TOWARDS IDENTIFYING ARTICULATORY GESTURES WITH PIXEL DIFFERENCE AND AUDIO SEGMENTATION

Giulia S. Orlando[1] and Pertti Palo[2]

[1] University of Helsinki, [2] Indiana University
[1] giulia.orlando@helsinki.fi, [2] pertti.palo@taurlin.org

ABSTRACT

We report results from using Pixel Difference – a metric based on Euclidean distance – to analyze articulatory gestures. Our data consists of mono- and disyllabic English words recorded with tongue ultrasound from four speakers. We analyze the shapes of Pixel Difference curves based on acoustic segmentation. The motivation is to see when and how articulatory and acoustic events align. Further, in cases where acoustic segmentation is difficult, e.g. approximants, articulatory events may provide potentially more robust anchor points.

The current study aims to provide a basis to analyze larger datasets. Since this use of Pixel Difference is a novel way to look at speech data, we present a qualitative analysis of time behaviour of Pixel Difference. This will provide a basis for quantitative analysis in the next phase of our project. As the first result, we describe how Pixel Difference curves in VCV sequences depend on the intervening consonant.

Keywords: Acoustic-articulatory analysis, tongue ultrasound, speech gestures, Euclidean distance.

1. INTRODUCTION

The present study investigates the relation between tongue movement and acoustic production. Seeking to find ways to identify speech gestures in ultrasound data in a way similar to how they are found in point tracking data [1, 2], we investigate the behaviour of Pixel Difference (PD) in relation to acoustic segmentation. PD and related metrics have been used for analysing speech data in various ways but not for gesture identification [3, 4, 5].

The analysis starts from the observation of the PD plot, with the aim of mapping the observed shapes and recognizing a set of behavior patterns. On the basis of such patterns, the paper outlines a threefold categorization. The main scope is to observe the degree of reliability of these behaviors and the consistency of the relation between PD shape and sound. Rooted herein, in understanding whether such presumed relation between movement and sound is consistent, is the intention to possibly consider using PD as an aiding tool for acoustic analysis. In closing, the paper also discusses the seeming deviations from the observed patterns, and lays out a structure for a possible generalisation of the relation discussed here.

2. MATERIALS AND METHODS

The data is from a separate study of speech initiation which uses a delayed naming task. The participants were instructed to remain at rest until they heard the go signal – a 1 kHz pure tone – and then produce the target word as soon and as accurately as possible. However, we concentrate on analysing utterance medial phenomena and are not concerned with the original study's questions. For the purpose of this study, the data is just words read out in isolation.

The data was recorded with tongue ultrasound controlled with Articulate Assistant Advanced [6]. Ultrasound was captured at 80 fps, and FOV was 92 degrees with synchronized audio sampled at 44.1 kHz.

The dataset consists of English lexical monosyllabic /[C]V[C]/[1] and disyllabic /[C]VCV[C]/ words produced by L1 speakers of General American English – 1 female, 1 non-binary, 2 males. The words vary onset, medial and coda consonants, as well as vowels, with a total of 180 words in the set. We analyse a subset of these words limited to words where the first and second vowels are phonologically identical. More specifically, we concentrate on the /VCV/ sequence in the disyllabic words and use only /V/ words from the monosyllabic set as an example of a vowel produced in isolation. In the subset we analyze, the medial consonant was one of /b, d, f, g, h, k, l, ɹ, ɾ, m, n, p, r, s, ʃ, t, θ, w, z/ and the vowels were /a, i, u/.

2.1. Acoustic annotation

After generating and populating tiers for utterance, word, phonological segment and phonetic

detail automatically with CAST[2], the data was acoustically segmented in Praat [7] by one of the authors.

2.2. Articulatory analysis

The articulatory analysis is based on a Euclidean metric called Pixel Difference (PD) [5]. PD is defined as the Euclidean distance between consecutive raw (uninterpolated, probe return) ultrasound frames. Expressed as an equation for frames k and $k+1$:

$$(1) \quad PD(k) = \sqrt{\sum_{i=1}^{n_x} \sum_{j=1}^{n_y} (im_k(i,j) - im_{k+1}(i,j))^2}$$

where indices i and j iterate over the pixels in x and y direction, $im_k(i,j)$ denotes the pixel in frame k, at row i and column j, and $k = \{1,2,\ldots n_{frames} - 1\}$.

