, C. and Tassinary, L. (1981). Natural measurement criteria for speech: The anisochrony illusion. In Long, J. and Baddeley, A., editors, *Attention and Performance IX*, pages 521 – 535. The International Association For The Study of Attention and Performance.

Jouen, A., Lancheros, M., and Laganaro, M. (2021). Microstate erp analyses to pinpoint the articulatory onset in speech production. *Brain Topography*, 34:29 – 40.

Kawamoto, A. H., Liu, Q., Mura, K., and Sanchez, A. (2008). Articulatory preparation in the delayed naming task. *Journal of Memory and Language*, 58(2):347 – 365.

References II

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. and Moisik, S. R. and Faytak, M. (2020). SATKIT: Speech Articulation ToolKIT [Python software package]. Available in a public software repository, accessed 1 Feb 2021. `https://github.com/giuthas/satkit`.

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. and Davis, M. H. (2002). On the complexities of measuring naming. *Journal of Experimental Psychology: Human Perception and Performance*, 28(2):307 – 314.

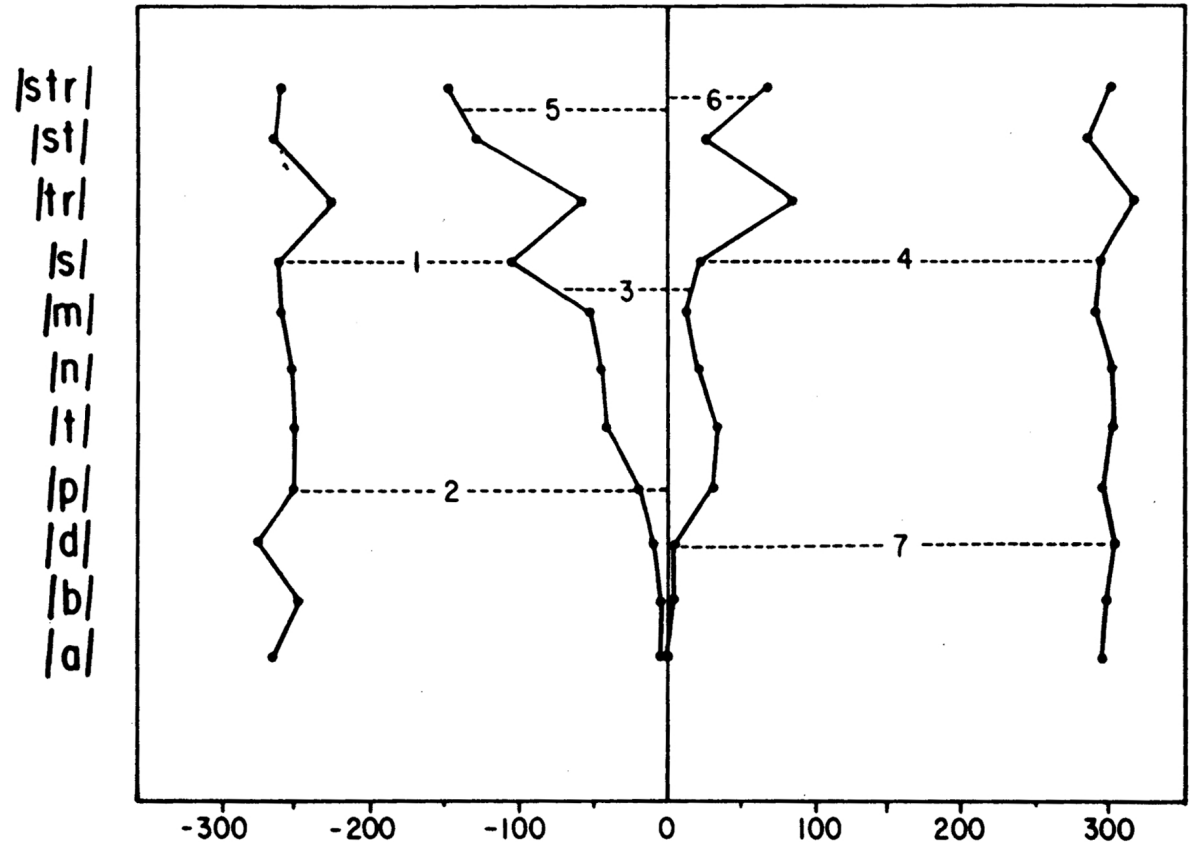
Rastle, K., Harrington, J. M., Croot, K. P., and Coltheart, M. (2005). Characterizing the motor execution stage of speech production: Consonantal effects on delayed naming latency and onset duration. *Journal of Experimental Psychology: Human Perception and Performance*, 31(5):1083 – 1095.

Roon, K. D. (2013). *The dynamics of phonological planning*. PhD thesis, New York University.

Thank you!

