SAY AGAIN? INDIVIDUAL ARTICULATORY STRATEGIES FOR PRODUCING A CLEARLY-SPOKEN MINIMAL PAIR WORDLIST

James M. Scobbie and Joan Ma

Queen Margaret University, Edinburgh
jscobbie@qmu.ac.uk, jma@qmu.ac.uk

ABSTRACT

We describe articulatory differences (lingual and labial) between two versions (neutral and clear) of a CVC wordlist of 12 targets (V = /ieaɔou̟/; C_C = /p_p/ or /m_m/). A companion paper describes the background; the participants, materials and tasks; the impressionistic and acoustic results.

Labial measures reflect vowel opening (and edge-spreading) and consonant compression using fleshpoint markers captured by head-mounted video. Consonant closure and total word duration are based on visual judgement of complete closure. Ultrasound data provides the absolute area between neutral and clear mid-sagittal tongue-surface splines at the maximum of each vowel target, and a qualitative description of tongue shape and location.

Strong and systematic interspeaker variation was evident in how articulation, acoustics and functional clarity were enhanced. Some large phonologically motivated segmental hyperspeech enhancements were observed, but they were not related straightforwardly to the phonological oppositions in the material nor consistently used by all speakers. Differences in utterance initiation are also discussed.

Keywords: Lombard speech, hyperspeech, ultrasound, intelligibility, labials.

1. INTRODUCTION

A functional drive for speech to be clearer, i.e. more intelligible than it would otherwise be, can arise in a range of contexts more challenging than the norm, particularly through attenuation of the acoustic signal. A speaker can increase their clarity using a variety of means, which can be studied experimentally using a variety of approaches. A large body of research has found some common mechanisms (potentially universal in origin) but also a fair amount of variation [3] [6] [7] [12].

Research into the *articulation* of clear speech is more unusual. It primarily uses flesh-point tracking e.g. with electromagnetic articulography (EMA) [7] or motion capture [12], or video analysis of the face and lips [9]. The non-invasive character of the latter is appealing, particularly when the research involves fieldwork, child speakers, naturalistic settings, clinical interventions or existing video corpora.

When articulatory instrumentation is used, speech is often elicited in relatively short and tightly structured tasks, e.g. reading short wordlists or sentences aloud. Such experimental demands might be thought to make speech more formal, clear, and hyperarticulated than natural everyday speech. And though vernacular and naturalistic speech occurs during articulatory investigation of dialogue, less work also addresses single words elicited in isolation through picture naming or reading aloud [10]. Interactive discourse is better for the study of global aspects of clear speech [7] [9] [12], but it is easier to focus on specific phonological contrasts using more traditional experimental methods.

Speech therapy and language learning also involve interactions very different from day-to-day conversation, rich in meta-linguistic feedback. They elicit clear speech, albeit "clear" in a different sense to speech in noise. The goal may be to introduce a perceptible phonemic contrast, or to enhance an established one, to be more accurate and intelligible.

While audio recordings of such interactions are relatively easy to obtain in principle, articulatory data is not. It has almost entirely been studied in the clinical domain, mostly with electropalatography [14], because it has been being used therapeutically for real-time feedback. More recently, ultrasound tongue imaging has emerged as a feedback tool with cost and ease-of-use advantages. Longitudinal datasets of clinical interaction can be collected during therapy [4], and are being made available [5], incorporating ultrasound, audio and videos of the lips. Similar language learning corpora will follow.

Therefore we think it useful to examine clearly spoken wordlists using the non-invasive articulatory techniques of ultrasound tongue imagine and facial video. Here and in a companion paper [13] we compare neutral (non-interactive) wordlist-speech with a clearly-spoken (interactive) alternative, with a primarily phonological goal: understanding how and whether phonological contrasts are enhanced. We adapt existing methods and use them to explore if and how clear speech differs from standard baseline productions of wordlists read aloud.

2. METHOD

The elicitation protocol, materials, and acoustic segmentation and analysis are described elsewhere [13], as are the ultrasound hardware and audio-ultrasound synchronisation [15]. Here we focus on more novel aspects of the method.

