AUDIBLE ASPECTS OF SPEECH PREPARATION

James M Scobbie[a], *Sonja Schaeffler*[a] *& Ineke Mennen*[b]

[a]CASL Research Centre, Queen Margaret University, UK; [b]University of Wales, Bangor, UK

jscobbie@qmu.ac.uk; sschaeffler@qmu.ac.uk; imennen@bangor.ac.uk

ABSTRACT

Noises made before the acoustic onset of speech are typically ignored, yet may reveal aspects of speech production planning and be relevant to discourse turn-taking. We quantify the nature and timing of such noises, using an experimental method designed to elicit naturalistic yet controlled speech initiation data. Speakers listened to speech input, then spoke when prompt material became visible onscreen. They generally inhaled audibly before uttering a short sentence, but not before a single word. In both tasks, articulatory movements caused acoustic spikes due to weak click-like articulatory separations or stronger clicks via an ingressive, lingual airstream. The acoustic onset of the sentences was delayed relative to the words. This does not appear to be planned, but seems a side-effect of the longer duration of inhalation.

Keywords: articulation, speech preparation, clicks, breathing, discourse

1. INTRODUCTION

How do speakers start to speak? In activities like reading aloud, segmental and prosodic speech planning occurs. In spontaneous speech, moreover, speakers plan what to say. In conversation, speakers listen to their interlocutor and dynamically and collaboratively create discourse.

Speaker-generated vocal-tract noises sometimes occur due to non-linguistic activity. Yet breathing, swallowing and other movements, noise-generating or not, may be integrated into speech production, so can give insight into prosodic and segmental speech-motor planning and the time course and nature of its implementation. Noise-making or visible pre-speech activity is, furthermore, relevant to turn-taking because it may function to signal the speaker's intention to speak.

The phonetics of pre-speech has been studied previously in articulatory research. Wilson concludes [7] that the articulatory system can be "speech-ready" in a language-specific sense, or at absolute rest in non-speech postures, or have some intermediate stages of pre-speech activity.

This acoustic paper is part of a bigger *articulatory* study, and three factors of that wider context must be mentioned. First, the study aims to examine the timing of the articulatory motion that occurs for the first segment(s) before the acoustic onset of speech. Secondly, it will explore the lingual postures which speakers adopt before speaking [2, 5, 7]. Thirdly, the speakers are bilinguals, to aid the measurement of what might be language-specific settings [2, 7]. In some of the older, more physiological literature cited in [7], performance of an oral or nasal inhalation was instructed. The task was then often to read sentences aloud [2, 7]. Spontaneous speech has now also been studied [5]. Here we describe a new elicitation paradigm in detail, and present acoustic timing results. We aim to elicit a more natural transition from listening to speaking than simply reading sentences, but still control the segmental content of what is said.

Some of the pre-speech noises studied here are extended frication, caused by breath inspiration through nose or mouth. Some are acoustic spikes, on a continuum between strong stand-alone clicks resulting from an ingressive "lingual" [1] airstream comparable to those found in click languages, and weak "spit-spikes" caused by the rapid mechanical separation of the articulators presumably via small localised pockets of negative pressure. "Weak clicks" arising in German consonant clusters have previously been described by Fuchs, et al. [3], and by Simspon, whom they cite.

2. METHOD

2.1. Speakers and language blocking

The speakers were eight native adult German speakers, all highly fluent in English. Recordings were obtained in the UK, at QMU. The first block of the data collection was in German, facilitated by a native German researcher. Following a break of a couple of minutes, involving some free conversation in English, the English block began. Each block took about 10 minutes to complete. In both blocks, the picture-naming word task preceded the sentence task.

2.2. Materials and their presentation

Single words were elicited via black and white line-drawn picture prompts (targets in Table 1), sentences via black-on-white text. Each word was repeated in a randomised list, giving four tokens.

Table 1: Materials.

English	Mice	Ducks	Fish	House
German	Mais	Dachs	Fisch	Haus
gloss	maize	badger	fish	house
	[mais]	[dʌks]	[fɩʃ]	[haus]

The sentences *began* with one of these same four words (Table 1), but were not identical throughout: there were five variants, and each appeared once. They were all statements, and were similar in prosody and length: a mean 8.2 syllables (s.d. 1.7) in English and 8.5 (s.d. 1.8) in German.