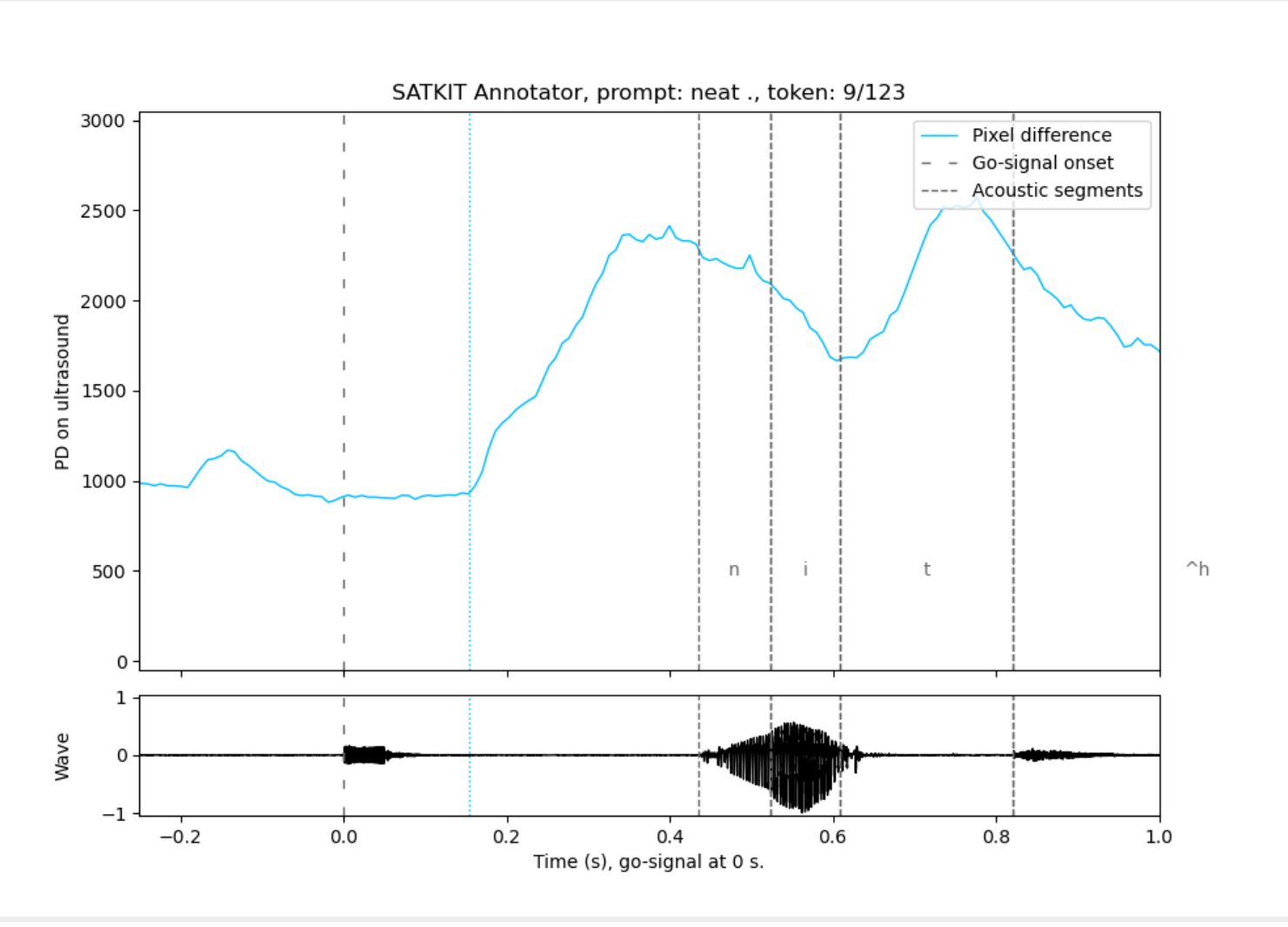
Also many thanks to everybody who's helped along the way:
Alan Wrench and Steve Cowen with data and recordings,
my PhD supervisors Sonja Schaeffler and Jim Scobbie,
my wonderful participants, and many many more.

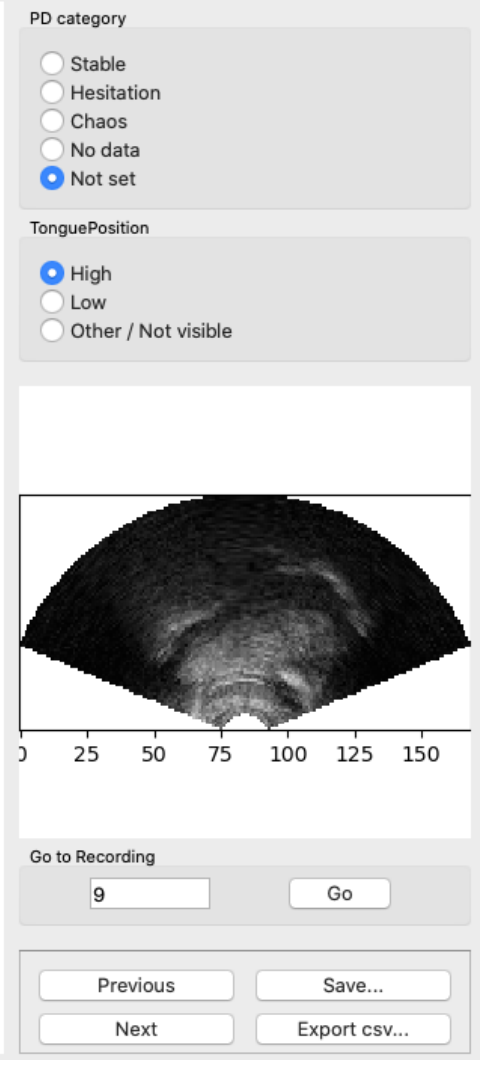
Connection with speech cycling



- Figure from Fowler and Tassinary (1981).
- Metronome click at 0 ms.
- Relevant intervals are: 1. silence between words, 3. onset consonant, and 4. vowel.

