An Evaluation of Inter-Speech Postures for the Study of Language-Specific Articulatory Settings

Sonja Schaeffler
Speech Science Research Centre, Queen Margaret University, Edinburgh, UK

James M. Scobbie
Speech Science Research Centre, Queen Margaret University, Edinburgh, UK

Ineke Mennen
ESRC Centre for Research on Bilingualism in Theory and Practice, University of Wales, Bangor, UK

E-mail: sschaeffler@qmu.ac.uk

Abstract

We present a methodological study evaluating Inter-Speech Postures, i.e. vocal-tract configurations achieved in the silent preparation for speech which have been claimed to be indicative of articulatory settings. The term articulatory setting refers to a characteristic use of the articulators believed to shape the overall phonetic realisation of a language and with that, possibly, its 'typical sound' [3].

1. Introduction

The departure point for the current study is the experimental protocol presented in Wilson [6], which in turn builds on work from Gick et al [2]. Wilson used ultrasound and the motion-capture system Optotrak to gather articulatory data. His results suggest that in inter-utterance pauses, in preparation for speech, different languages may exhibit a specific physical oral posture (cf. [6]). Being language-specific, these postures must be learned by native speakers, and so have been claimed to be a plausible physical instantiation of the language's articulatory setting.

The longer-term objective of our research is to develop capacity for analysing cross-linguistic differences in articulatory settings systematically, in a large number of speakers and for a broad range of languages. Data on these silent "Inter-Speech Postures" (henceforth ISPs), is suitable for this purpose as it appears to provide vocal tract information without confounding effects from lexis, phonotactics or phonological inventory, which greatly simplifies cross-linguistic comparisons. However, it is necessary to explore ISPs in more detail, particularly the environments which give rise to them, and their role in mediating between non-speech and the active (and often silent) movements which are observable in the run-up to speech. Understanding the nature and limitations of data gathered from ISPs will help to obtain more readily interpretable data in future studies. Another issue is that Wilson [6] used a read-speech paradigm making it necessary to examine ISPs in more natural speech.

In order to determine an operational definition of ISPs as well as to investigate their ecological elicitation we obtained electropalatographic (EPG) data. We opted for EPG data because it provides useful information not just on overall tongue-palate contact patterns during speech but also on non-speech activities (swallowing, bracing etc.). To supplement our EPG data we also carried out informal inspections of already existing ultrasound data [5]. Both data sets only comprised of data from one language, English, as monolingual data are sufficient to determine the methodology to be used in future cross-linguistic comparisons. For the purpose of the current paper, we will only to a very limited extent report on the ultrasound data.

2. Method

2.1. Speakers and Instrumentation

Three speakers (1 female, 2 male, 28 to 48 years of age) were recorded in a sound-attenuated room at Queen Margaret University, Edinburgh. To obtain the EPG data we employed the WinEPG system (Articulate Instruments Ltd, cf. [1]) and its associated software for data capture and analysis (Articulate Assistant Advanced) at a sampling rate of 200Hz. EPG was synchronised with acoustic data, sampled at 22.1kHz. Acoustic data was used to keep track of what was being uttered by the speaker, and to determine the relative timing of audible speech to silent coarticulatory movements associated with particular segments and to the start of the zone in which an ISP itself can be found.

2.2. Speech Tasks and Material

Speakers were administered three different speech tasks in order to elicit ISPs. The tasks were performed in the same order by all speakers.

(1) A read-speech task comparable to that in [6], though with single words rather than sentences presented orthographically on screen.
(2) A Picture Naming Task with simple pictures presented on screen.
(3) A semi-spontaneous Map Task in which speakers described a simple route to a listener.

In Tasks 1 and 2, the length of pauses between speech prompts was systematically varied at 2, 5 and 8 seconds, as was the segmental context following the pauses (alveolar vs. non-alveolar vs. vocalic onset). Task 3 did not enable any such control.

2.3. Procedure and Measurements

In Tasks 1 and 2, data was gathered following an audible beep for 2, 5 or 8 seconds before the orthographic or picture prompt appeared, continually until after the end of the acoustic output. Data was gathered continuously during Task 3. Analysis of EPG data was carried out by using standard algorithms implemented in the AAA software to determine *'alveolar contact'* as well as *'total overall contact'* and *'centre of gravity'* (cf. [7]).

For annotation of ISPs, we first identified the presence of any notable change in the overall contact pattern from the time where the prompt appeared and the acoustic onset of speech. An ISP zone was labeled when the transition between the pre-prompt (non-speech) position and the first speech gesture was neither a mere interpolation nor random movement. An ISP was identified in this phase (see Figure 1) if the kinematic record indicated a motion towards some configuration ①, followed by smooth movement away from it towards the first segment ②, or a clear pause during a continuous motion. In Figure 1 the change at ② is due to jaw raising for the acoustic [m] at ③.