A commercially-available stabilising headset was used to stabilise the ultrasound probe [1] while permitting natural head-movement during interaction with interlocutors. The headset comes with an option to mount a micro-video-camera (interlaced colour VGA NTSC output rated at ~30 frames per second, de-interlaced to 60 fps). Previous work has used this type of fixed-perspective camera for various purposes. The consistent viewpoint makes automatic speech-recognition (e.g. for a silent speech interface) much more tractable [8]. It assists in qualitative analysis (including transcription). It is useful also to evaluate ultrasound probe stability within the mid-sagittal plane, through comparison of upper incisors with visible parts of the headset, because their relative positions should be fixed.

Typically, a single camera is mounted on an adjustable sagittal frame, positioned to give a profile view of the lips. Sometimes a second camera is used, e.g. in the Dynamic Dialects website [11], which used a video mixer to pre-mix images from the two cameras together before digitisation.

Figure 1: American Speaker 44, Georgia, USA. "Goose". From [11]. Screen grab.

We adapted this procedure, additionally affixing small strips of sterile adhesive white tape (Micropore™, manufactured by 3M). One strip was attached in a mid-sagittal location onto the upper lip either entirely within the vermillion zone or crossing over the vermillion border onto the face. Another was likewise placed onto the lower lip in such a way that one quasi-horizontal edge of each strip was always visible (Fig 2). The inner white straight edges of the tape provided the main reference for the analysis of upper-lower lip mid-sagittal aperture and constriction, as viewed by the frontal camera.

An additional tape strip was attached just superior to the corner of the mouth and in view of the profile camera. It was hard to find an optimal or replicable location for this location, and the orientation and size of the reference tape strip in the video image was much more affected by skin distortions than the mid-sagittal lip tapes.

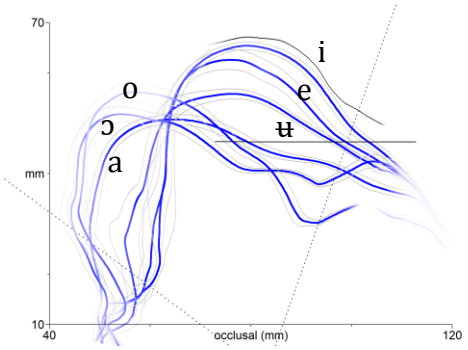
Figure 2: Edges of white adhesive surgical tape provided labial references on the lips. Small adhesive 3D hemispheres (3mm diameter) and/or a blue pen-drawn × or + mark specific fleshpoints.

Linear measurements (arbitrary units) were made from the frontal images (Fig 2 upper panel) between the edges of the strips at time-points corresponding to maximal opening (for the vowel targets) and maximal constriction (for the consonant targets). A linear measurement of the distance in the profile image from the centre of the upper lip 3D hemisphere to the centre of the corner hemisphere (Fig 2 lower panel) estimated maximum lip spread.

Ultrasound data capture, measurement and analysis used Articulate Assistant Advanced™ [2], using Ultrasonix hardware. A micro-convex probe scanned a 135° field of view using 63 hardware scanlines at a frame rate of 121fps. Automatic edge-tracking was performed in AAA for each frame in and around a word, using a vowel-specific guide template that set an envelope within which the bright regions corresponding to the tongue surface were tracked as a spline. For a typical CVC word with around 500ms of articulatory activity (Table 1), approximately 60 frames were auto-tracked.

For quantitative analysis, extreme anterior and posterior regions of the image not corresponding to the tongue were discarded. Parts of the image containing parts of the tongue tip and lower-root data that were not well-imaged or tracked were also excluded (Fig 3). Thus analysis was limited to a sector of interest, comprising 23 AAA analysis fanlines for S2-4, and 21 for S1 (around 76° and 69°, respectively). AAA recorded confidence levels for the edge-tracked splines within this sector.

Figure 3: Example (S4) average mid-sagittal tongue splines (thick lines) rotated to occlusal plane. The sector of interest is between the two dashed radial fanlines. Lower confidence areas of splines are paler. These six neutral vowels illustrate the whole vowel space.



