COMPLEX PATTERNS IN SILENT SPEECH PREPARATION: PREPARING FOR FAST RESPONSE MIGHT BE DIFFERENT TO PREPARING FOR FAST SPEECH IN A REACTION TIME EXPERIMENT

Sonja Schaeffler, James M. Scobbie and Felix Schaeffler

CASL, Queen Margaret University, Edinburgh
sschaeffler@qmu.ac.uk, jscobbie@qmu.ac.uk, fschaeffler@qmu.ac.uk

ABSTRACT

This paper presents articulatory data on silent preparation in a standard Verbal Reaction Time experiment. We have reported in a previous study [6] that Reaction Time is reliably detectable in Ultrasound Tongue Imaging and lip video data, and between 120 to 180 ms ahead of the standard acoustics-based measurements. The aim of the current study was to investigate in more detail how silent speech preparation is timed in relation to faster and slower Reaction Times, and faster and slower articulation rates of the verbal response. The results suggest that the standard acoustic-based measurements of Reaction Time may not only routinely underestimate fastness of response but also obscure considerable variation in actual response behaviour. Particularly tokens with fast Reaction Times seem to exhibit substantial variation with respect to when the response is actually initiated, i.e. detectable in the articulatory data.

Keywords: Ultrasound Tongue Imaging, video lip data, articulators, Reaction Time, speech preparation

1. INTRODUCTION

Silent articulatory movements in preparation for audible speech constitute a well-known although until now largely unquantified aspect of speech production, despite this transition occurring at the start of every single utterance, word or phrase.

The Verbal Reaction Time (RT) paradigm is perhaps the clearest case where the discrepancy between articulatory and acoustic onset of speech is not just observable in each and every data point, due to the nature of the paradigm's one-word utterance elicitation, but could have theoretical implications if this silent speech interval patterns in a way that cannot be predicted from standard Acoustic RT.

In standard psycholinguistic procedure the observable response is usually the acoustic onset of verbal feedback, and here, RT is usually determined via "voice key", a device which is triggered automatically as soon as the monitored sound pressure reaches a pre-defined level. However, what is detectable as the onset of acoustic output is only one, the final, stage in the speech production process. Before anything becomes audible, the articulators have already moved into place. This movement shows that the motor plan for the response has been put into action and is an earlier observable response than audible speech. Silent movements of some of the articulators (e.g. the lips and jaw) are clearly detectable by human observers, and so are relevant in natural conversation, with visual cues in turn-taking being an obvious example.

1.1. The articulatory advantage

In a previous study [6] we investigated Acoustic and Articulatory RTs for two speakers, and showed that articulatory measurements could capture Verbal RT reliably at a much earlier time point than acoustic measurements. Table 1 summarises our results, showing that while the speakers behaved very differently in how fast they responded overall, across speakers the duration of the Articulatory RT amounted to only around 75 - 80% of the Acoustic RT, giving the articulatory measures a 20 - 25% advantage. We also found that there was no pattern in the data that suggested a significant impact of phonetic target types on Articulatory RT measurements. This was crucial in light of findings that *Acoustic* RTs derived via the commonly used voice key are fairly susceptible to differences in onset type (cf.[3]).

Table 1: Means (SD) in ms for the three RT types for both speakers, as reported in [6].

	Speaker 1	Speaker 2
Acoustic RT	851 (251)	586 (127)
Tongue RT	670 (237)	442 (137)
Lip RT	677 (237)	466 (127)

1.2. Exploring silent articulation in relation to audible response in more detail

The current study investigated the existing data of the two pilot speakers in more detail in order to better understand the timing of speech preparation in relation to audible speech. One aim was to get a better idea of the size and nature of the error that

might be introduced by using standard acoustic measurements rather articulatory ones for determining RT, as is routinely done. The other aim was to explore whether measures of silent preparation, if used, could in fact serve as predictors for audible response.

Because our previous between-speaker comparison [6] suggested that a faster RT may result in a shorter duration of silent articulation (i.e. Speaker 2 exhibiting overall 'compressed' silent articulation durations in keeping with her overall shorter RTs) we wanted to test here whether this pattern would hold within-speaker and per token. *Does a shorter duration of silent preparation reliably predict a faster onset of response?*

It also seemed conceivable that a faster articulated verbal response should be preceded by a faster articulated, i.e. shorter, silent target-specific articulation phase. *Does a shorter duration of silent preparation reliably predict a faster speech rate of audible response output?*

Finally we wanted to explore how non-linguistic movements of the articulators that are not yet target-specific but may aid preparation for speech (e.g. parting the lips for a following bilabial gesture) are tied in in time with overall fastness of response. We had observed these movements routinely in the data but had them not yet quantified. *Would such additional movements of the tongue and lips slow down overall response times?* That might indicate that the speaker was not quite speech-ready, or maybe not quite speech-ready for the target-relevant motor plan.