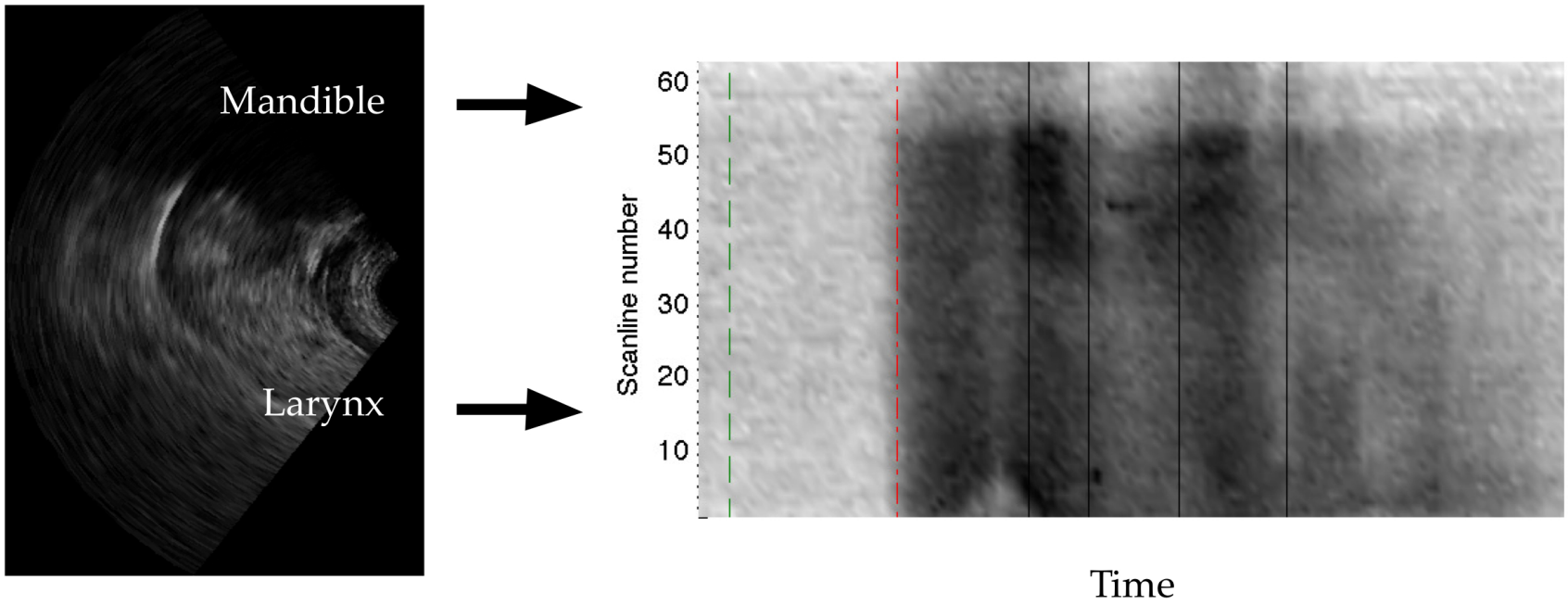
Annotating Pixel Difference (PD)