Seven splines at equal increments during each acoustic vowel were exported to the workspace from the six CVC tokens of that target. Thus 42 splines were averaged to create a profile for each neutral and each clear vowel. For each vowel, the built-in AAA difference function was used to measure a clear-neutral difference and estimate its significance (correcting for the non-independence of the time-normalised splines from within each vowel by adjusting p-values by a factor of 7). For each of the 23 (or 21) analysis fanlines, the two conditions were compared, and flagged as "different" if 5 or more contiguous fanlines (16.45° or more) were significantly different. The absolute linear radial distance between the clear and neutral vowels was averaged for *all* fanlines in any case. Given so few participants (n=4) and so many measures, this pilot study reports indicative results descriptively.

3. RESULTS

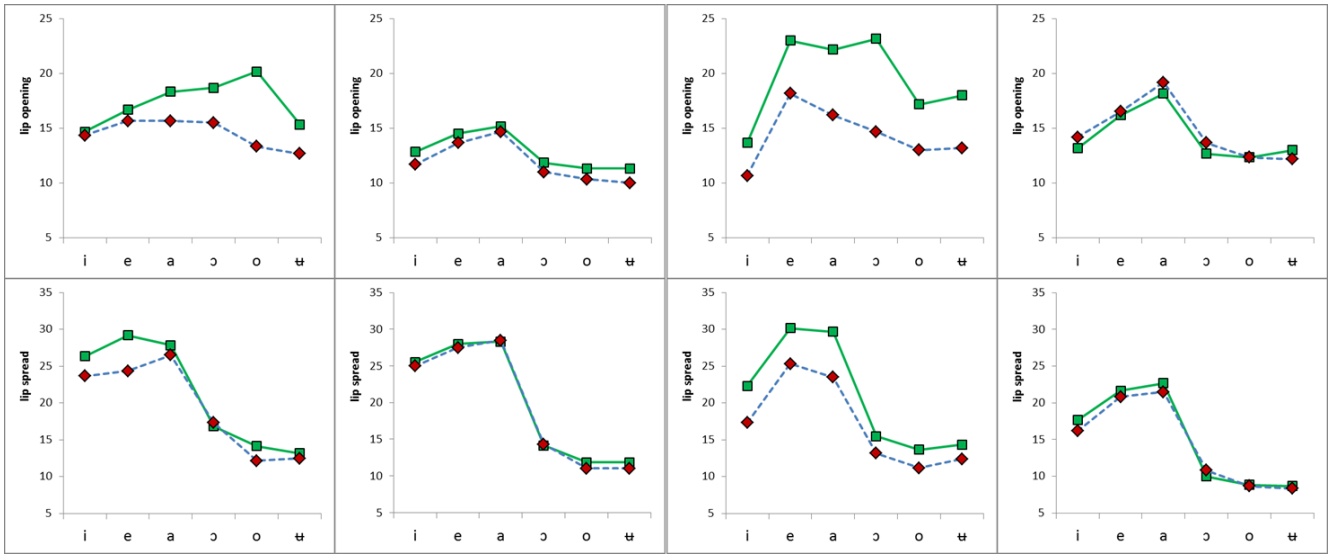
For three speakers, articulatory word duration ($C2_{end}$-$C1_{start}$) appeared longer in the clear condition (Table 1), for /m/-words and /p/-words alike, and for almost individual vowel pairs. See below for S2.

Table 1: Articulatory word duration (ms) from the start of C1 closure to the end of C2 closure.

	/m_m/		/p_p/	
	neutral	clear	neutral	clear
S1	487	571	515	532
S3	571	630	559	658
S4	378	504	321	559

S1 and S3 showed a big increase in lip opening for some or all vowels in clear speech (Fig 4 upper panel). Perhaps S2 showed a very slight change. No consistent difference was observed for S4. The vowel with most opening varied. For lip spreading, the effect of the phonological specification of roundness was much more clear than vowel height (Fig 4 lower panel). There was no clear speech change (apart probably for S3) in spreading. /m/-words and /p/-words behaved alike, so were pooled.