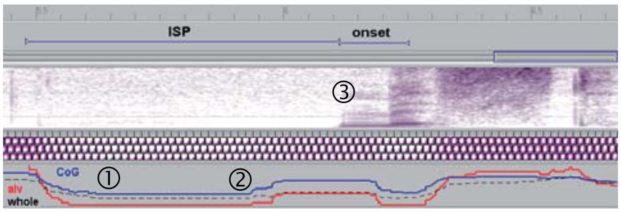

Figure 1. *ISP zone before acoustic onset of 'mist'. ISP would be found between ① and ②.*

EPG is particularly useful for investigating the distribution of ISPs in the subset of cases in which speakers hold part of their tongue against the roof of the mouth during non-speech. As speakers vary greatly in their non-speech activities, other techniques such as ultrasound are required for a fuller understanding. We cannot tell from EPG alone whether the ISPs which follow linguo-palatal contact differ (through coarticulation, perhaps) from tokens in which the speaker is open-mouthed during non-speech.

3. Results

We set out to explore how different conditions impact on the formation and dynamic structure of the ISP. We were specifically interested to see if ISPs occurring in spontaneous pauses occur more or less often than those occurring in prompted pauses, and how long a pause has to be to give rise to a measurable ISP. We found that speakers exhibit very different habits in the transition to speech.

3.1. Timing of Pauses in Prompted Speech

In prompted speech (Tasks 1 and 2) the prompts appeared at least 2, 5 or 8 seconds after the end of the preceding utterance was completed and saved to disk (which took about 6 seconds), so all pauses were long. Still, there were important differences between conditions. The pause in the 2-second condition often seemed too short for speakers to adopt a non-speech rest position (and from there an ISP) following the beep, especially when the task had gone on for a while. Speakers frequently missed the prompt or performed random non-speech movements (such as groping the palate) at the time of the prompt. Such movements are often difficult to tell apart from ISPs and thus can impede the selection of valid tokens. A delayed start is equally problematic as it makes it more difficult to decide where pausing ends and speech preparation begins. Similar problems occurred in the 8-second condition where the pause seemed too long. Speakers often stopped paying attention and missed the prompt or, when the prompt finally appeared, performed random non-speech movements.

The 5-second condition was eventually deemed most suitable, with the pause being long enough for speakers to get into rest position and from there into ISPs, but short enough to keep speakers to task. Based on this finding all data from the 2- and 8-second conditions was discarded, and only data from the 5-second condition underwent further analysis.

3.2. Proportion of ISPs in Prompted Speech

We identified 86 ISP zones in 140 pause tokens (i.e. 61.4% of tokens). ISP zones began approximately half a second after presentation of the prompt, 454 ms before the acoustic onset (s.d. 275 ms). There were significant differences between speakers (one-way ANOVA, F(2,85) = 6.892, p=.002; cf. Table 1) but not between speech tasks (i.e. picture naming vs. word list) or following segmental context (i.e. alveolar vs. non-alveolar vs. vocalic onset). Between-speaker differences in speech rate might have played a role, but we suspect habitual differences between speakers (perhaps merely occasion-specific) are far more likely to explain the result.

Table 1. *Means and standard deviation of ISP zone duration per speaker*

	ISP zone duration (in ms)
Speaker 1	388 (178)
Speaker 2	286 (225)
Speaker 3	572 (335)

3.2. Comparison with Semi-Spontaneous Speech

In the semi-spontaneous speech condition (Task 3) ISP zones were considerably more difficult to identify than in Tasks 1 and 2. Many possible pause sites needed to be considered and frequently rejected, which was time-consuming. Pauses often seemed too short, and were, for example, discarded when no clear zone was discernable between two closely sequenced utterances. Instead of a quasi-stable phase just before the onset of active speech production we often found simple interpolation between gestures of the pause-preceding word and those of the pause-succeeding word. This was reflected in a constant interpolation in the EPG measures (i.e. alveolar contact, total overall contact and centre of gravity).

As Figure 2 shows there was no task that seemed consistently successful in eliciting likely examples of ISPs across all three speakers. There is also no clear trend for either spontaneous or prompted elicitation to work better. As an effect of task order can be excluded, reasons for this variation have to be sought in habitual differences between speakers.

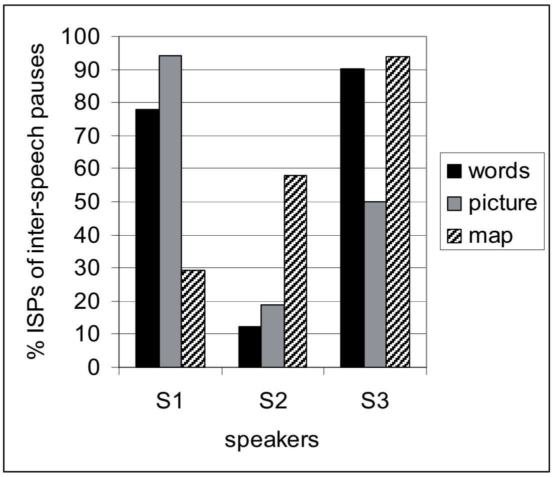

Figure 2. *The proportion of pauses with ISP zones to total number of inter-speech pauses*