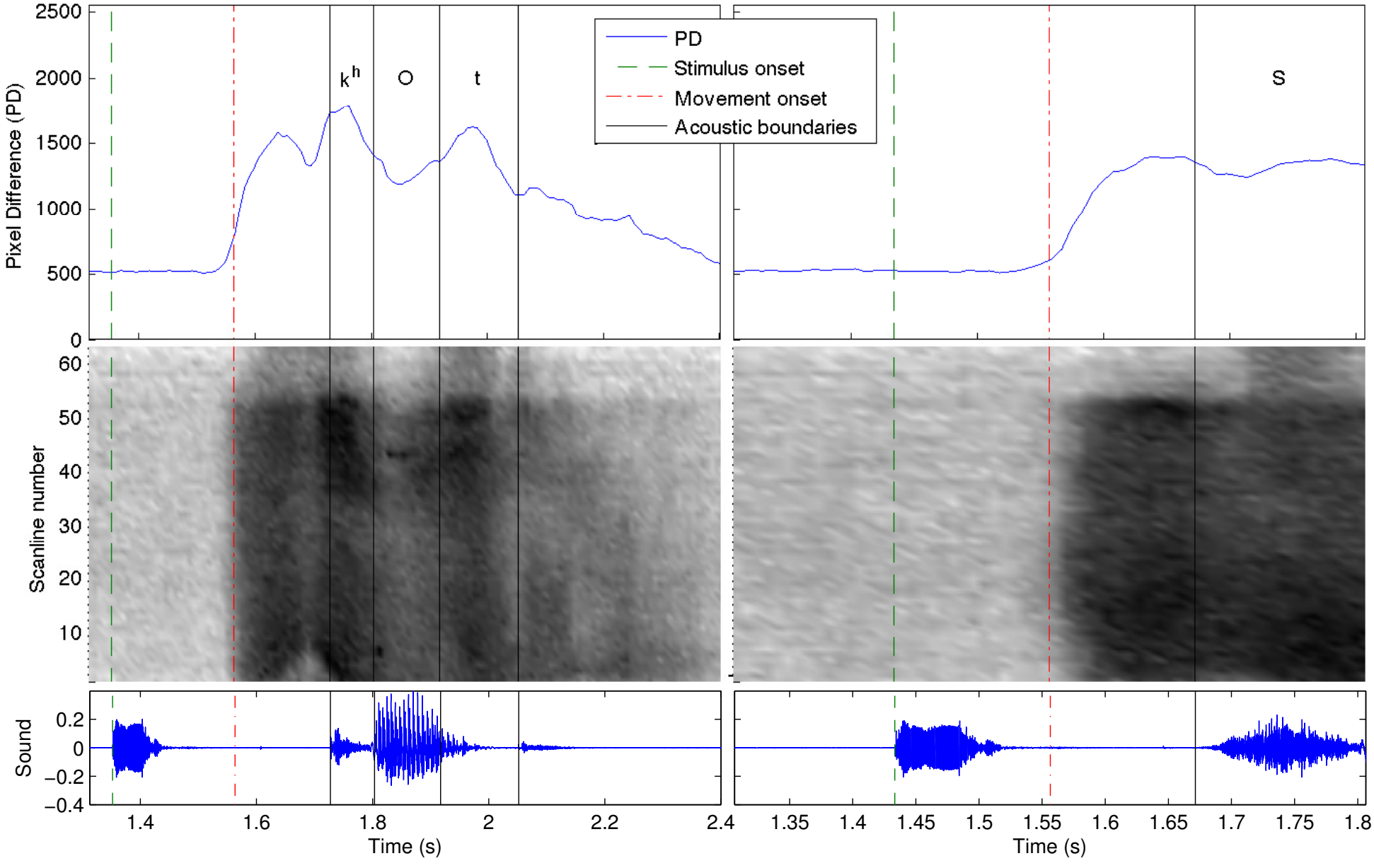


SBPD and anatomy

The SBPD-gram on the rigth shows change over time for each scanline.

