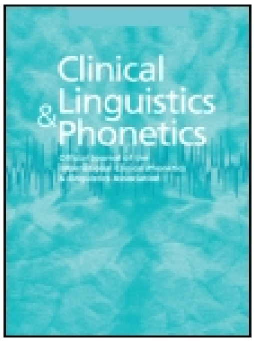

Clinical Linguistics & Phonetics

ISSN: 0269-9206 (Print) 1464-5076 (Online) Journal homepage: https://www.tandfonline.com/loi/iclp20

Tongue reading: Comparing the interpretation of visual information from inside the mouth, from electropalatographic and ultrasound displays of speech sounds

Joanne Cleland, Caitlin Mccron & James M. Scobbie

Clinical Linguistics & Phonetics, April 2013; 27(4): 299–311
© 2013 Informa UK Ltd
ISSN: 0269-9206 print / ISSN 1464-5076 online
DOI: 10.3109/02699206.2012.759626

Tongue reading: Comparing the interpretation of visual information from inside the mouth, from electropalatographic and ultrasound displays of speech sounds

JOANNE CLELAND, CAITLIN MCCRON, & JAMES M. SCOBBIE

Clinical Audiology, Speech and Language Research Centre, Queen Margaret University, Musselburgh, Edinburgh, Scotland

(*Received 14 September 2012; revised 7 December 2012; accepted 12 December 2012*)

Abstract

Speakers possess a natural capacity for lip reading; analogous to this, there may be an intuitive ability to "tongue-read." Although the ability of untrained participants to perceive aspects of the speech signal has been explored for some visual representations of the vocal tract (e.g. talking heads), it is not yet known to what extent there is a natural ability to interpret speech information presented through two clinical phonetic tools: EPG and ultrasound. This study aimed to determine whether there is any intuitive ability to interpret the images produced by these systems, and to determine whether one tool is more conducive to this than the other. Twenty adults viewed real-time and slow motion EPG and ultrasound silent movies of 10 different linguo-palatal consonants and 4 vowels. Participants selected which segment they perceived from four forced-choice options. Overall, participants scored above chance in the EPG and ultrasound conditions, suggesting that these images can be interpreted intuitively to some degree. This was the case for consonants in both the conditions and for vowels in the EPG condition.

Keywords: *electropalatography*, *ultrasound*, *speech reading*, *visual feedback*

Introduction

It is well known that being able to view the face of a speaker enhances the intelligibility of an utterance by the virtue of lip-reading (Benoît & Le Goff, 1998). A small number of studies have looked at whether a natural capacity might also exist for "tongue-reading." The tongue is a major articulator, involved in the production of most English consonants and all vowels, yet it is highly inaccessible and listeners are able to view it only partially at best. Despite this, listeners can copy someone else's speech characteristics from sound alone easily. Acquiring speech from sound exposure is a natural process. Even in the absence of any visual cues (i.e. in blind individuals) speech is acquired, suggesting that learners can acquire articulatory information from the acoustic speech signal alone. According to the Motor Theory of Speech Perception, listeners perceive speech sounds as the "intended phonetic gestures of the speaker" (Liberman & Mattingly, 1985), meaning listeners use articulatory knowledge, albeit at a subconscious level. Evidence for Motor Theory has been

Correspondence: Dr Joanne Cleland, Clinical Audiology, Speech and Language Research Centre, Queen Margaret University, Queen Margaret University Drive, Musselburgh EH21 6UU, Edinburgh, Scotland. Tel: +0131 474 0000. E-mail: jcleland@qmu.ac.uk

mixed; however, the discovery of mirror neurones, or specifically echo neurones, has reignited interest in this theory (Lotto, Hickok, & Holt, 2008). There is a vast literature supporting the view that mirror neurones are responsible for the imitation system, which may be the root of learning, but most of this literature investigate the visual domain in non-speech tasks. Evidence from the studies of primates now exists that auditory perception of a sound (and theoretically a speech sound) is directly related to the action required to make that sound. That is, upon hearing a sound, echo neurones responsible for the action required to generate that sound will fire (Kohler et al., 2002), so in primates the noise of a stick dropping will fire the neurone involved in the actual action.