2. METHOD

2.1. Speakers and Material

We report data from the same two female native speakers of Scottish varieties of English (aged between 20-35) as in [6]. Participants had no visual or hearing impairments. Verbal responses were collected in a standard picture naming task, with the 260 items from the well-tested and frequently used Snodgrass-and-Vanderwart picture inventory [8], presented in colour (cf. [5]).

2.2. Instrumentation and data synchronisation

All articulatory and acoustic recordings were obtained simultaneously and synchronised carefully (cf. [6]) using AAA software from Articulate Instruments Ltd [2]. The participants were fitted with a purpose-built headset to ensure stabilisation of the ultrasound probe [1, 7]. Attached to the helmet was a small Audio Technica AT803b microphone for high-quality acoustic recordings, plus a NTSC micro-camera to capture recordings of the speakers' lips.

Ultrasound recordings were obtained at a rate of 201 frames per second from a SonixRP system (cf. [9]). Video was captured then deinterlaced to an effective rate of 59.95 fps. Recordings started 1.5 seconds before prompt presentation so that the whole speech production process was captured.

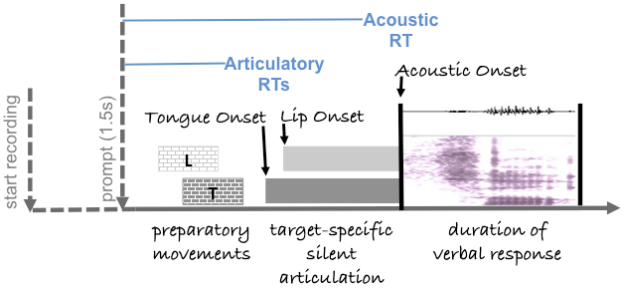
3. TOKEN SELECTION, ANNOTATION AND MEASUREMENTS

Of the 260 targets presented, 19 elicited erroneous responses in Speaker 1 and were excluded, leaving 241 tokens for annotation. Speaker 2 stopped after recording of 156 tokens. Of the recorded tokens, four elicited erroneous responses, leaving 152 tokens for annotation for Speaker 2. Targets that elicited British English variants instead of the American English target (e.g. 'waistcoat' instead of 'vest') were not excluded but were re-coded phonemically.

3.1. Annotations

For [6] we had annotated for each token a) the onset of target-specific lip movement b) the onset of target-specific tongue movement and c) the acoustic onset. For the current study we annotated in addition d) non-target specific preparatory lip-movements (preL), e) non-target specific preparatory tongue-movements (preT) and f) the acoustic duration of the verbal response (cf. Fig. 1).

Figure 1: Schematic view of relevant landmarks for annotation and analysis



3.1.1. Articulatory annotations and criteria

Only articulatory movements that occurred after 1.5s, i.e. after prompt presentation, were considered.

To be target-specific (and with that part of the response), the relevant articulator had to be seen to move in a single, smooth, contiguous manner towards one of the initial articulatory targets for the word. As reported in [6] we gave all our articulatory annotations confidence ratings from '1 = very unsure' to '5 = absolutely sure', and only used tokens for further analysis with a 4 or 5 rating for

both lips *and* tongue. This left 132 tokens (55%) for Speaker 1 and 82 tokens for Speaker 2 (54%).

For the purpose of this study new annotation categories of preparatory but not yet target-specific lip and tongue movements (preL and preT) were used in cases where movements were observable that clearly facilitated a subsequent target-specific gesture but that themselves did not contiguously move towards the articulatory target. A non-target specific movement might have been e.g. a slow, vague movement of the tongue downwards from a resting position that was subsequently followed by a fast, unambiguous tongue movement into velar closure for a velar target. A preL or preT could also be followed by a static holding phase. In either case, the end of movement marked the endpoint of the annotation (cf. Fig.1). Movements presumably related to false starts were not coded as preL/preT.

3.1.2. Acoustic criteria and measurements

The acoustic onset (and offset for response duration) was determined by visual inspection of the speech signal (cf. [6]). For speech rate measurements, we coded for each token number of syllables (1-4) as well as number of segments (1-11) as produced by the speaker (affricates and diphthongs counted as two segments given that they have more gestural activity). Divided by the overall duration of the audible response this yielded average syllable and segment rates for audible parts of each token.