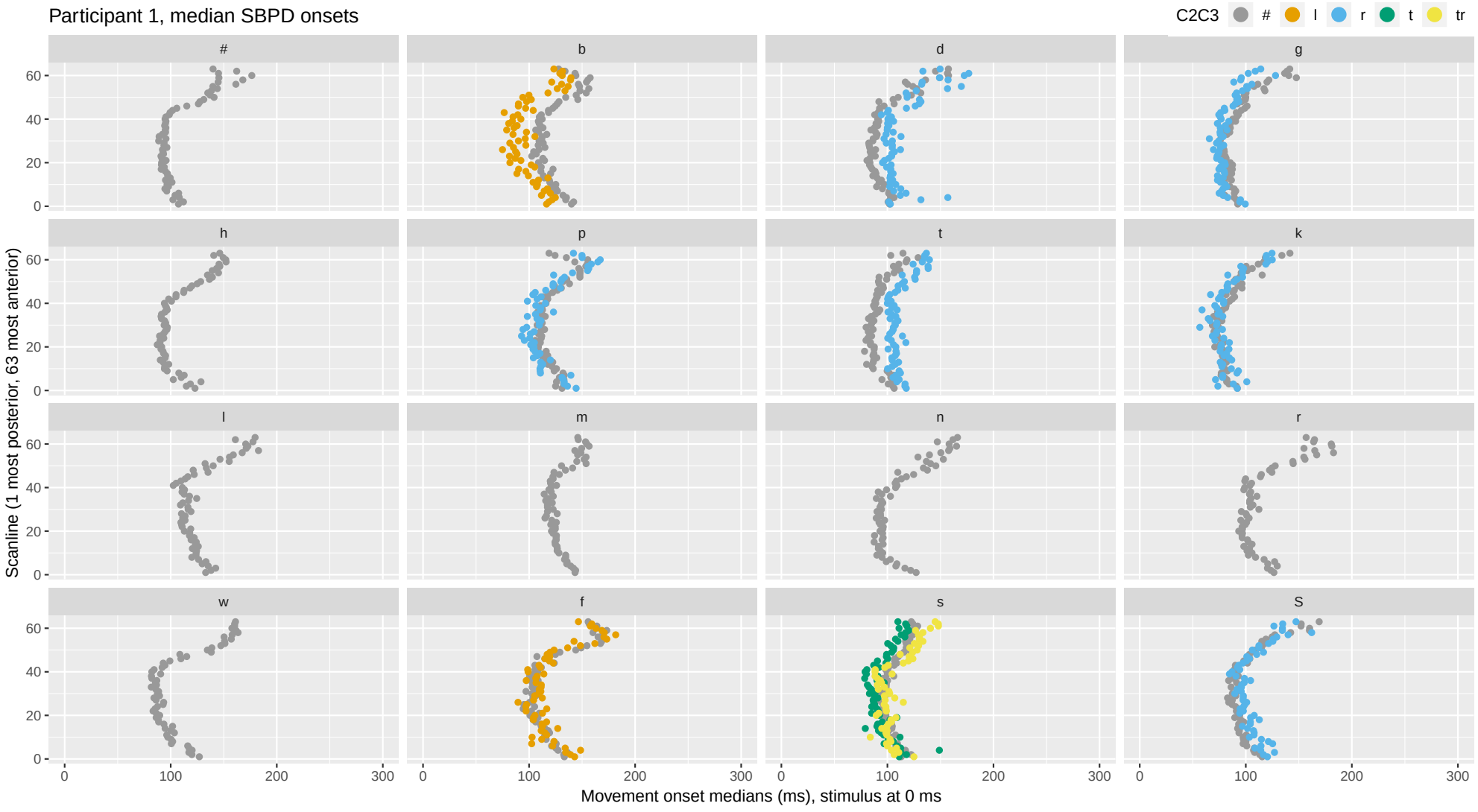


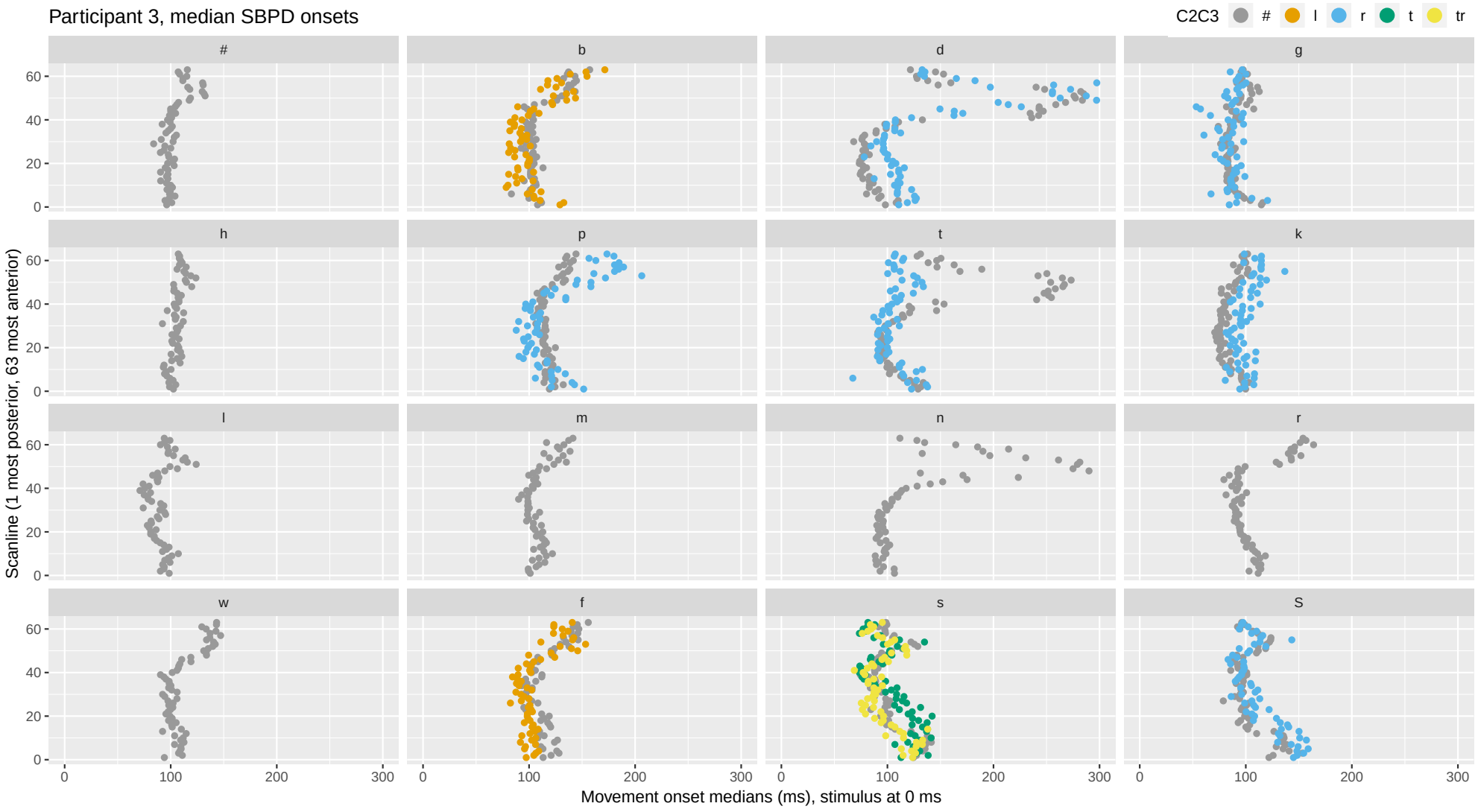
Examples of PD and SBPD

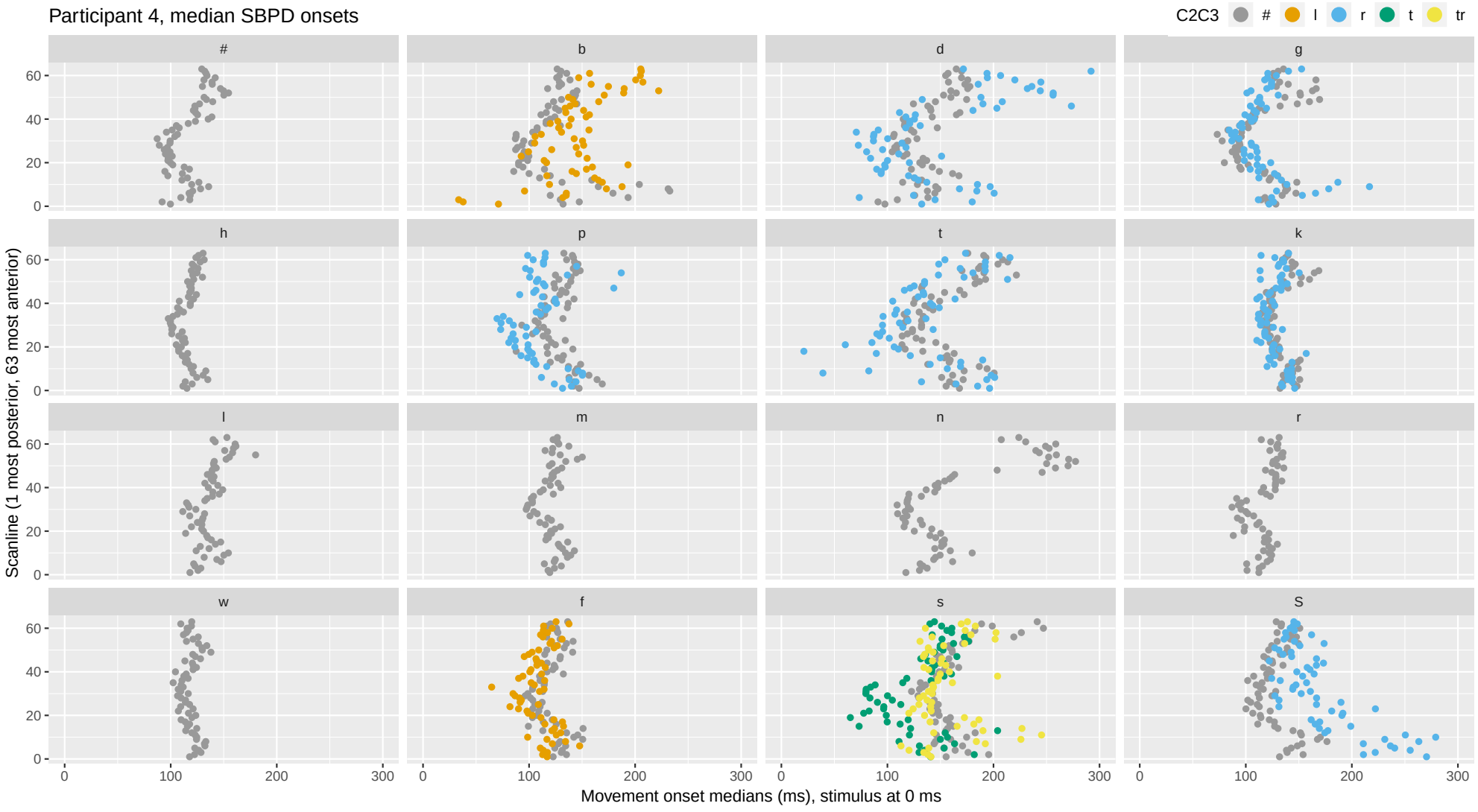


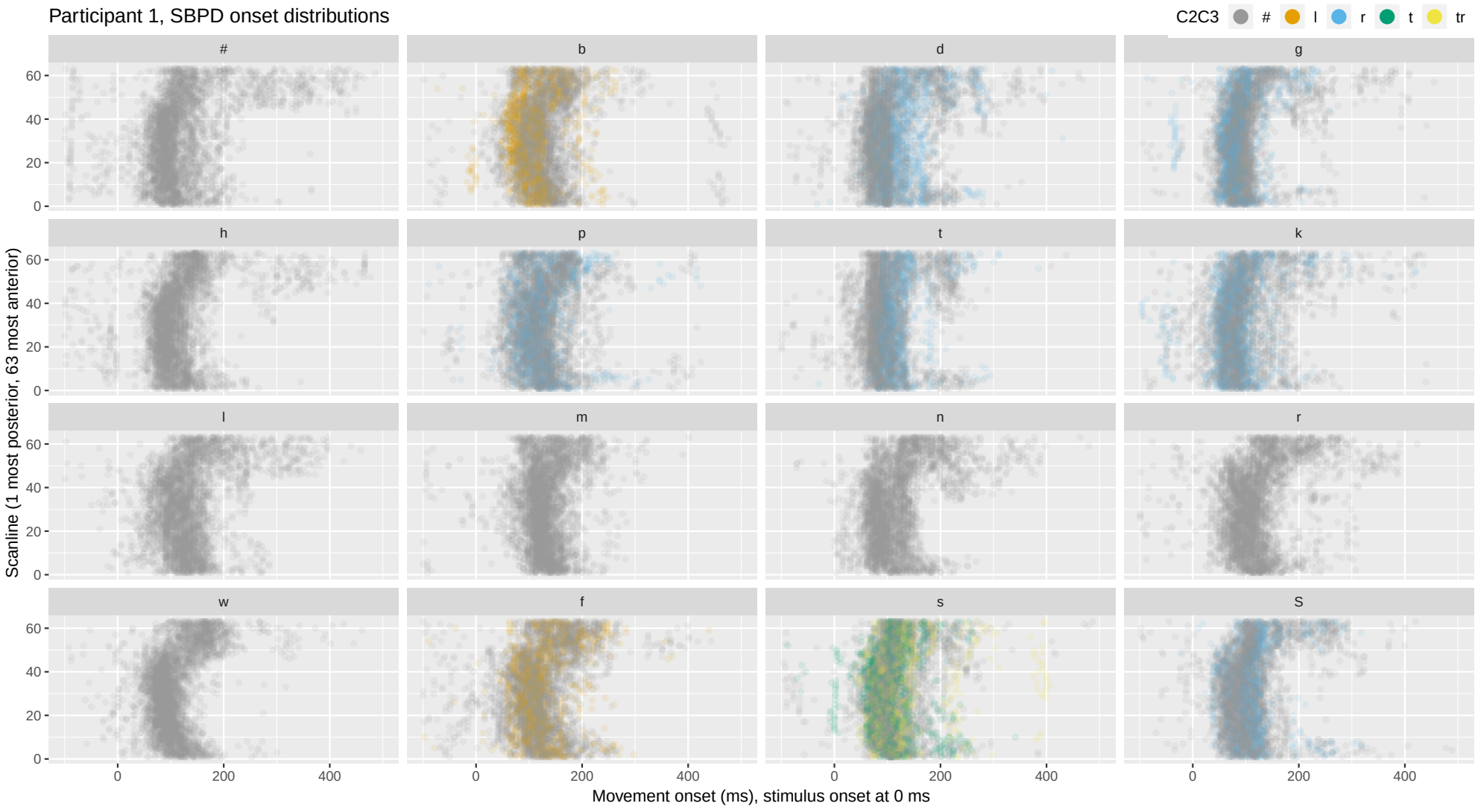
