Can we detect initiation of tongue internal changes before overt movement onset in ultrasound?

Pertti Palo

CASL Research Centre, Queen Margaret University, Edinburgh, Scotland, UK

pertti.palo@taurlin.org

Abstract

In order to understand speech articulation, we need to understand not only what movements of the articulators are used to produce a given sound, but also how those articulator movements are produced by muscle actions. This paper approaches this problem by analysing ultrasound data with three methods. First, Pixel Difference accounts for all change apparent in tongue ultrasound data (Palo 2019; Palo, P. and Moisik, S. R. and Faytak, M. 2020), second, two methods which evaluate the distance between tongue contour splines: Average Nearest Neighbour Distance (Zharkova and Hewlett 2009) and a novel method called Median Point-by-Point Distance. The results show that while there maybe a small delay between tongue internal changes and movement of the tongue contour it lies within the margin of error in the current data and is unlikely to be significant. Further details are provided on the performance of the two spline metrics.

Keywords: Pre-speech articulation, ultrasound tongue imaging, pixel difference, nearest neighbour distance, automatic data analysis

1. Introduction

Speech initiation can be used as a window into how the articulator movements of speech are produced by muscle actions. Furthermore, direct measurement of articulation or muscular activation gives a more detailed (Kawamoto et al. 2008; Linden et al. 2014) and a more appropriate method of evaluating speech reaction times than acoustics.

Among articulatory measurement methods tongue ultrasound is currently one of the most popular. While tongue contour extraction is the most common method of analysing tongue ultrasound (Stone 2005; Davidson 2006; Mielke 2015), recently methods that analyse the whole ultrasound image have received attention (McMillan and Corley 2010; Drake, Schaeffler, and Corley 2013; Palo 2019; Palo, P. and Moisik, S. R. and Faytak, M. 2020; Faytak, Moisik, and Palo 2020; Saito et al. 2020).

In seeking to make measurement of articulatory reaction times fast and reliable Palo (2019) used a Euclidean distance based metric called Pixel Difference (PD) to analyse ultrasound data. PD based reaction times are on average almost 40 ms shorter than those measured by manually annotating ultrasound videos (Figure 3.9 in Palo 2019). Figure 1 shows a typical example from Palo (2019). We can see that the red dashed-dotted line which marks the manual video based movement onset is well after the point where the PD curve starts to rise.

In this context, the question arises if the difference in the reaction time measures is due to manual annotators relying on contour movement, while PD based reaction times are the beginning of any significant change. In other words, can we detect tongue internal changes before movement of the tongue contour?

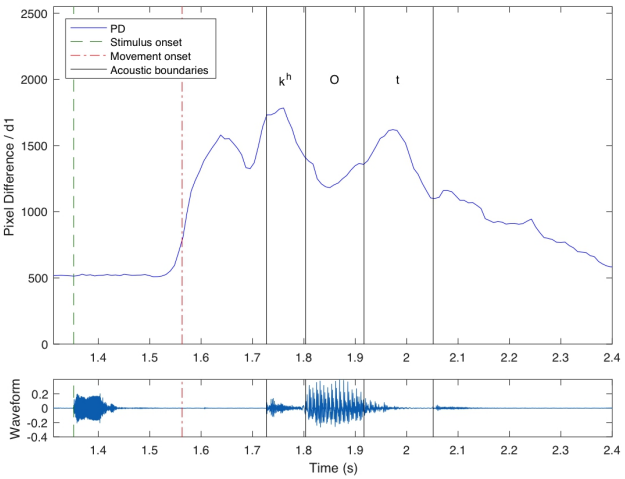

Figure 1: *An example of Pixel Difference (PD): the utterance 'caught'. Green dashed line indicates the 'go'-signal, red dashed-dotted line movement onset as annotated on the ultrasound video by the author, and black solid lines the acoustic segmentation.*

Working from the hypothesis that the difference is because human annotators react to *tongue contour movement*, while PD measures *all change* including changes in the speckle, we need a method for evaluating tongue contour movement. The Average Nearest Neighbour Distance (ANND) (Zharkova and Hewlett 2009) was the first candidate for a spline change metric, but eventually a novel metric called Median Point-by-point Distance (MPBPD) proved better suited for this problem.