The articulatory analysis was performed with a Python software package called SATKIT [8, 9]. Figure 1 shows an example of data display in SATKIT.

Given the exploratory nature of this study and the limited dataset size, we chose a qualitative approach to analysing the data. Each author went through the data independently and listed their observations. After this, we discussed our findings to identify systematicities that we could agree on.

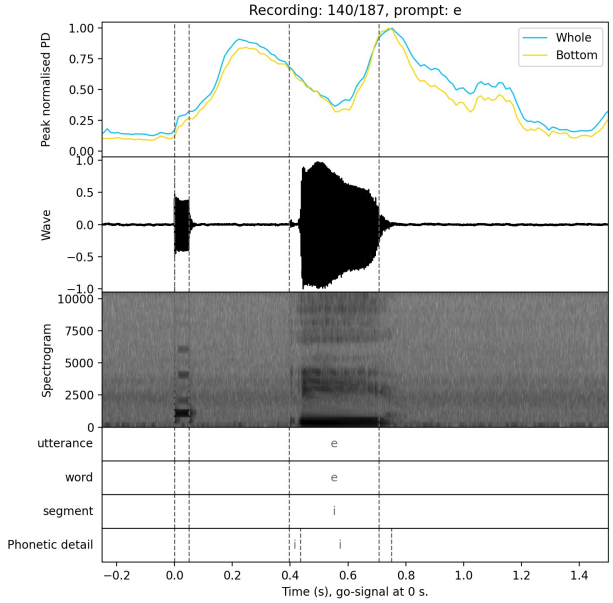


Figure 1: Participant 2 reading 'e', realised as [i]. Pixel Difference (PD) on top shows the typical trough shape of a vowel surrounded by an onset and an offset gesture.

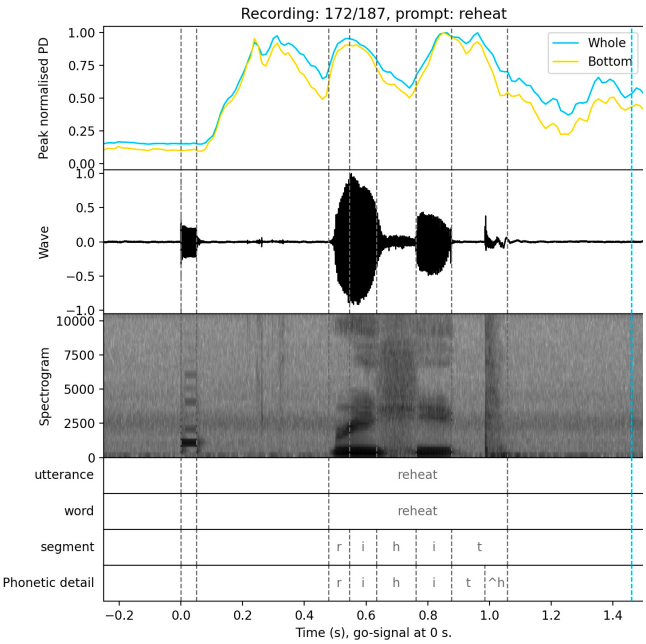
3. RESULTS

The observations of the PD plot seem to show a consistent relation between tongue movement and sound, where the tongue movement appears to follow certain patterns, specific to certain sounds. Specifically, we have been able to identify three categories in PD. Since the present study focuses on /VCV/, the resulting categories are built on the observed deviation in the PD trajectory of what has been identified as the most basic pattern, namely the tongue movement associated with a vowel (Figure 1). The consonants have been grouped to that do not particularly affect the base trajectory (Group 1, Figure 2a), those that generate a gestural peak (Group 2, Figures 2b and 2c), and those whose gestural peak bifurcates in two (Group 3, Figure 2d). Within the categorization proposed, there is an element apparently inconsistently represented in PD, or rather showing a high degree of variability: the allophones represented in American English by the phoneme /h/.