Figure 4: Upper panel = lip opening, lower panel = lip spreading. Clear (solid) vs. neutral (dashed), S1-4 (left to right). Arbitrary units. /m/-words and /p/-words were pooled, so each data point is based on six tokens.



For C1 duration (not shown), S1 and S3 had a consistent increase across all vowel contexts for /m/-words. S4's pattern was not clear. S2 could not be measured: this speaker tended to start each trial with a closed mouth that could not be differentiated from the initial labial consonant (Table 2). Final /p/ was generally released with a burst and final /m/ was usually released (silently) after the offset of voicing.

Only S3 seemed to make a difference in lip compression in clear speech (not shown), by compressing both /m/ and /p/ in both C1 and C2.

Table 2: Number of tokens (of 18) in which C1 articulatory duration could not be measured due to lack of lip opening prior to word production.

	/m/		/p/	
	neutral	clear	neutral	clear
S1	10	0	0	0
S2	15	11	17	16
S3	0	0	0	0
S4	13	2	11	1

In the ultrasound data, it is worth noting the extent of fronting and lowering of /ʉ/ (GOOSE and FOOT) vowel (e.g. Fig 3). S1 and S2 distinguished clear from neutral versions of the vowels (Table 3), while S3 and S4 seemed not to. See also [13]. In absolute terms, the radial distance differences were extremely small.

Table 3: Number of analysis fanlines flagged with a neutral vs. clear difference in AAA. (Bracketed if fewer than 5 were contiguous.) Diff is the average radial difference (mm).

	i	e	a	ɔ	o	ʉ	Diff
S1	18	16	(9)	6	6	10	1.26
S2	7	7	16	20	21	12	1.33
S3	(3)	14	5	(2)	(3)	(3)	1.12
S4	(7)	(2)	15	(3)	(5)	(0)	0.85

Speech initiation showed interesting speaker-specific effects. All speakers waited for the prompt to appear on the screen with a closed mouth, and three then nearly always opened their mouth (with or without an in-breath) before initiating word production, which included labial closure needed to produce the word-initial labial consonant. Speaker S2, on the other hand, was tight-lipped, in the sense that 72% of /m/ words and 92% of /p/ words were initiated from a closed-mouth resting position without any intervening lip-opening (Table 2). S1's 10 tokens of this "stay-closed" type were all /m/ in the neutral condition, suggesting an interaction of segmental and stylistic planning. S3 on the other hand always first opened their mouth, initiating their *segmental* labial closure from an open mouth starting point. S4 had 24 out of 27 cases lacking any mouth opening, but these were strongly pattern by task: "stay closed" initiations were all were in the neutral condition. Unlike S1, this applied equally to /m/ and /p/. At offset, S4 was unusual in not releasing C2 every time: five cases all involved /m/.

4. DISCUSSION

As well as more participants, a control experiment is needed, comprising two neutral conditions. This will provide useful information on zero-effects.

The small adhesive 3D hemispheres provided a reference point reliably visible to a profile camera. Surgical tape provided a safe, non-interfering and discomfort-free bed for fixing them, and the edge of the adhesive tape was in fact very easily tracked (in manual measurement) in the frontal camera data.

Automatic analysis of small 3D objects of a contrastive colour to facial and vermillion-lip skin tone (and white tape) ought to be achievable, e.g. with feature-extraction methods [8]. This would facilitate the analysis of protrusion, constriction and compression by adding fleshpoints within the vermillion zone to measures of the cross-sectional area of lip aperture, wireframe 3D models of the lips, and the kinematics of more familiar fleshpoints lying outwith the lips themselves (e.g. [6] [7] [9]).

Here we were limited to 2D planar analysis and did not correct for changes in depth (hence arbitrary units of measurement were used, particularly relevant for the profile camera). The frontal camera gave a reasonable view of complete closure as well as fleshpoint minima and maxima from tape.