2. Materials

The speech materials come from two delayed naming experiments, which were recorded with the high-speed ultrasound facility at Queen Margaret University. In the first experiment – Experiment 3 of Palo (2019) – all phonotactically valid Finnish /CV/ syllables were produced by the author. In the second experiment – Experiment 2 of Palo (2019) – lexical /CVC/ words were produced by speakers of Standard Scottish English. The materials analysed here come from a young adult male speaker.

In both experiments the participants were asked to

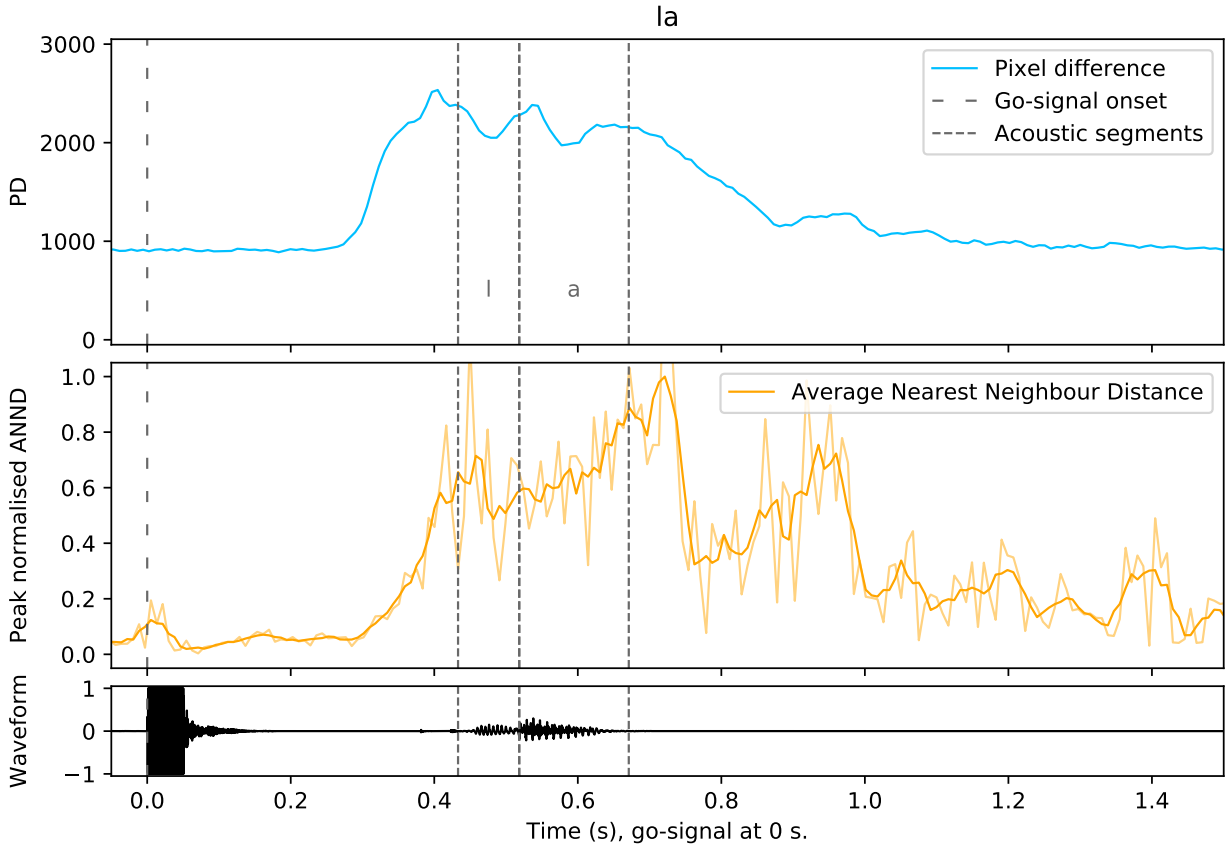

Figure 2: *PD in the top panel, peak normalised Average Nearest Neighbour Distance (ANND) in the middle panel, and acoustic waveform in the bottom panel. In all panels, the widely spaced dashed line at 0 s marks the 'go'-signal and the tightly spaced dashed lines mark acoustic segment boundaries. In the middle panel the fainter, less stable curve is the non-filtered ANND.*

