Pre-Speech Tongue Movements Recorded with Ultrasound

Pertti Palo[1], Sonja Schaeffler[1], James M. Scobbie[1]

[1]*CASL Research Centre, Queen Margaret Uiniversity, Edinburgh, United Kingdom*
ppaloqmu.ac.uk, sschaeflerqmu.ac.uk, jscobbieqmu.ac.uk

Abstract

We analyse Ultrasound Tongue Imaging (UTI) data from five speakers, whose native languages (L1) are English (3 speakers), German (1 speaker), and Finnish (1 speaker). The data consist of single words spoken in the subjects' respective native tongues as responses to a picture naming task. The focus of this study is on automating the analysis of ultrasound recordings of tongue movements that take place after the subject is presented with a stimulus. We analyse these movements with a pixel difference method (McMillan and Corley 2010; Drake, Schaeffler, and Corley 2013a; Drake, Schaeffler, and Corley 2013b), which yields an estimate on the rate of change on a frame by frame basis. We describe typical time dependent pixel difference contours and report grand average contours for each of the speakers.

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

1. Introduction

Study of speech preparation has come under increasing interest in recent years. Studies based on the acoustic modality analysing e.g. questions of phonological preparation (Rastle et al. 2000), and conversation turn taking (Heldner and Edlund 2010) have been complemented with studies based on articulatory data such as the recent analysis of organisation of speech preparation processes by Tilsen and Goldstein 2012.

The most readily accessible modality of speech production – and therefore of speech preparation – is sound. However, recording speech sounds and sounds produced during speech preparation is an indirect way of observing the physical processes that produce them. Furthermore, the speech preparation movements are mostly silent and thus best observed by directly recording the articulation itself.

By varying the complexity of the tasks performed by the participants of a reaction time experiment light can be shed on the organisation of cognitive speech preparation processes (see e.g. Rastle et al. 2000; Tilsen and Goldstein 2012). For this purpose researchers are usually concerned with the onset time of acoustic speech, onset of phonation, or the onset of phonetically meaningful articulatory movements.

Ultrasound Tongue Imaging (UTI) has a long history in speech studies in general see Minifie, Kelsey, and Zagzebski 1971, for one of the first studies. Its applications range from speech therapy see e.g. Bernhardt et al. 2005 to silent speech interfaces Hueber et al. 2010. In comparison with the most viable alternative, i.e. Magnetic Resonance Imaging (MRI), UTI is relatively cheaper and simpler to use and has a far better temporal resolution. On the down side UTI data is fairly noisy, often contains artefacts and the imaged area is limited to the tongue and its surface.

McMillan and Corley 2010; Drake, Schaeffler, and Corley 2013a; Drake, Schaeffler, and Corley 2013b have automated the processing of UTI data by considering the difference or amount of change between consequtive ultrasound frames. This difference has been defined as the Eucledian distance of two ultrasound frames or images as they are taken to be N dimensional vectors, with each pixel presenting a dimension.

2. Materials and methods

The experiment used the Snodgrass-Vanderwart picture naming task (Snodgrass and Vanderwart 1980) and it had five subjects: Four females (P1, P2, P3, and G1) and one male (S1). All of the speakers did the experiment in their native tongue – three in English (participants P1, P2, and P3), one in German (participant G1) and one in Finnish (participant S1).

The experiment was run with synchronised ultrasound, lip imaging and sound recording controlled with Articulate Assistant Advanced (AAA) software (*Articulate Assistant Advanced User Guide: Version 2.14* 2012) which will also be used for the analysis. The participants were fitted with a purpose-built headset to ensure stabilisation of the ultrasound probe (*Ultrasound Stabilisation Headset Users Manual: Revision 1.4* 2008). Attached to the helmet was a small Audio Technica AT803b microphone for high-quality acoustic recordings. Ultrasound recordings were obtained at a frame rate of 201 frames per second.

The recording was initiated 1.5 seconds before the subjects were shown the stimulus on a computer screen, thus capturing any movements related to speech preparation as well as making it possible to spot cases where the subject was moving already before the onset of the stimulus. These cases as well as ones where the subject trouble naming the picture were excluded from further analysis resulting still in well over 200 analysed tokens per speaker.

