COMPUTER ASSISTED SEGMENTATION OF TONGUE ULTRASOUND AND LIP VIDEOS

Pertti Palo *[1]
[1]Indiana University, Bloomington, Indiana

1 Introduction

In many speech analysis tasks, such as evaluating reaction times, combining direct measurement of articulation or muscular activation with audio data is preferable to using only audio data [1, 2]. However, compared to segmenting acoustic speech data, time domain analysis of tongue ultrasound data and lip videos is challenging and there is yet to develop a consensus on best tools for the task. The most widely used method is to select time points for articulatory analysis on the basis of audio segmentation. In contrast, for acoustic analysis the spectrogram provides an easy way of analysing time and frequency domain characteristics of the speech signal in one glance.

Among articulatory measurement methods tongue ultrasound is currently one of the most used. While analysing extracted tongue contours is perhaps the most popular method of analysing tongue ultrasound data, recently methods that analyse the whole ultrasound image have received attention [3–5].

One such method is an analysis tool called Pixel Difference (PD), which can be used for direct phonetic analysis of tongue ultrasound data [6, 7]. The tool is an application of the Euclidean distance metric to the whole ultrasound image. It can be used to easily visualise over all change in the data. This study extends PD for simultaneous analysis of synchronised tongue ultrasound and lip videos. Analysis results of a sample data set of synchronously recorded ultrasound and lip video from a single speaker will be discussed in the presentation.

2 Materials

The speech materials come from a delayed naming experiment, which were recorded with the high-speed ultrasound facility at Queen Margaret University. The data is described in more detail as Experiment 2 of the author's PhD thesis [6]. In it lexical /CVC/ words were produced by speakers of Standard Scottish English. The materials analysed here come from a young adult male speaker designated P1.

In the experiment the participants were asked 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. Ultrasound was captured at 120 fps and FOV was 137 degrees. And lip videos from a side view camera mounted on the ultrasound helmet at 29.97 fps and de-interlaced to 59.94 fps. Results and further details have been published in the thesis [6].

∗. pertti.palo@taurlin.org

3 Method : 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. top panel of Figure 1) PD provides a clear view of tongue gestures and is particularly useful in identifying articulatory utterance onset.

4 Results

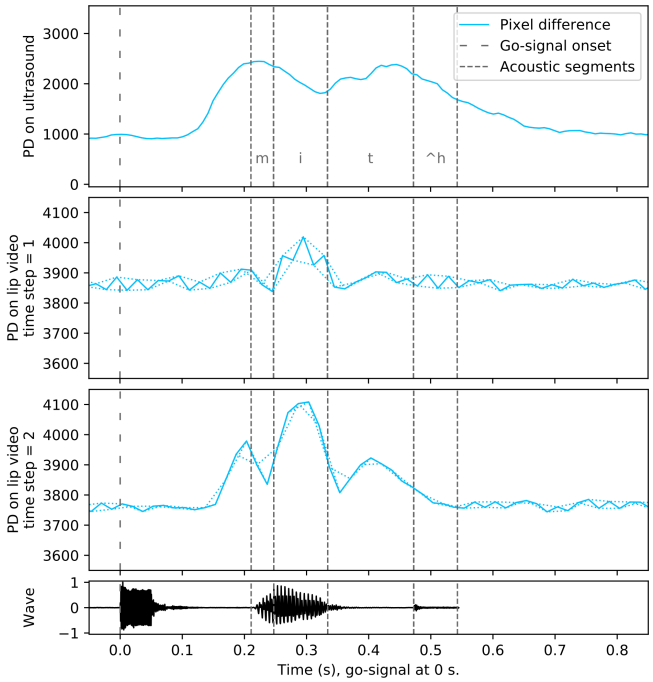

FIGURE 1 – Scottish English speaker pronouncing /meet/. PD on raw ultrasound (top panel) and on lip video data : second panel – time step = 1 ; third panel – time step = 2).

4.1 Selection of time step

One of the choices available in fine tuning PD to a given data source or set, is selection of the used time step. Given no other constraints it is preferable to compare consecutive data frames without skipping any [6]. This means using a time step of 1. In other words, that the individual pixel differences are calculated as $\Delta_i(t+.5) = pixel_i(t) - pixel_i(t+1)$ where i spans the pixels in a frame and $\mathrm{PD}(t+.5) = \sqrt{\sum_{i=1}^{n} \Delta_i^2}$. Using a longer step means we loose time locality of PD : $\delta_i(t+s/2) = pixel_i(t) - pixel_i(t+s)$, where $s > 1$. Yet qualities of the data may force us take this choice [6].