remain at rest until they heard the go signal – a 1 kHz pure tone – and then produce the target syllable as soon and as accurately as possible. Ultrasound was captured at 120 fps and FOV was 137 degrees. Results and further details of both experiments have been published in Palo (2019). We present selected examples from the first data set and statistics from about 189 automatically splined tokens in the second data set. In both datasets, the ultrasound data was automatically splined in AAA (*Articulate Assistant Advanced User Guide: Version 2.14* 2012) and the splines were hand corrected up to the point where movement had clearly begun.

3. Methods

3.1. Pixel Difference (PD)

Pixel Difference (PD) is a change metric which can be used on any pixelated data. In this study, we use PD on raw ultrasound frames (probe return data). PD is the Euclidean distance between consecutive frames where each frame is interpreted as an N-dimensional vector (N is the number of pixels in the raw ultrasound frames). In many cases (e.g. Figures 1 and 2) PD provides a clear view of articulatory gestures and is particularly useful in identifying articulatory onset.

3.2. Choosing a spline change metric

Average Nearest Neighbour Distance (ANND) is a distance metric for two groups of points. It is based on the Nearest Neighbour Distance (NND), which is calculated for an individual point in relation to a comparison group of points (Zharkova and Hewlett 2009). The Nearest Neighbour Distance of a point to the comparison group is defined as the distance between the point and the nearest point in the comparison group. In general use, the distance can be defined by any distance metric. In our case, we use Euclidean distance. ANND is the average of the Nearest Neighbour Distances of group one when compared with group two. In this study, the point groups are 2D spline sample points of individual ultrasound frames (Figure 2).

To produce change curves that are analysable, the spline metrics were computed using a time step of 3 – that is comparing each frame to the third one after it instead of the immediately following one like in the case of PD. The curves were also smoothed with a moving average filter with a window length of 5 frames. To facilitate comparison of different metrics, the spline metrics in Figures 2 and 3 have been peak normalised by scaling the highest peak of each metric to 1.

ANND rarely behaves as well as it does in the example in Figure 2. Instead, in many tokens it shows a lot