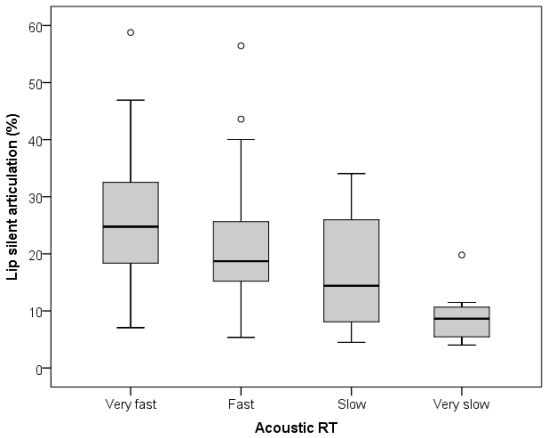
4. RESULTS

4.1. Silent articulation durations related to overall RTs

To test whether the duration of silent articulation could predict speed of response we investigated the relation of Lip and Tongue RTs to the respective silent articulation durations (SAD). If silent articulation constituted a relatively stable percentage of the Acoustic RT we would expect a positive correlation between these two durations. However, the observed effect was negative [Speaker 1: Lip RT-Lip SAD, Pearson's r=-.42, p<.001, Tongue RT-Tongue SAD, Pearson's r=-.42, p<.001; Speaker 2: Lip RT-Lip SAD, Pearson's r=-.28, p.012, Tongue RT-Tongue SAD, Pearson's r=-.35, p=.001]. Tokens with *longer* Articulatory RTs showed a tendency to exhibit *shorter* silent articulation durations.

This effect can be further illustrated by looking at the duration of silent articulation duration as a percentage of the Acoustic RT. We divided Acoustic RTs into four classes (upper limits: very fast <0.75 s, fast < 1 s, slow <1.25 s and very slow < 1.5 s). Fig. 2 shows the distribution of Lip silent articulation percentages over the four Acoustic RT classes for Speaker 1. It can be clearly seen that the percentage of silent articulation decreases with longer RTs. This effect is mainly a consequence of a reduction in the upper limits of percentages, not the lower limits. For tokens with fast RTs, the silent articulation durations can make up a sizeable percentage of the Acoustic RT (up to 40-60%), while for tokens with slower RT the maximum percentages found are smaller, barely exceeding 15-20% for tokens with the slowest RTs.

Figure 2: Percentage of Lip silent articulation across four categories of tokens with very fast to very slow Acoustic RT in Speaker 1.



4.2. Silent articulation durations in relation to audible response speech rate

To test whether silent articulation durations could predict the articulation rate of audible response we explored the effect of speech rate on the onset and duration of silent, target-specific articulation. Speaker 1 had an average syllable rate of 4.8 (SD=3.0) syllables/second, and an average segment rate of 11.8 (SD=4.2) segments/second. Speaker 2 had an average syllable rate of 4.6 (SD=1.9) syllables/second and an average segment rate of 14.5 (SD=7.1) segments/second, showing that while Speaker 2 reacted faster in the current experiment she did not speak faster, at a syllable rate. Importantly, there was no correlation in either speaker to suggest that the duration of silent articulation could predict the articulatory rate of the verbal response. The two speech rate measures did not show any significant correlations with Acoustic RT, Tongue or Lip RT, or Tongue or Lip silent articulation duration.

4.3. Preparatory but not yet target-specific movements in relation to overall response times

Preparatory but not yet target-specific preL and/or preT could only be observed in some of the tokens. In Speaker 1, 59 of the analysed 132 tokens

exhibited detectable pre-preparatory movements of the lips (n=14), tongue (n=31), or both (n=14). In Speaker 2, 17 of the analysed 82 tokens exhibited detectable pre-preparatory movements of the lips (n=6), tongue (n=5), or both (n=6). The presence of these movements may well be indicative of speakers having to make adjustments to articulators that are not quite in place for the start of target-specific articulatory onsets, but at the same time these movements do not generally seem to influence overall RTs (cf. Figs. 3 and 4). As sample sizes were very imbalanced, we refrained from statistical testing, but descriptive values do not indicate major influences of preparatory movements on Articulatory and Acoustic RTs. One exception is Speaker 2's 'preL only' category, but dispersion in this category is large as indicated by the error bars. However, this category warrants further testing in a larger sample, as even for Speaker 1 this is the category with the longest mean RTs, even if only by a small margin.

Figure 3: Occurrence and timings of preL and preT in Speaker 1.