2.3. Prompts

Tongue position and movement during speech preparation was recorded via Articulate Assistant Advanced™. This multichannel system records audio and articulatory channels, and presents timed prompts audibly and visually to the participant.

The speaker had been instructed to either name a picture prompt, or read a prompt sentence, with no specific time pressure, and these were revealed on screen at 2.5 seconds. Pilot work had found that speakers' articulators were un-naturally restless if they were left to sit staring at a blank screen waiting for the prompt to appear, whereas observations of natural discourse had revealed that interlocutors tended to keep their tongue still while they were listening in a real dialogue. Each elicitation therefore began with a range of pseudo-discourse audio pre-prompts being played over headphones to the speaker. A voice (German or English, as appropriate, to enhance the language mode) was heard uttering a randomised list of task-appropriate utterances, like "And the next picture, please", or "And what do you call what's on the next picture?" Thus, during this preliminary non-speaking phase, speakers were treated as if they were listeners in a mini-dialogue. The audio pre-prompt varied in length, but always ended at 1.7 seconds, leaving an 800ms gap between pre-prompt and prompt.

2.4. Annotation criteria

The acoustic onset was marked by hand, and we considered it easy to annotate consistently (Fig. 1). Annotations of pre-speech noise were made at their onset and offset (Fig. 1). We used both waveform and spectrogram to annotate. Labels (Table 2) indicate the acoustic quality and its likely cause.

Figure 1: Example from Speaker S7, "Ducks won't look you in the eye". "Ducks" (A), at 3.815s, starts 150ms after 270ms of pre-speech noise (qi).

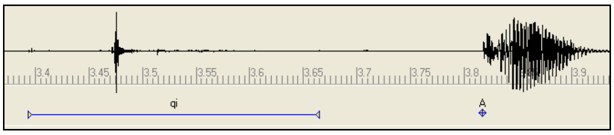


Table 2: Pre-speech noise annotations, purely mechanical lingual airstream (M) or (also) involving pulmonic breath (B).

M	qq	Acoustic spikes only
B	qi	Breath noise following acoustic spikes
	qb	Breath preceding and following spikes
	ib	Breath frication noise only

In /f/ and /h/ the annotation was made at the appearance of broadband spectral friction, but some tokens began with a slow build-up of contiguous frication, so the annotation point was placed relatively early, at its start. In /m/, the annotation was placed at the sudden appearance of voiced energy, except for a few tokens of pre-aspirated /m/ ([ᵐ̥m]), where, again the acoustic onset was placed early. In the case of /d/, onset was marked as the burst of the stop, and not, in the three relevant cases, at a short non-contiguous period of pre-voicing.

All spikes seemed to occur due to opening of the vocal tract, and sounded lingual, labial, or a mixture. Many were acoustically weak and appear to lack any appreciable airstream but some clearly involved a lingual ingressive airstream causing a loud labial, coronal or labial-then-coronal click sequence. No clear demarcation between lingual click and weak mechanical separation types was noted and there was a lot of variation. However, there were some clear patterns due to sequencing. Weak spikes tended to precede louder clicks token-internally, and when these spikes preceded inhalation (e.g. Fig. 1), the breath sounded oral, as could be expected. All in-breaths preceding spikes sounded nasal, presumably because the oral tract was, at that point in time, closed. There was then some further oral or oral-nasal pulmonary inspiration during and/or after the spikes, so these tokens was labeled as one qb event (Table 2).

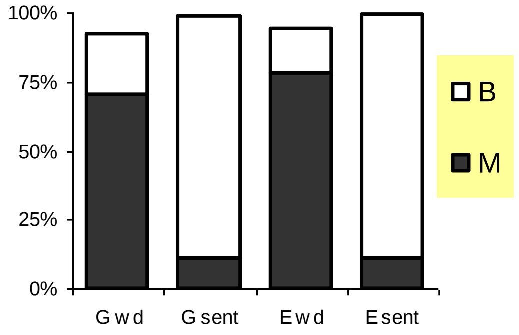
Complete datasets were gathered from six participants. Speaker S4 only participated in the sentence conditions, S5 only in the German condition. 501 of 506 possible tokens were analysable. S1 is missing one English sentence, S2 two German and one English sentence, and one English word.