4. CONCLUSION

Our articulatory results (also [13]) support previous observations that when enhancing intelligibility in difficult communicative environments, we should expect speaker-specific behaviour. Individuals seem to vary *systematically* in how they approach the discriminability of phonological contrasts, even within-task and within-dialect. It is difficult to conclude that clear speech enhances phonemic oppositions in a straightforward general way such as a uniformly maximised dispersal in multi-dimensional phonetic space. Even for binary feature oppositions e.g. /m/ vs. /p/, let alone the multiple oppositions in a vowel inventory, many options are possible. Though additional participants will enable meaningful statistical analysis, we do not expect this conclusion to change.

In addition to global phonetic augmentation (e.g. greater intensity and overall duration), holistic aspects of clear speech also involve the ways in which speakers plan and implement transitions from non-speaking pre-speech resting positions into speech itself. We therefore agree that there can be phonologically-relevant responses to difficult communicative conditions. The specific enhancements used, however, and the phonological oppositions they relate to, seem likely to vary.

5. REFERENCES

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

[2] Articulate Instruments Ltd. 2018. *Articulate Assistant Advanced* Version 2.17. Software.

[3] Castellanos, A., Benedi, J. M., Casacuberta, F. 1996. An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect. *Speech Comm*. 20, 23–35.

[4] Cleland, J., Scobbie, J.M., Heyde, C., Roxburgh, Z., and Wrench, A.A. 2019. Enabling new articulatory gestures in children with persistent speech sound disorders using ultrasound visual biofeedback. *Jou. Sp. Lang. Hear. Res.*. 62 (2): 229-246

[5] Eshky, A., Ribeiro, M.S., Cleland, J., Richmond, K., Roxburgh, Z., Scobbie, J., Wrench, A. 2018. UltraSuite: a repository of ultrasound and acoustic data from child speech therapy sessions. *Proc. 19th Interspeech* Hyderabad.

[6] Garnier, M., Ménard, L. Alexandre, B. 2017. Hyper-articulation in Lombard speech: An active communicative strategy to enhance visible speech cues? *J. Acoust. Soc. Am*. 144(2), 1059–1074.

[7] Hazan, V., Kim, J. 2013. Acoustic and visual adaptations in speech produced to counter adverse listening conditions. *Proc. AVSP'13* Annecy, 93–98.

[8] Hueber, T., Benaroya, E.-L., Chollet, G., Denby, B., Dreyfus, G., Stone, M. 2010. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Comm*. 52(4), 288–300.

[9] Kim, J., Davis, C. 2014. Comparing the consistency and distinctiveness of speech produced in quiet and in noise. *Comp. Speech Lang*. 28, 598–606.

[10] Lawson, E., Stuart-Smith, J., Scobbie, J.M. 2008. Articulatory insights into language variation and change: preliminary findings from an ultrasound study of derhoticisation in Scottish English. *NWAV* 36, 102-110.

[11] Lawson, E., Stuart-Smith, J., Scobbie, J.M., Nakai, S. 2018. *Dynamic Dialects: An Articulatory Web Resource for the Study of Accents*. University of Glasgow. Accessed 7/12/2018. https://www.dynamicdialects.ac.uk/

[12] Mixdorff, H., Pech, U., Davis, C., Kim, J. 2007. Map task dialogs in noise: a paradigm for examining Lombard speech. *Proc. ICPhS* Saarbrücken, 1329–1332.

[13] Scobbie, J.M.,. Ma, J. 2019. Say again? Individual acoustic strategies for producing a clearly-spoken minimal pair wordlist. *19th ICPhS* Melbourne.

[14] Scobbie, J.M., Wood, S.E., Wrench, A.A. 2004. Advances in EPG for treatment and research: an illustrative case study. *Clin. Ling. Phon.* 18(6-8), 373–89.

[15] Wrench, A.A., Scobbie, J.M. 2016. Queen Margaret University ultrasound, audio and video multichannel recording facility (2008-2016). *CASL Working Papers, Queen Margaret University*, WP-24. Accessed 30/11/2018. http://eresearch.qmu.ac.uk/4367/