2.1. Pixel differences

We analyse the samples by automatically calculating the amount of change in the UTI data as a function of time. We use pixel difference (McMillan and Corley 2010; Drake, Schaeffler, and Corley 2013a; Drake, Schaeffler, and Corley 2013b) as the change metric and evaluate it over each recording comparing raw ultrasound frames in sequence.

UTI commonly uses ultrasound probes which produce a fan shaepd image of the tongue. Ordinary or interpolated ultrasound data refers to the form commonly displayed by ultrasound imaging systems as seen in Fig. 1 a). The fan image of the ordinary ultrasound data is produced by linear interpolation between the actual raw datapoints produced by the ultrasound system as it images the tissues. The raw datapoints are distributed along radial scanlines with the number of scanlines and the number of data points imaged along each scanline depend-

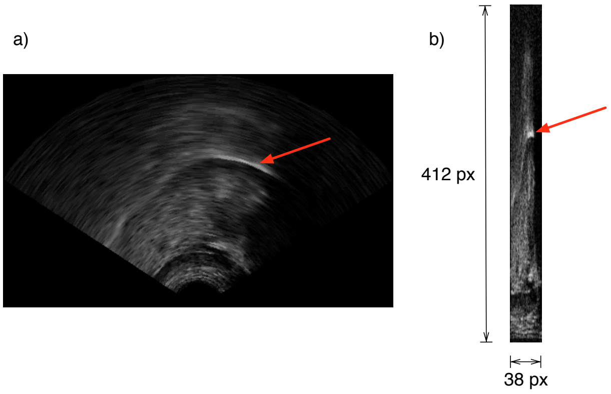

Figure 1: *a) 1st ultrasound frame from participant P2 naming a picture of a lemon. b) Raw (uninterpolated) version of the same ultrasound frame as in a). The participant is facing right. Red arrow points to the upper surface of the tip of the tongue.*

ing on the setup of the ultrasound system.

Fig. 1 b) shows a raw ultrasound frame. In our experiments each frame has 38 scanlines (x dimension of the image) and 412 datapoints of pixels along each scanline (y dimension of the image).

To calculate the pixel difference between two UTI frames we interpret each raw frame as a $N = n_x \times n_y$ dimensional vector. The pixel difference $d1$ between consecutive UTI frames is then defined as the Eucledian distance between the two frames im_k and im_{k+1} with indeces i and j iterating over the pixels in x and y direction:

$$d1(k) = \sqrt{\sum_{i=1}^{n_x} \sum_{j=1}^{n_y} (im_k(i,j) - im_{k+1}(i,j))^2} \qquad (1)$$

for $k = \{1, 2, \ldots n_{frames} - 1\}$. The difference can be calculated as readily for images further removed and are defined as

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

for $k = \{1, 2, \ldots n_{frames} - L\}$. In this paper we will use the differences d1 and d3. The time stamp $t_{dL(k)}$ corresponding to a single difference value $dL(k)$ is defined as the average of the time stamps of the corresponding UTI frames:

$$t_{dL(k)} = \frac{1}{2}(t_{im_k} + t_{im_{k+L}}), \qquad (3)$$

where the time stamp of image k is the time its acquisition ends.

3. Results

[Explain the pictures here:]

Fig. 2 shows a typical clear production. The participant has held still before starting linguistic articulation. From comparison with the corresponding waveform at the bottom of the figure, it can be seen that the articulation starts a good while before acoustic onset.