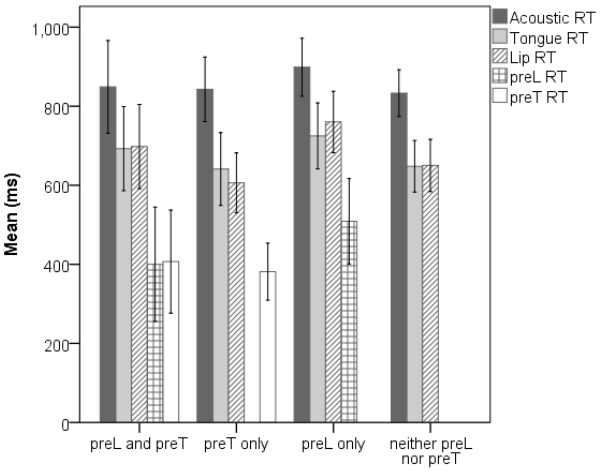
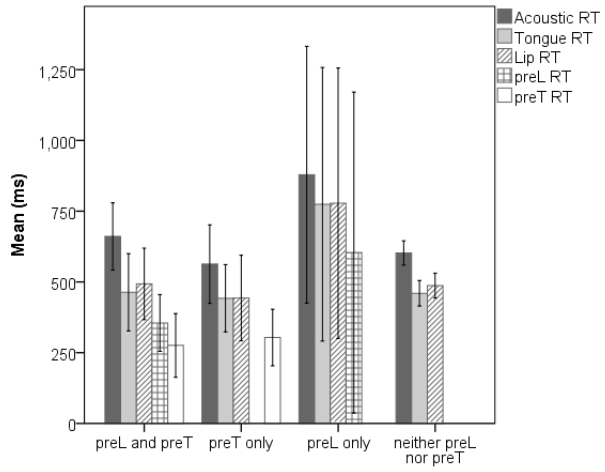


Figure 4: Occurrence and timings of preL and preT in Speaker 2.



5. DISCUSSION AND CONCLUSION

In [6] we had found a robustly measurable, non-negligible and consistently present discrepancy between articulatory and acoustic onset of speech in an experimental setting where earliest possible detection of verbal response is crucial. While the two speakers behaved very differently (with Speaker 2 reacting much faster than Speaker 1), the average advantage of the articulatory measures amounted to a very similar 20 to 25 % for both. Investigating the data here in more detail we found a number of patterns that may have implications for our understanding of speech preparation, and speech planning more generally.

Contrary to what we expected, across tokens faster RTs did not equal proportionally shorter silent articulations - so here we did not observe a compression effect as could have been maybe implied by the averages across speakers. There was even a tendency for tokens with particularly *slow* Articulatory RTs to exhibit particularly *short* silent articulations. One could speculate that speakers may have various strategies available to compensate during implementation of the motor plan, maybe, for example, by shortening or speeding up the silent articulation phase when the onset of verbal response is already self-monitored as being delayed. Whatever the reasons are for these more complex patterns, using only acoustic measurements to determine RT may well obscure considerable variation in actual response behaviour.

We also found that the duration of the silent articulation phase was not related to the articulation rate of the audible verbal response. We simply do not know enough about the optional and the obligatory parts of preparation to take this as evidence that speakers can vary articulation rate between silent and audible parts of speech, but this is something that could be investigated in more depth in a larger-scale study.

Lastly, we found that the presence of additional preparatory but not yet target-specific lip and tongue movements does probably not slow down articulatory or Acoustic RTs. Particularly here, more data is needed, but it seems to again suggest that speakers have various ways of compensating in the preparation of response.

To investigate these issues more thoroughly we need a systematic manipulation of relevant variables such as speech rate, task, speaker, and also linguistic complexity of the elicited material (cf. [4]). Taken together the results of this pilot study suggest that speakers employ a number of distinct strategies that are all part of speech preparation and planning, and which interact in complex ways.

6. REFERENCES

[1] Articulate Instruments Ltd. 2008. *Ultrasound Stabilisation Headset Users Manual: Revision 1.4*. Edinburgh, UK: Articulate Instruments Ltd.

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

[3] Kessler, B., Treiman, R., & Mullennix, J. 2002. Phonetic biases in voice key response time measurements. *Journal of Memory and Language*, 47, 145–171.

[4] Lehiste, I. 1972. The timing of utterances and linguistic boundaries. JASA 51, 2018-2024.

[5] Rossion, B., & Pourtois, G. 2004. Revisiting Snodgrass and Vanderwart's object set: The role of surface detail in basic-level object recognition. *Perception*, 33, 217-236.

[6] Schaeffler, S., Scobbie, J.M., & Schaeffler, F. 2014. Measuring Reaction Times: Vocalisation vs. Articulation. *Proc. 10th ISSP*, Cologne, 379-382.

[7] Scobbie, J.M., Wrench, A.A., & van der Linden, M. 2008. Head-Probe Stabilisation in Ultrasound Tongue Imaging Using a Headset to Permit Natural Head Movement. *Proc. 8th ISSP*, Strasbourg, 373-376.

[8] Snodgrass, J.G., & Vanderwart, M. 1980. A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 174-215.

[9] Wrench, A.A. and Scobbie, J.M. 2011. Very high frame rate ultrasound tongue imaging. *Proc. 9th ISSP*, Montreal, 155-162.

ACKNOWLEDGEMENTS

We are grateful to our participants, as well as Steve Cowen and Alan Wrench for their ongoing technical and moral support. We also gratefully acknowledge funding from the ESRC (RES-000-22-3032) which helped to lay the foundation for our work on speech preparation.