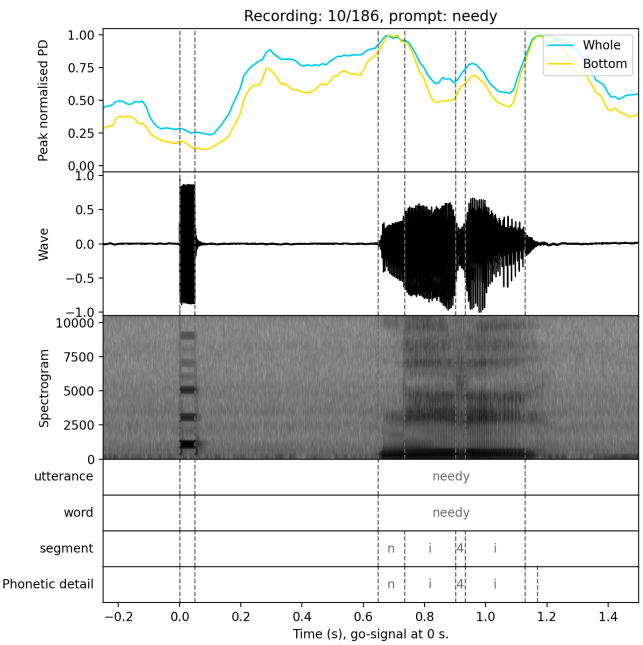
GROUP 0 – vowels A vowel in isolation in PD is shaped as a trough. Initial or final aspiration added to an isolated vowel does not modify the curve observed in PD.

GROUP 1 – glottal fricatives, bilabial nasals They do not affect the vocalic trough or create just a minor disturbance in it, because by their nature they do not require tongue movement, which if happens it is of minor entity. Figure 2 (a) is a good example of how /h/ does not affect the PD trajectory. The PD curve forms a long v-shaped trough over the acoustic /ihi/ sequence with the PD peaks at each end of the sequence aligning very well with the acoustic segment boundaries (beginning of first /i/ and end of second /i/). Similar examples have been found in the dataset for the bilabials nasal, where /m/ does not affect at all the PD trajectory. And even in the instances when /h/ and /m/ do affect the trajectory, it is just a very subtle fluctuation, consisting of a minor gestural peak in the vocalic trough.

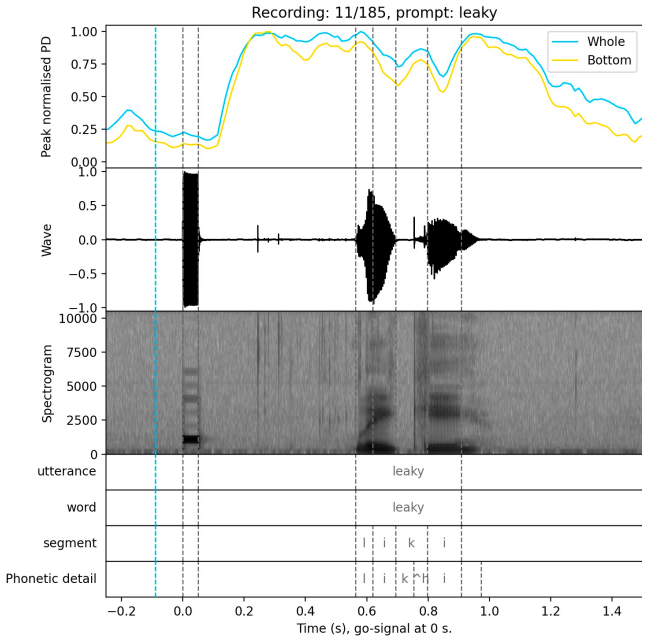
GROUP 2 – plosives Plosives create a gestural peak in the vocalic trough. The amplitude of the peak appears to depend on the point of articulation, with frontal plosives generating smaller gesture peaks, and velars generating bigger ones. The categorization proposed in this paper finds good reasons in grouping all the plosives together, since regardless of their type, all of them produce the same single gestural peak in PD. A good example of the