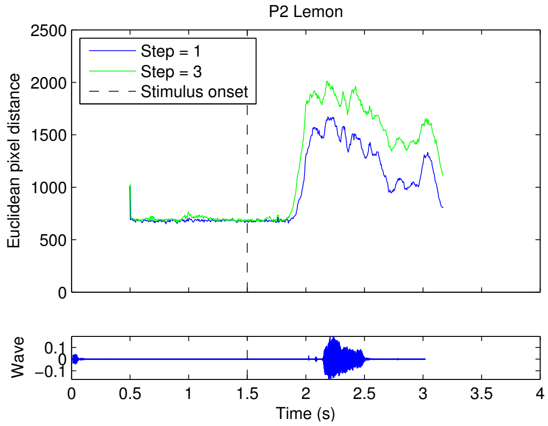

Figure 2: *A clear sample: P2 naming a picture of a lemon.*

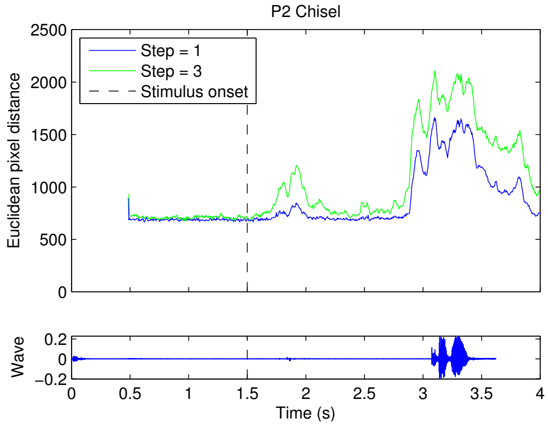

Figure 3: *A hesitation: P2 naming a picture of a chisel.*

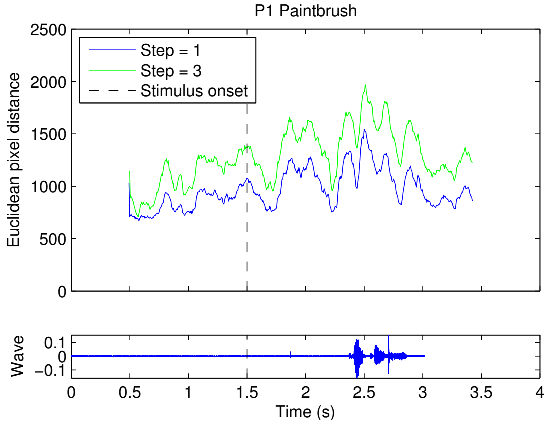

Figure 4: *Chaos: P1 naming a picture of a paintbrush.*

Fig. 3 shows an example of a hesitation. The participant moves their tongue as if they were going to speak, but returns to rest and finally speaks later.

Fig. 3 shows a chaotic example. The participant is moving already at the time the recording starts and continues to move through out the whole recording.

Then explain the grand averages and thus figures 5, 6, 7, 8, and 9.

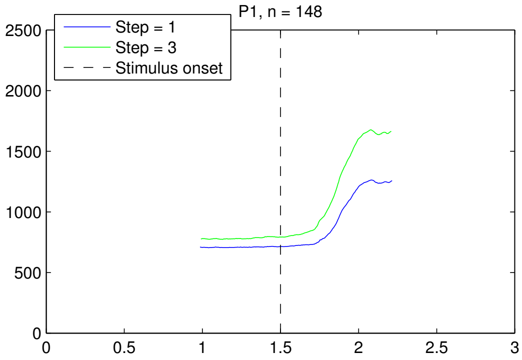

Figure 5: *Grand averages of all participant P1's UTI recordings longer than 2 seconds.*

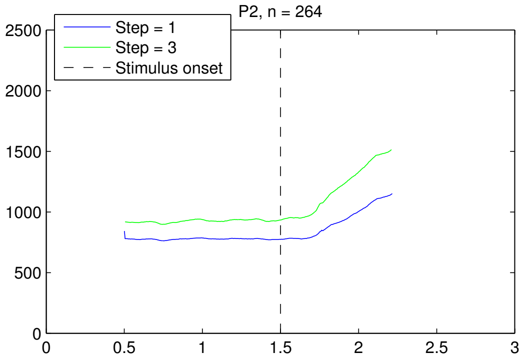

Figure 6: *Grand averages of all participant P2's UTI recordings longer than 2 seconds.*