3. PRE-SPEECH NOISE TYPE

3.1. Group results

Overall, high rates of non-breathing pre-speech noise were found, with almost all tokens containing spikes. In the sentence-reading task, an in-breath noise B was the norm (Fig. 2), with (qi), a spike then oral in-breath, being the most common sub-type, in about ~60% of tokens (cf. Table 3). In the picture-naming task (where speakers knew they would produce just single words) they tended not to take a breath before speaking (Fig. 2). The remainder (8% and 5% in German (G) and English (E) words, and 1% in the sentences) had no audible (or spectrographically visible) pre-speech noise. Of the spikes, around 5% are impressionistically strong lingual airstream clicks.

Figure 2: Mean percentage of tokens preceded by audible spikes only (M) or with a component of audible breath (B: qi, qb & ib).



3.2. Individual results

Speakers varied in their pre-speech noise (Table 3), though the general distinction between single words and sentences is shown by all. Some other results are worth noting. S2 had far less pre-speech noise than the others. S3 was also unusual: B types appeared quite often in the word conditions, compared to other speakers' high rates of M noise.

4. PRE-SPEECH NOISE TIMING

The acoustic onset of speech was over a second (Table 4, n=6, those completing all tasks). There was a trend for slower times in L2 English; and in sentences by about 260ms (Tables 4, 5). The onset of pre-speech noise does not seem task-dependent.

Pre-sentence B were 269ms longer than pre-word M (covering 84% of all tokens) (Table 5). Individual results (Fig. 3) show that the fastest responder (S7) does not have these trends, but it is not clear if this is due to her greater overall speed. For the three slowest responders, English "reaction time" appeared longer than German. Pre-speech, as expected (Table 4), was more consistent (Fig. 4).

S2 had a near-significant difference in word vs. sentence delay, and S1 had such a difference in English (t(30)=3, p=0.005). Pre-speech timing was fairly constant overall because of S7, S8, S3, S6.

Table 3: Noise type (% of tokens) and duration (ms). M has only mechanical spikes; B has additional breathing noise (qi, qb) or breath alone (ib). Blank cells represent zero occurrences, grey rows mean no data was collected.

		M (%)	B (%)			none	ms
		qq	qi	qb	ib		
S1	G wd	94%				6%	72
	G s	20%	40%	30%	10%		259
	E wd	100%					41
	E s	16%	63%	11%	11%		318
S2	G wd	56%	6%			38%	91
	G s		72%	6%	11%	11%	247
	E wd	60%	13%			27%	88
	E s		84%		16%	10%	315
S3	G wd	63%	13%	25%			246
	G s		10%	90%			750
	E wd	44%	25%	31%			259
	E s		40%	60%			550
S4	G wd						
	G s	5%	70%	25%			410
	E wd						
	E s		65%	35%			402
S5	G wd	19%	75%	6%			224
	G s	5%	68%	26%			641
	E wd						
	E s						
S6	G wd	80%	20%				118
	G s	5%	95%				387
	E wd	79%	21%				119
	E s		95%		5%		430
S7	G wd	94%		6%			223
	G s	20%	75%		5%		217
	E wd	100%					92
	E s	35%	65%				218
S8	G wd	88%				12%	89
	G s	35%	25%	35%	5%		257
	E wd	88%		6%		6%	105
	E s	20%	20%	55%		5%	318

Table 4: Mean "reaction time" (ms) from prompt until acoustic onset of speech (top), or pre-speech noise (mid), with their average difference (bottom).

G wd	G sent	E wd	E sent	Sent-wd
1358	1639	1522	1889	324
1072	1103	1230	1327	64
286	535	292	562	260

Table 5: Mean pre-speech durations (ms) and counts.

G wd M	G sent B	E wd M	E sent B	Sent-wd
118	375	77	357	269
n=71	n=100	n=22	n=99	

Figure 3: Mean reaction time, speech (ms).