As we can see in the second panel of Figure 1, PD on lip video data has a clear sawtooth wave riding on it when we use a time step of 1. In some extreme cases any analysable change in the curve is masked by this sawtooth effect. The ripple alternates between consecutive frames. Envelope curves have been drawn on the image with dotted lines to aid visual analysis.

To find a solution to this problem, time steps between 1 and 5 were used on a small test data set. The results provided evidence for the effect being a consequence of odd vs. even frame comparison as it was smallest on even time steps (comparing even with even and odd with odd), but largest with odd time steps. Since using a time step larger than 2 did not bring much additional clarity to the curves, this step was chosen as recommended one.

Third panel from top in Figure 1 shows the same data as the previous panel, but now analysed with time step of 2. As we can see, almost all of the sawtooth is gone and the changes in the curve are much clearer to the eye. Some sawtooth noise does still remain.

4.2 Other observations

As we can see by looking at the y-scale of the middle panels of Figure 1, PD on lip videos required zooming to make the changes visible. Fortunately, the PD signal level is constant within a recording session with a relatively stable noise floor, which in this example session was about 3750 PD units. Due to changes in image content the noise floor and signal levels change from session to session making it necessary to adjust the y-scale each time data is processed.

As is visible in the displayed example, the tongue and lip PD curves at times move in synchrony and at times they are out of phase. In particular, in this data set lip PD with very few exceptions trailed behind tongue PD at the start of an utterance. This is true even for /s/ onsets, which based on previous results should have broken the pattern (see Experiment 3 in [6], /r/ is not comparable here as participant of Experiment 3 was a Finnish speaker).

5 Discussion

The tool presented here provides a method of viewing overall change in both ultrasound and lip video data as a time varying function. As has been previously reported, having access to such functions makes movement onset detection a fast and simple task [8]. It changes the work flow from time consuming back-and-forth viewing of video frames to just inspecting a single curve.

However, for detailed analysis of within-utterance-movements it would be useful to be able to access the videos – both the ultrasound sequence and the lip video – based on the PD curves. This would make it possible to not only find movement maxima and minima, but also inspect the quality of movement at a given point. Implementing a suitable GUI that will display the video frames that correspond to selections on the PD curve is a project for the near future.

Code availability

All PD analysis code was written in Python 3.7.4, NumPy 1.17.2, Scikit-video 1.11.1 and plots drawn with Matplotlib 3.1.1. The code is available as open source code under the GPL license as part of the Speech Articulation ToolKIT (SATKIT) [7, 9].

Acknowledgments

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 and lip videos from AAA and subsequent post-processing of the data. This work has been in part supported by a grant from the Emil Aaltonen Foundation.

References

[1] Alan H. Kawamoto, Qiang Liu, Keith Mura, and Adrianna Sanchez. Articulatory preparation in the delayed naming task. *Journal of Memory and Language*, 58(2) :347 – 365, 2008.

[2] Lotje van der Linden, Stephanie Kathleen Ries, Thierry Legou, Boris Burle, Nicole Malfait, and F.-Xavier Alario. A comparison of two procedures for verbal response time fractionation. *Frontiers in Psychology*, 5(1213) :1 – 11, 2014.

[3] 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.

[4] T. G. Csapó, K. Xu, A. Deme, T. E. Gráczi, and A. Markó. Transducer misalignment in ultrasound tongue imaging. In *Proceedings of the 12th International Seminar on Speech Production (ISSP 2020)*, pages 166 – 169, Online / New Haven, CT, 2020.

[5] M. Saito, F. Tomaschek, C.-C. Sun, and R. H. Baayen. An ultrasound study of frequency and co-articulation. In *Proceedings of the 12th International Seminar on Speech Production (ISSP 2020)*, pages 206 – 209, Online / New Haven, CT, 2020.

[6] P. Palo. *Measuring Pre-Speech Articulation*. PhD thesis, Queen Margaret University, Edinburgh, Edinburgh, 2019.

[7] Palo, P. and Moisik, S. R. and Faytak, M. SATKIT : Speech Articulation ToolKIT [Python software package]. Available in a public software repository, accessed 28 Aug 2021, 2020. `https://github.com/giuthas/satkit`.

[8] P. Palo. Can we detect initiation of tongue internal changes before overt movement onset in ultrasound ? In *Proceedings of the 12th International Seminar on Speech Production (ISSP 2020)*, pages 242 – 245, Online / New Haven, CT, 2020.

[9] M. Faytak, S. R. Moisik, and P. Palo. 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, 2020.