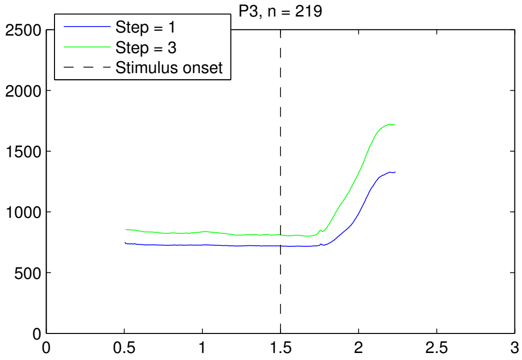

Figure 7: *Grand averages of all participant P3's UTI recordings longer than 2 seconds.*

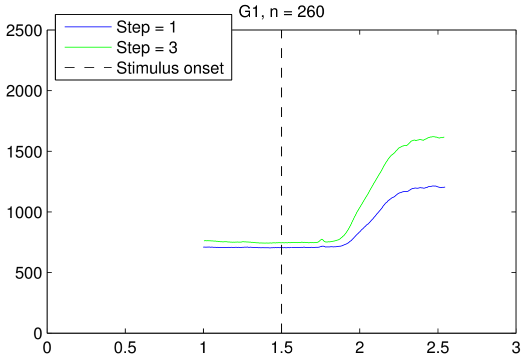

Figure 8: *Grand averages of all participant G1's UTI recordings longer than 2 seconds.*

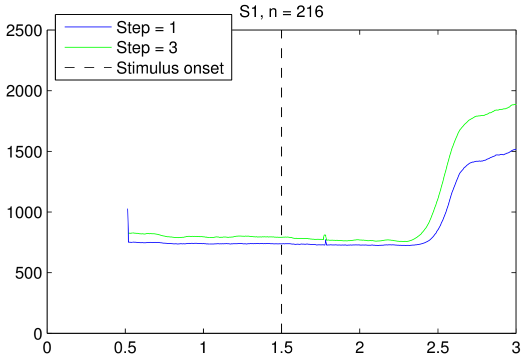

Figure 9: *Grand averages of all participant S1's UTI recordings longer than 2 seconds.*

4. Discussion and conclusion

Points:

- Different types of productions are present in the data. Some are easier to analyse than others.
- Individual differences are evident in the averages. For example, S1 is slower to respond than the other participants and P2 has more variation (i.e. probably more movement) in the time leading up to linguistic articulation. Particularly here the use of $d1$ in conjunction with $d3$ or some other higher order pixel difference is essential.
- The use of $d3$ probably needs more explanation. Could be done by pointing to how it enhances the visibility of hesitation in the chisel sample.
- The timeline hypothesised in the original abstract, does not correspond with the majority of the data.
- A word or two about noise floor would be good too.

5. Acknowledgements

The authors 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.

6. References

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

Ultrasound Stabilisation Headset Users Manual: Revision 1.4 (2008). Edinburgh, UK: Articulate Instruments Ltd.

Bernhardt, B., B. Gick, P. Bacsfalvi, and M. Adler-Bock (2005). "Ultrasound in speech therapy with adolescents and adults". In: *Clinical Linguistics & Phonetics* 19.6-7, pp. 605–617.

Drake, E., S. Schaeffler, and M. Corley (2013a). “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.

— (2013b). “Does prediction in comprehension involve articulation? Evidence from speech imaging”. In: *11th Symposium of Psycholinguistics (SCOPE)*. Tenerife.

Heldner, M. and J. Edlund (2010). “Pauses, gaps and overlaps in conversations”. In: *Journal of Phonetics* 38.4, pp. 555–568. DOI: `http://dx.doi.org/10.1016/j.wocn.2010.08.002`. URL: `http://www.sciencedirect.com/science/article/pii/S0095447010000628`.

Hueber, Thomas, Elie-Laurent Benaroya, Gérard Chollet, Bruce Denby, Gérard Dreyfus, and Maureen Stone (2010). “Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips”. In: *Speech Communication* 52.4, pp. 288–300. DOI: `http://dx.doi.org/10.1016/j.specom.2009.11.004`. URL: `http://www.sciencedirect.com/science/article/pii/S0167639309001733`.

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

Minifie, F., C. Kelsey, and J. Zagzebski (1971). “Ultrasonic Scans of the Dorsal Surface of the Tongue”. In: *Journal of the Acoustical Society of America* 49.6, pp. 1857–1860.