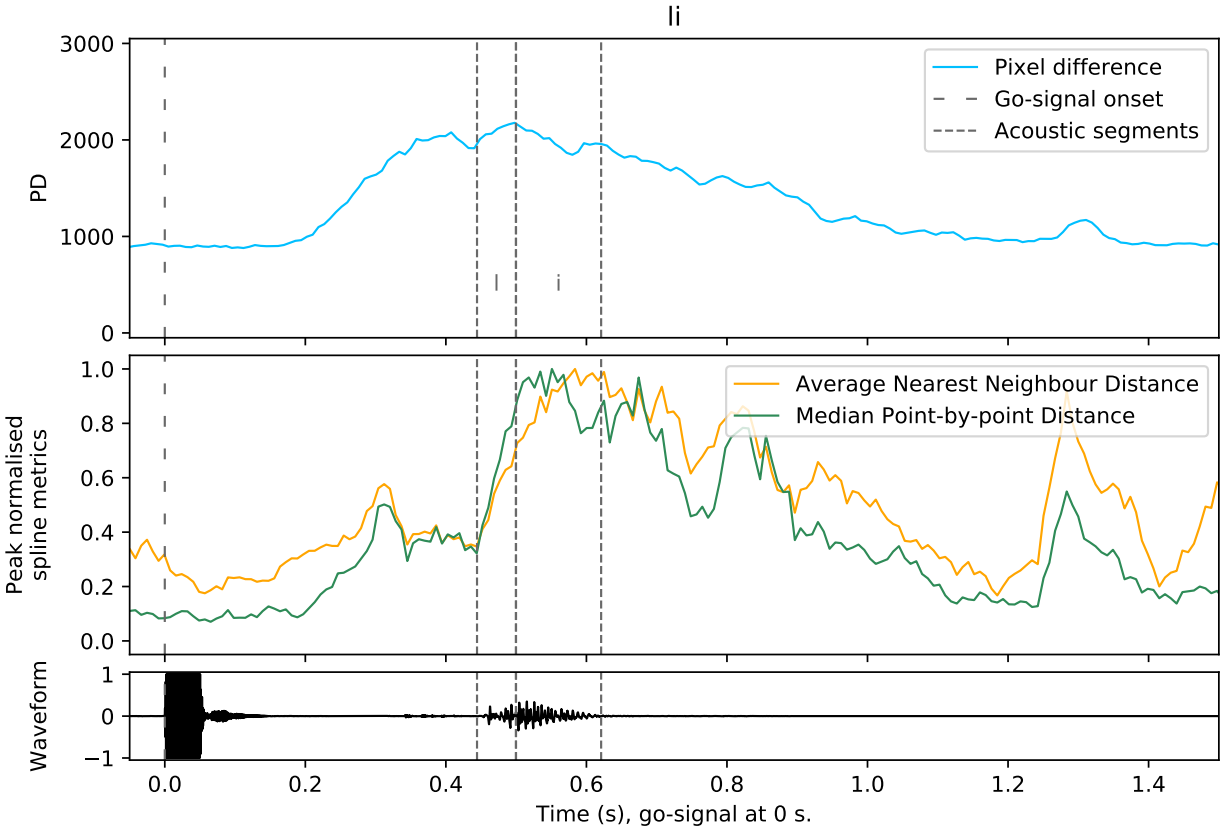

Figure 3: *PD in the top panel, peak normalised ANND and Median Point-by-Point Distance (MPBPD) in the middle panel, and acoustic waveform in the bottom panel. In all panels, the widely spaced dashed line marks the 'go'-signal and the tightly spaced dashed lines mark acoustic segment boundaries.*

of noise while the corresponding PD curve is very steady. Figure 3 shows an example where ANND (orange) shows significant activation before PD shows movement. After trying out median NND and average point-by-point distance, Median Point-by-Point Distance (MPBPD) was selected for use in this study because it was the most conservative of the tried metrics.

MPBPD calculates first the Euclidean distance between each corresponding sample point of the two splines being compared. (Correspondence is defined radially in probe centred coordinates.) The metric is then defined as the median of the individual distances.

3.3. Code availability

All analysis code was written in Python 3 and will be available as open source code under the GPL license in the near future. PD is already available in Python as part of the Speech Articulation ToolKIT (SATKIT) (Faytak, Moisik, and Palo 2020; Palo, P. and Moisik, S. R. and Faytak, M. 2020), which also includes other ultrasound analysis tools. ANND, MPBPD, and all other metrics that were tested in this study will be included in SATKIT in early 2021. The code also includes a simple GUI for annotating onsets on the different metrics.

4. Results

Figure 4 shows violin plots (box plot-like density distribution plots) of the difference between PD onset and MPBPD onset as well as the onsets themselves. For the difference between onsets the mean was $\approx$ 7.8 ms and sd $\approx$ 21.1 ms.

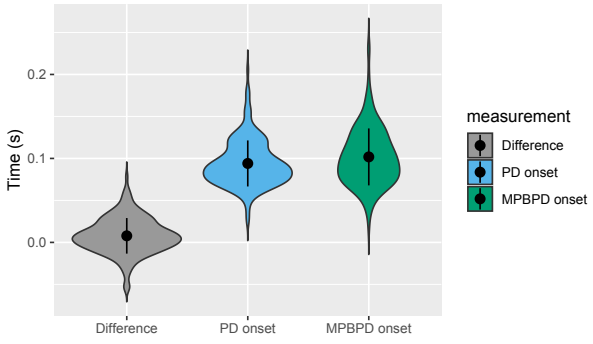

Figure 4: *Distributions of the difference between PD onsets and MPBPD onsets and the distributions of the onsets themselves. The black dots on the graphs represent means with the lines showing the extent of* $\pm$ **one** *standard deviation.*

The original sample consisted of 192 tokens. Out of these 179 were analysed. Three tokens were excluded due to technical problems with the data, nine tokens were excluded because they had either unclear patterns in either PD or MPBPD or both, and finally, one token was omitted for a considerably longer onset time than the others (close to .6 seconds in both modalities). As we can see the onset distributions are quite similar and most importantly the distribution of their differences overlaps 0 s as does the very conservative estimate of mean ± one standard deviation.