4. Discussion

To explore what might have caused the between-speaker variability, we inspected the EPG patterns of the speakers in more detail. We found that Speaker 3, who was least accustomed to wearing an artificial palate, kept his tongue pressed against his palate while waiting for the next prompt and when pausing naturally in Task 3. This made the detection of ISP zones comparatively easy. Speaker 2, on the other hand, often adopted a rest position with no or hardly any tongue-palate contact, making their detection considerably harder. Speaker 1, who is very accustomed to wearing an artificial palate, slipped easily into a rest position between prompts (thereby facilitating ISP zone detection) but did not seem to pause long enough in semi-spontaneous speech to exhibit an equally high percentage of zones in Task 3. We have no explanation as yet as to why Speaker 2 exhibited ISP zones to such a larger extent in the spontaneous compared to the prompted tasks.

It was more difficult to estimate the impact of spontaneous vs. prompted speech on the formation of ISPs, and no clear conclusion could be drawn as to which task most reliably elicits ISPs. However, though ISP zones are observable in spontaneous, discourse-like speech, we think that their analysis yields practical and theoretical problems.

In a prompted speech task paradigm the speaker has no way of knowing what he or she is going to say next; an unprompted speech task will necessarily allow the speaker to plan ahead. We found in our Task 3 data quasi-smooth transitions between audible speech gestures where the speaker seemed to wait for some turn-taking or turn-yielding feedback from the conversational partner [4], but where the speaker had already planned how to continue if this was the negotiated outcome. This suggests that in spontaneous speech tasks it is not just difficult to control segmental context but also to control the pragmatic function of the utterance, which might well have an impact on ISPs and their analysis.

To check that these issues are not specific to EPG, we undertook a pilot inspection of spontaneous dialogue ultrasound data, focusing on movement of the tongue root. We sampled 150 seconds of data from subject MCYF1 (a female SE speaker aged 12) from the ECB08 Corpus [5]. Forty-one tokens of pauses greater than 500 ms were found, of which only 23 were candidates for containing an ISP. Of these, only ten clearly showed an ISP zone, judging by the presence of root movement (anterior in nine cases) from a non-speech position to an observable ISP then reversing towards the first segment. Otherwise, we found interpolation where the speaker appeared to be waiting for feedback (see above). Spatial analysis of these ISPs is currently under way.

5. Summary and Conclusions

We looked at pause length in two prompted speech tasks to establish how long inter-speech pauses should be to give rise to a detectable ISP using EPG. We found that speakers seemed to fall most easily into a rhythm of speaking mode, rest position and in various cases ISPs when we set the pause between prompts to 5 seconds. ISP zones could be found in such data relatively efficiently. Wilson in his study [6] decided on a 1-second pause arguing that speakers exhibited too much non-speech activity (such as bracing) when given a 2-second pause instead. He does not report on trials with longer pauses. We believe that 2 seconds is in fact not long enough for speakers to settle between prompts. Interestingly, and maybe because of the relatively short pauses, speakers in [6] only exhibited ISPs in around 50% of cases, while in our prompted tasks the proportion was 61.4% across all three speakers.

The current study has shown that various factors assist in the elicitation and detection of zones in which ISPs may exist and be measured. Understanding these is crucial in developing a suitable experimental methodology. Overall, inter-speaker variability based on habitual differences has to be considered as a crucial factor for future research. Finally, prompted speech enables more experimental control without, we think, being unrepresentative of natural discourse.

Acknowledgements

We gratefully acknowledge support from the ESRC (RES-000-22-3032 and RES-000-22-2032), as well as the British Academy (SG-48460).

References

[1] Articulate Instruments Ltd. *Articulate Assistant User Guide: Version 1.16*. Edinburgh, UK: Articulate Instruments Ltd., 2007.

[2] B. Gick, I. Wilson, K. Koch & C. Cook (2004). Language-Specific Articulatory Settings: Evidence from Inter-Utterance Rest Position. *Phonetica*, 61: 220-233, 2007.

[3] J. Laver, J. The Concept of Articulatory Settings: A Historical Survey. *Historiographia Linguistica, 5*, 1–14, 1978.

[4] J. Local. Phonetic Detail and the Organisation of Talk-in-Interaction. *Proceedings of the XVIth ICPhS*, 1-10, 2007.

[5] J.M. Scobbie, J. Stuart-Smith, and E. Lawson. *Looking Variation and Change in the Mouth: Developing the Sociolinguistic Potential of Ultrasound Tongue Imaging*. ESRC Final Report (http://www.esrcsocietytoday.ac.uk), 2008.

[6] I. Wilson. *Articulatory Settings of French and English Monolingual and Bilingual Speakers*. PhD dissertation, University of British Columbia, 2006.

[7] S. Wood, W. Hardcastle. Instrumentation in the Assessment and Therapy of Motor Speech Disorders: A Survey of Techniques and Case Studies with EPG, In: I. Papathanasiou (Ed.) Acquired Neurogenic Communication Disorders: a Clinical Perspective. London: Whurr. 2000.