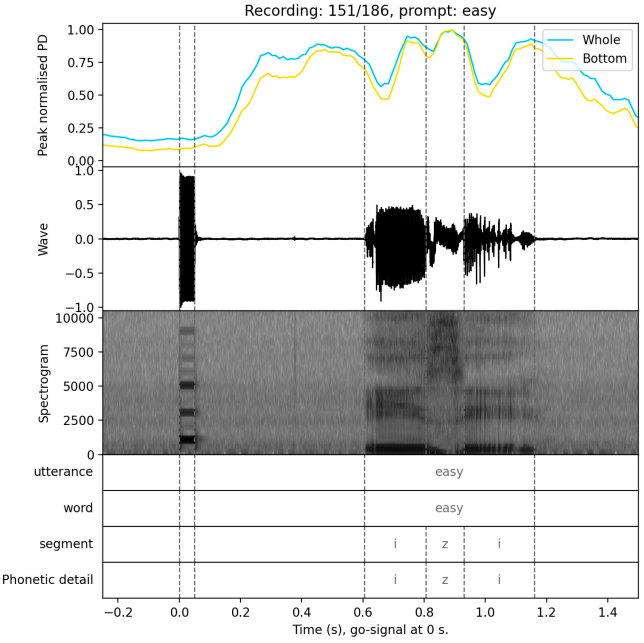
(a) Example of Group 1: Participant 2 reads 'reheat'.



(b) Example 1 of Group 2: Participant 3 reads 'needy'.



(c) Example 2 of Group 2: Participant 4 reads 'leaky'.



(d) Example of Group 3: Participant 3 reads 'easy'.

Figure 2: Examples of Groups 1-4.

plosive peak can be observed in Figures 2b and 2c. PD shows a falling line starting at the onset of the first vowel, and a rising one ending at the offset of the second vowel, with a gestural peak starting at the onset of the velar /k/.

The need to contain a build-up of pressure leading to the plosive release also seems to affect the amplitude of the peak. There were examples in the dataset, where between the falling movement of the first vowel, and the rising one of the second vowel, there is a big gestural peak, when the bilabial is produced. It was also possible to find counter examples of the same plosive without buildup pressure, where the bilabial appears just as a small peak in the vocalic trough.

GROUP 3 – laterals, sibilants, approximants
This group of consonants creates a significant

fluctuation in the vocalic trough, namely a bifurcated peak within its trajectory. Figure 2d illustrates this behavior. Here the intervocalic phoneme considered is /z/. PD shows a falling line starting with the onset of the first /i/, then a rising one just when the spectrogram starts to change, then a small drop corresponding to the production of the sibilant, a falling movement ending when the spectrogram changes again, and then another rising movement ending with the offset of the second /i/. Considering /izi/ altogether, this fluctuation can be seen as a double bump placed halfway in the vocalic trough.

Summary of sounds by Group:

Group 0 vowels

Group 1 bilabial nasal /m/

Group 2 mainly plosives /b, d, g, k, ɾ, n, p, t, θ/

Group 3 /l, ɹ, s, ʃ, t, θ, w, z/, possibly also f, r

In more than one group /h/

4. CONCLUSION

The qualitative inspection of data suggests a consistent relation between movement and sound. At this stage, we observed, in intervocalic position, consistency in the behavior of the described categories, which is coherent with the tongue movement involved in their production. Based on the analysis presented in this paper, we believe PD will prove helpful in supporting acoustic segmentation. This seems especially helpful when spectral analysis results a bit unclear (i.e. with laterals, rhotics, approximants). In order to suggest an application of PD analysis for supporting acoustic segmentation, however, further investigation to map the tongue movement patterns in other places, and a larger dataset to generalise our observations are required. In particular, a larger data set is required to be able to quantify the role of peak height in plosive articulations. With the possibility to precisely measure the amplitude, we would be able also to propose a more fine-grained description of the plosives, and thus subcategorise them.