Analysing location of first movement

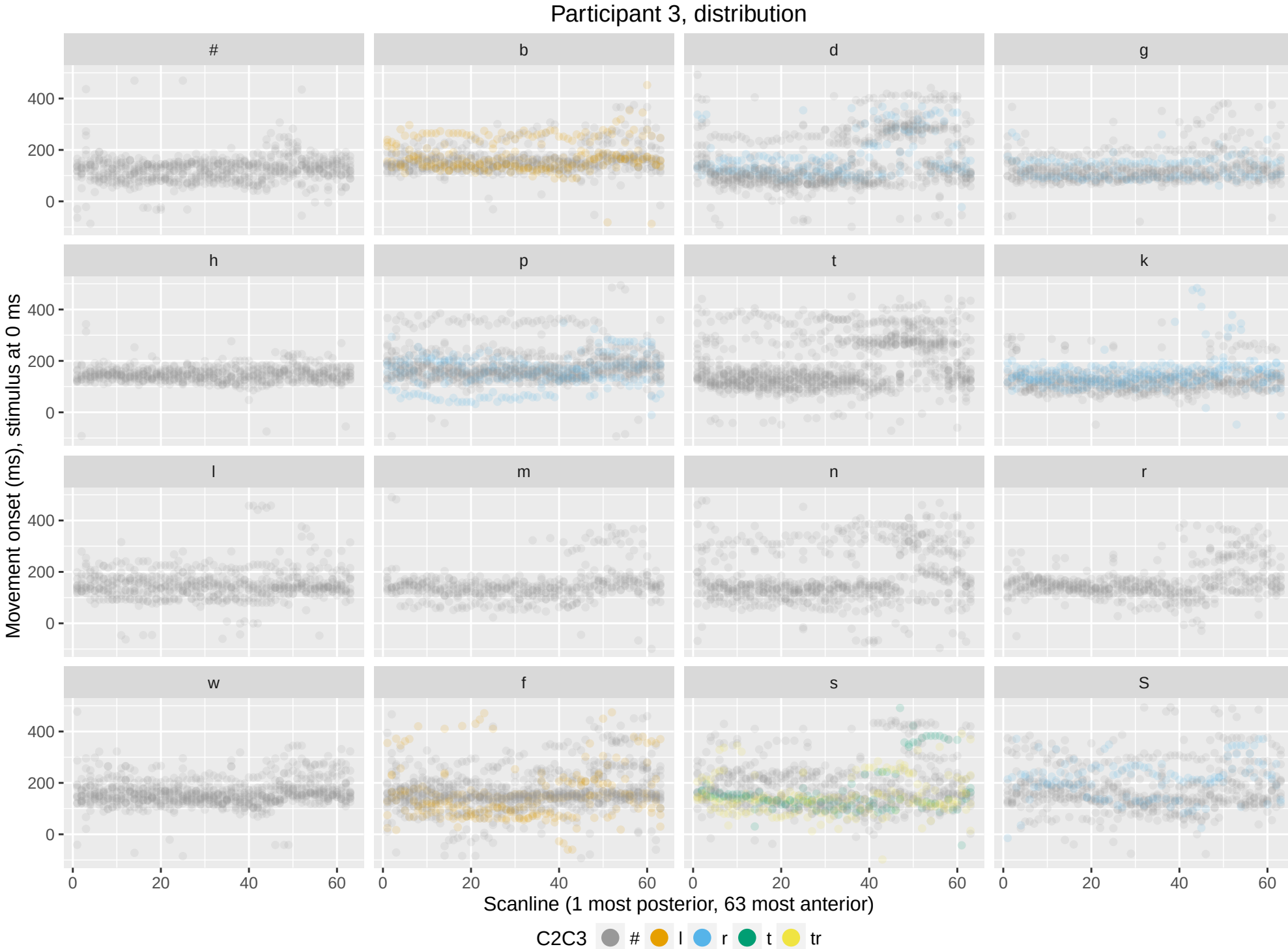
- Calculate SBPD for each token.
- Run DTW for each scanline of each token.
- Merge data from acoustic annotation, manual annotation and SBPD/DTW analysis.
- Remove tokens that were manually identified as false starts.
- Take median of onsets identified by DTW over tokens for each scanline.
- Plot results.
- (Scratch head.)