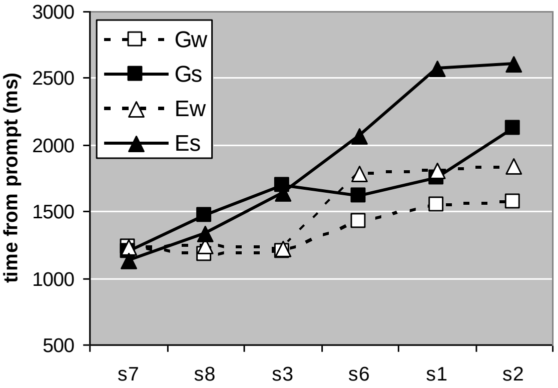
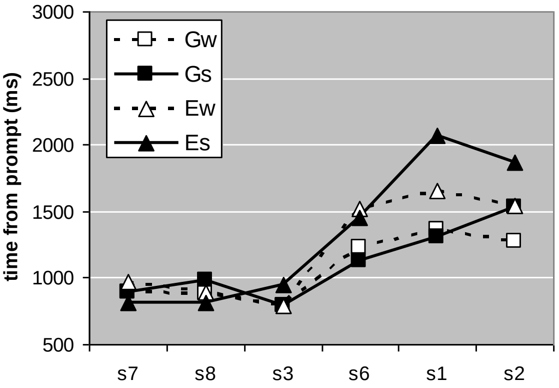


Figure 4: Mean reaction time, pre-speech noise (ms)



5. GENERAL DISCUSSION

Pre-speech noises occurred a quarter to half a second sooner than acoustic lexical content. This pre-speech noise was caused mostly by the articulators pulling away from a contact resting position, lingual clicks, or pulmonic inhalation. The audibility (and visibility) of such movements may signal speech early to an interlocutor, in which case such information could function in discourse to facilitate turn-taking [4, 6, 8], and at least will be influenced by the speaker/listener's discourse planning. Pre-speech is far more variable in spontaneous discourse than our listen-and-respond experiment [5]. There are similarities, however, e.g. in click location [4, 6, 8]. To find out how speech and pre-speech is planned, experimental control of segments, phrase length and speech task are key.

Since inhalation noise lasts longer than purely mechanical spike sequences (Table 5), the trend for sentential delay is probably caused by the inherent durations of different pre-speech behaviours. The sentences begin with the same words used in picture-naming, so segmental planning and execution is unlikely to be the primary cause here. In addition to the actual inhalation time, reading the materials and planning for longer utterances probably both matter. To determine the role of these factors, future work should elicit phrases of different length, ones which start alike lexically, to examine the prosodic effects of utterance length on pre-speech event type, timing and duration. It would be also be interesting to put time-pressure on speakers to make them respond as fast as possible. Clearer L1/L2 differences may emerge, revealing language dominance and ability.

Putative language-specific articulatory settings may be true postural linguistic targets, a neutral underpinning for speech output, detectable in inter-utterance pauses [2, 5, 7]. Other, earlier, pre-speech vocal tract behaviours occur, and are known to be relevant [7]. Listening position (e.g lingual-palatal contact), its release, and inhalation are all important, and their distribution, nature and timing must be taken into account.

6. ACKNOWLEDGEMENTS

Thanks to Jane Stuart-Smith for the idea to make our speakers listen to real speech instead of just sitting, waiting, and two reviewers for comments. Thanks also to Steve Cowen, Xaver Koch, Richard Ogden. Funding from ESRC RES-000-22-3032.

7. REFERENCES

[1] Fuchs, S., Koenig, L., Winkler, R. 2007. Weak clicks in German? *Proc. 16th ICPhS* Saarbrücken, 449-452.

[2] Gick, B., Wilson, I., Koch, K., Cook, C. 2004. Language-specific articulatory settings: Evidence from inter-utterance rest position. *Phonetica* 61, 220-233.

[3] Miller, A., Brugman, J., Sands, B., Namaseb, L., Exter, M., Collins, C. 2009. Differences in airstream and posterior place of articulation among N|uu clicks. *JIPA* 39, 129-161.

[4] Ogden, R. 2006. Clicks in York English. *BAAP Colloquium* QMU, Scotland.

[5] Ramanarayanan, V., Byrd, D., Goldstein, L., Narayanan, A. 2010. Investigating articulatory setting – pauses, ready position and rest – using realtime MRI. *Proc. Intespeech*.

[6] Stuart-Smith, J. 2009. Social distribution of clicks. *Click Workshop* University of York, England.

[7] Wilson, I. 2006. *Articulatory Settings of French and English Monolingual and Bilingual Speakers*. Ph.D. Thesis, University of British Colombia.

[8] Wright, M. In press. On clicks in English conversation. *JIPA*.