Rastle, K., J. Harrington, M. Coltheart, and S. Palethorpe (2000). “Reading Aloud Begins When the Computation of Phonology Is Complete”. In: *Journal of Experimental Psychology: Human Perception and Performance* 26.3, pp. 1178–1191.

Snodgrass, J. G. and M. Vanderwart (1980). “A Standardized Set of 260 Pictures: Norms for Name Agreement, Image Agreement, Familiarity, and Visual Complexity”. In: *Journal of Experimental Psychology: Human Learning and Memory* 6.2, pp. 174–215.

Tilsen, S. and L. Goldstein (2012). “Articulatory gestures are individually selected in production”. In: *Journal of Phonetics* 40, pp. 764–779.

7. Stuff from the template. Will remove these on Monday

LaTeX users: the document should compile successfully after

- an initial call to `pdflatex *.tex`
- followed by a call to `biber *.bcf`
- followed by two more calls to `pdflatex *.tex`

8. Page layout and style

All papers must be submitted in compliance with the provided template. Check details of your final PDF submission against the template example file.

8.1. Basic layout features

- Two columns are used except for the title section.
- Left margin is 20 mm.
- Column width is 80 mm.
- Spacing between columns is 10 mm.
- Top margin 25 mm.
- Text height (without headers and footers) is maximum 235 mm.
- Headers and footers must be left empty.

8.1.1. Headings

Section headings are in boldface with the first word capitalized and the rest of the heading in lower case. Sub-headings appear like major headings. Sub-sub-headings appear like sub-headings, except they are in italics and not boldface. No more than 3 levels of headings should be used.

8.2. Text font

Times or Times Roman font is used for the main text. Font size in the main text must be 9 points, and in the References section 8 points. Other font types may be used if needed for special purposes.
IMPORTANT: All fonts must be embedded when compiling the final PDF! Failure to embed fonts will result in missing text in the final publication.

LaTeX users: font types, font face and type size should be predefined in either the document body or the issp2014.sty style file. Authors must not use Type 3 (bitmap) fonts.

8.3. Figures and Tables

Figures should preferably be line drawings. If they contain gray levels or colors, they should be checked to print well on a high-quality non-color laser printer. Please use only *solid* fill colors.

Table 1: *This is an example of a table.*

Ratio	Decibels
1/1	0
2/1	≈ 6
3.16	10
10/1	20
1/10	-20
100/1	40
1000/1	60

8.4. Equations

Equations should be placed on separate lines and numbered. Examples of equations are given below. Particularly,

$$x(t) = s(f_\omega(t)) \tag{4}$$

where $f_\omega(t)$ is a special warping function

$$f_\omega(t) = \frac{1}{2\pi j}\oint_C \frac{\nu^{-1k}d\nu}{(1-\beta\nu^{-1})(\nu^{-1}-\beta)} \tag{5}$$

A residue theorem states that

$$\oint_C F(z)dz = 2\pi j \sum_k Res[F(z), p_k] \tag{6}$$

Applying (6) to (4), it is straightforward to see that

$$1 + 1 = \pi \tag{7}$$

Finally we have proven the secret theorem of all speech sciences. No more math is needed to show how useful the result is!

8.5. Submitted files

Authors are requested to submit PDF files of their manuscripts. Tools are available online to create the PDF such as:

- `http://www.pdfforge.org/products/pdfcreator`
- pdflatex

The PDF file should comply with the following requirements: (a) there must be no PASSWORD protection on the PDF file at all; (b) all fonts must be embedded; and (c) the file must be text searchable (do CTRL-F and try to find a common word such as 'the'). The proceeding editors will contact authors of non-complying files to obtain a replacement. In order not to endanger the preparation of the proceedings, papers for which a replacement is not provided timely will be withdrawn.

8.6. Page numbering

Page numbers will be added later to the document electronically. *Don't make any footers or headers!*

8.7. Abstract

The total length of the abstract is limited to 1000 characters. The abstract included in your paper and the one you enter during web-based submission must be identical. Avoid non-ASCII characters or symbols as they may not display correctly in the abstract book.