5. DISCUSSION

As observed in this dataset, the behavior of aspiration in PD varies greatly. This seems to find an explanation in the fact that aspiration can be produced with different phonemes, i.e. a glottal fricative (Figure 2a) or pharyngeal/velar which are allophones in American English. In the dataset, it can be seen that glottal fricatives do not affect the vocalic trough due to the nature of their gesture, namely because there is no tongue movement involved. However, when the pronunciation of this intervocalic /h/ is stronger, the difference shows in PD, potentially because the sound produced is actually a different phone. The produced fricative is not a glottal, but a pharyngeal/velar fricative, which involves the tongue in its production, thus unavoidably affects the PD plot. This would need further investigation to account for the specific variation in the tongue movement, but on the basis of the current data, the behavior of aspiration appears to be coherent with the categorization given and does not disprove the patterns outlined.

Another category which showed great variability was the bilabials: in our dataset, we observed that bilabial nasals have slight to no impact on the vocalic trough; as for bilabial plosives, their impact on the PD trajectory was variable, and potentially dependent on the intraoral pressure. In some instances, there was a peak in the PD trajectory in correspondence to the consonant, while in others, the consonant (/b/, /p/) did not perturb the vocalic trough in PD. These findings are related to the electromyography (EMG) trough effect, which is a deactivation of underlying articulator activity during the production of /VCV/ (identical vowels, bilabial consonant) [10, 11]. However, unlike EMG, PD analysis of tongue ultrasound does not differentiate between muscles. This means that it provides a different angle on the phenomenon, and the observed trajectories are not directly comparable.

But since PD analysis of tongue ultrasound does not differentiate between muscles, the results differ to some degree. It would be valuable to collect data with EMG and tongue ultrasound to observe how they relate and obtain a deeper perspective on the muscular activity involved.

From the present analysis, it seems reasonable to assume that the behavior observed in PD plot is correlated with the steadiness of the sound produced. Vowels have a steadier mid-phase, which shows in PD as the trough; while plosives have an acoustical steady state during the closure, in PD they show as just a peak; finally, the rest of the consonants have a steadier phase similar to the vowel's but shorter reflected in PD as a trough which bifurcates the gesture peak in two. This analysis suggests that PD can be used to select the time points when a phoneme reaches its target. For vowels and most consonants this would be the bottom of the PD trough and for plosives the peak of their gesture.

6. REFERENCES

[1] C. P. Browman and L. Goldstein, "Some notes on syllable structure in articulatory phonology," *Phonetica*, vol. 45, pp. 140 – 155, 1988.

[2] ——, "Articulatory gestures as phonological units." *Phonology*, vol. 6, pp. 201 – 251, 1990.

[3] C. T. McMillan and M. Corley, "Cascading influences on the production of speech: Evidence from articulation," *Cognition*, vol. 117, no. 3, pp. 243 – 260, 2010.

[4] E. Drake, S. Schaeffler, and M. Corley, "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, 2013.

[5] P. Palo, "Measuring pre-speech articulation," Ph.D. dissertation, Queen Margaret University, Edinburgh, 2019.

[6] *Articulate Assistant Advanced User Guide: Version 2.14*, Edinburgh, UK: Articulate Instruments Ltd, 2012.

[7] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," 2022, version 6.2.12, retrieved 14 April 2022 from http://www.praat.org/.

[8] P. Palo, S. R. Moisik, and M. Faytak, "SATKIT: Speech Articulation ToolKIT [Python software package]," Available in a public software repository, accessed 31 Aug 2022, https://github.com/giuthas/satkit.

[9] ——, "Analysing speech data with SATKIT," in *International Conference of Phonetic Sciences (ICPhS 2023)*, Prague, 2023.

[10] B. Lindblom, H. Sussman, G. Modarresi, and E. Burlingame, "The trough effect: implications for speech motor programming." *Phonetica*, vol. 59, no. 4, pp. 245 – 262, 2002.

[11] S. Fuchs, P. Hoole, J. Brunner, and M. Inoue, "The trough effect – an aerodynamic phenomenon?" in *From sound to sense: MIT Meeting*, 2004.

[1] Here [C] denotes a consonant being present in some but not all of the words in the indicated position.

[2] https://github.com/giuthas/cast