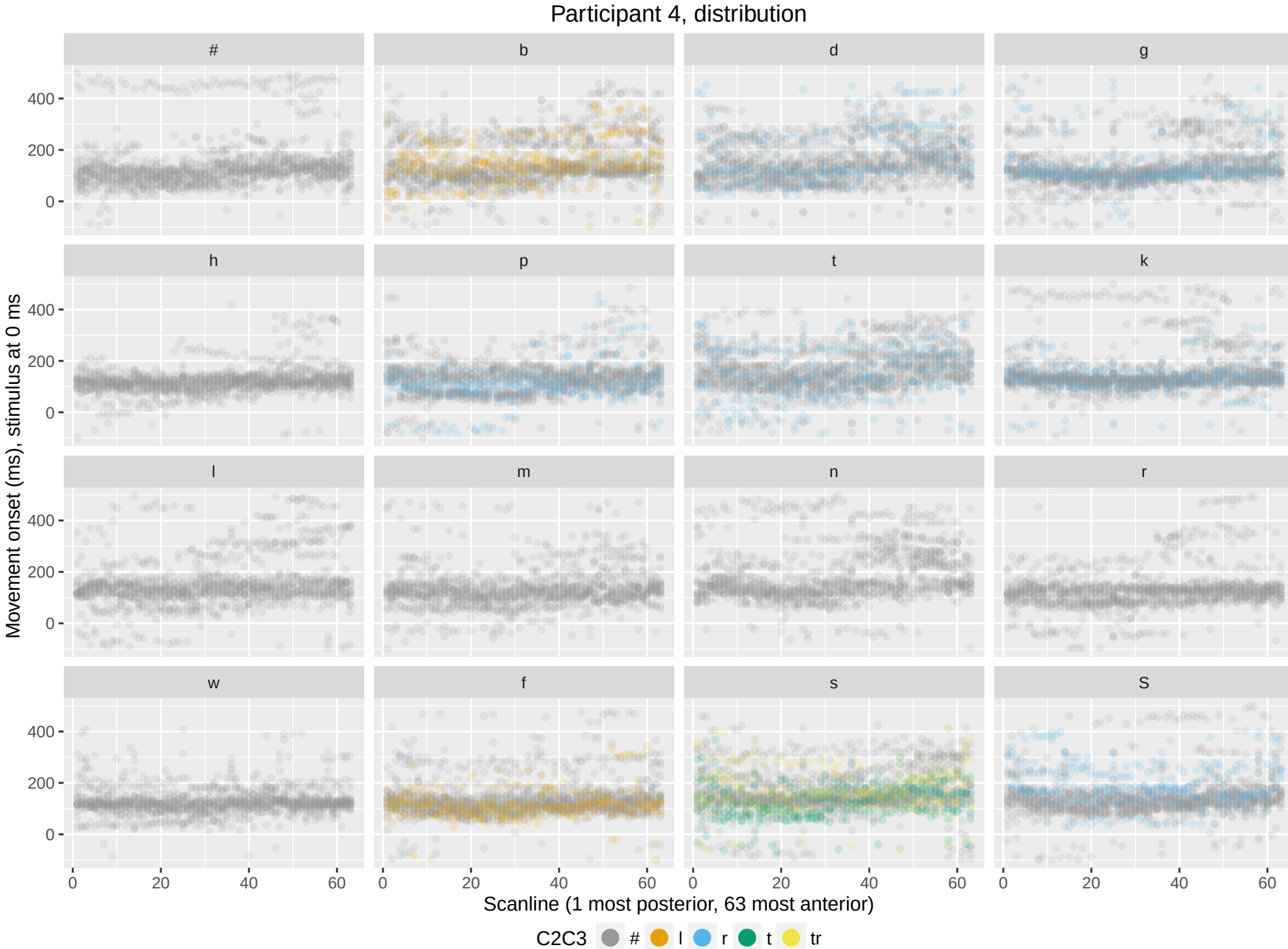

Participant 1, median SBPD onsets
C2C3
#
l
r
t
tr
#
b
d
g
h
p
t
k
l
m
n
r
w
f
s
S
Scanline (1 most posterior, 63 most anterior)
0
20
40
60
0
100
200
300
Movement onset medians (ms), stimulus at 0 ms


Participant 3, median SBPD onsets
C2C3
#
l
r
t
tr
#
b
d
g
h
p
t
k
l
m
n
r
w
f
s
S
Scanline (1 most posterior, 63 most anterior)
0
20
40
60
0
100
200
300
Movement onset medians (ms), stimulus at 0 ms


Participant 4, median SBPD onsets
C2C3
#
l
r
t
tr
#
b
d
g
h
p
t
k
l
m
n
r
w
f
s
S
Scanline (1 most posterior, 63 most anterior)
Movement onset medians (ms), stimulus at 0 ms
0
20
40
60
0
100
200
300


Participant 1, SBPD onset distributions
C2C3
#
l
r
t
tr
#
b
d
g
h
p
t
k
l
m
n
r
w
f
s
S
Scanline (1 most posterior, 63 most anterior)
0
20
40
60
0
200
400
Movement onset (ms), stimulus onset at 0 ms


Participant 3, distribution
#
b
d
g
h
p
t
k
l
m
n
r
w
f
s
S
400
200
0
0
20
40
60
Movement onset (ms), stimulus at 0 ms
Scanline (1 most posterior, 63 most anterior)
C2C3
#
l
r
t
tr


Participant 4, distribution
#
b
d
g
h
p
t
k
l
m
n
r
w
f
s
S
400
200
0
0
20
40
60
Movement onset (ms), stimulus at 0 ms
Scanline (1 most posterior, 63 most anterior)
C2C3
#
l
r
t
tr

Relation of an interpolated ultrasound frame to anatomy