4.1. Observations on splining

It is very difficult to have a splining template that would catch everything. As a result it is practically always necessary to check all the splines before applying metrics to them *if we are interested in anything but the speech onset*. Spline checking tended to be easy and fast within the utterance because the template was fitted to a central articulation position. However, high starting positions needed to be corrected more often, and the steady articulation following the utterance had frequent tracking problems around the velar region. These are presumably individual characteristics of the speaker in the first data set. Correcting splines was done at a rate of slightly more than 8 frames per minute. Given the long samples (on average 1.5 s) analysed in the first data set, the correction time for as single token was 15-25 minutes. Since the aim was only to identify the speech onset, using automatic splining with minimal checking – only seeing if the splines fit well at the rest position – was needed.

4.2. Annotating PD and MPBPD

A custom Python GUI was used by the author to annotate the PD and MPBPD curves. Marking movement onset on both PD and MPBPD (or excluding a token from analysis) took about 35 minutes for a sample of 189 tokens. In most cases the decision was easy to make, while some MPBPD curves were more challenging. A total of 9 tokens had to be excluded because either PD or MPBPD or both were unclear.

5. Discussion and conclusion

It looks unlikely that we can detect tongue internal changes significantly before tongue contour movement – at least with the methods presented here. It is postulated that the muscular hydrostat nature of the tongue causes local changes in muscle shape to have an almost immediate effect on the overall shape of the tongue. Thus, it also seems unlikely that the difference between PD onset and movement onset measured manually from ultrasound videos could be due to human annotators reacting to tongue contour movement. Rather, we need to look for a different explanation there.

The results also suggest that at least the spline change metrics tested here are poorly suited for evaluating changes over short time intervals because spurious changes in the contour fitting produce relatively large changes in the metrics. This will most likely not be the case if the splines come from very different articulatory positions.

6. Acknowledgements

I wish to thank Steve Cowen for assistance with the ultrasound recordings and Professor Alan Wrench for advice and help on extracting the raw ultrasound data from AAA and subsequent post-processing of the data.

7. References

Articulate Assistant Advanced User Guide: Version 2.14 (2012). Edinburgh, UK: Articulate Instruments Ltd.

Davidson, L. (2006). "Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance". In: *Journal of the Acoustical Society of America* 120.1, pp. 407–415.

Drake, E., S. Schaeffler, and M. Corley (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., S. R. Moisik, and P. Palo (2020). "The Speech Articulation Toolkit (SATKit): Ultrasound image analysis in Python". To appear in ISSP 2020.

Kawamoto, Alan H., Qiang Liu, Keith Mura, and Adrianna Sanchez (2008). "Articulatory preparation in the delayed naming task". In: *Journal of Memory and Language* 58.2, pp. 347–365.

Linden, Lotje van der, Stephanie Kathleen Ries, Thierry Legou, Boris Burle, Nicole Malfait, and F.-Xavier Alario (2014). "A comparison of two procedures for verbal response time fractionation". In: *Frontiers in Psychology* 5.1213, pp. 1–11.

McMillan, C. T. and M. Corley (2010). "Cascading influences on the production of speech: Evidence from articulation". In: *Cognition* 117.3, pp. 243–260.

Mielke, J. (2015). "An ultrasound study of Canadian French rhotic vowels with polar smoothing spline comparisons". In: *Journal of the Acoustical Society of America* 137.5, pp. 2858–2869.

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`.

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

Saito, M., F. Tomaschek, C.-C. Sun, and R. H. Baayen (2020). "An ultrasound study of frequency and coarticulation". To appear in ISSP 2020.

Stone, M. (2005). "A Guide to Analyzing Tongue Motion from Ultrasound Images". In: *Clinical Linguistics and Phonetics* 19.6–7, pp. 455–502.

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