In primates, the area of the brain containing echo neurones is analogous to Broca's area (Kohler et al., 2002). Although it is not possible to carry out comparable experiments in humans, it is hypothesized that this echo neurone system could be essential in learning to speak. If this is the case, then typical listeners/speakers may have access to the articulatory information involved in speech production and would be able to make use of visual information about normally invisible articulators to enhance perception of speech, and perhaps even to learn new speech sounds. While it is clear that a speech perception/production link must exist, it is far from clear the extent to which, if at all, listeners have access to the articulatory information of speakers. That is, just because a listener understands a speaker, that does not mean the listener has any intuition about what the speaker's tongue (and other parts of the vocal tract) are doing during speech.

A small number of studies have attempted to assess whether listeners have an intuitive tongue-reading ability, using various "Talking Heads." Talking heads are artificial animations of speech production usually based on instrumental data of real speech (often magnetic resonance imaging or electromagnetic articulograph). Some are 3D (e.g. Badin & Serrurier, 2006) and some are 2D (e.g. Kröger, 2003), but most attempt to model the movement of the tongue during speech by providing the user with a cut-away mid-sagittal view of the tongue (as in Figure 1). The main application of Talking Heads is usually as a teaching tool for pronunciation training in second language learning; however, few studies investigate the effectiveness of this.

However, there is increasing evidence that listeners are able to use information about the tongue to enhance the perception of native speech sounds. Badin, Tarabalka, Elisei, and Bailly (2010) investigated the ability of listener–viewers to use a Talking Head to enhance the perception of speech in various noise conditions. The mean phoneme identification rate was significantly greater for all conditions where audio-visual information was added (including a lip-only condition), but more importantly phoneme recognition was significantly greater (68.1%) when a mid-sagittal view with the tongue visible was compared with a mid-sagittal view with no visible tongue (63.7%). Badin et al. (2010) concludes that there is a natural, intuitive, capacity for listeners/viewers to tongue-read and suggest that this provides support for a perception/production link which could relate to the theory of mirror neurones. In a different study, Kröger, Gotto, Albert, and Neuschaefer-Rube (2005) show that children as young as 4;6 with articulation disorders

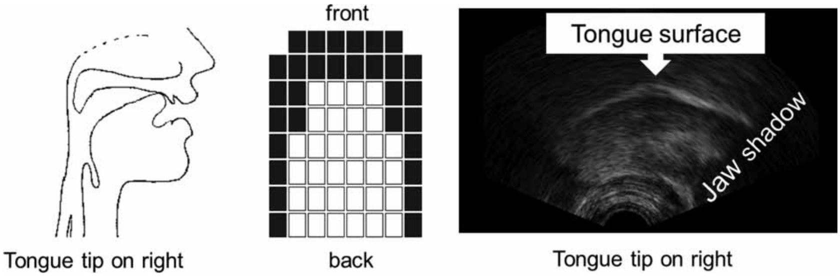

Figure 1. (Left to right) Comparison of typical mid-sagittal diagram, EPG, and ultrasound of [t].

show a similar ability. They tested phoneme recognition of silent animations (based on MRI with information about all articulators), whereby the children were asked to watch the animations and produce what they thought the speech sound was. Responses were rated on a scale for phonological feature correctness. There appears to be a confound here, with children potentially asked to produce sounds that were not in their phonetic inventories (since they had articulation disorders), but this is not explored in the paper. With this particular experimental design, it was impossible to identify cases where a child perceived the speech sound but was unable to indicate this because they were unable to articulate it. Since no child achieved 100%, it is possible that children were unable to correctly produce the speech sounds that were usually in error in their speech.

Nevertheless, these children with articulation disorders do show some ability to tongue-read. It does not, however, necessarily follow that Talking Heads are a useful tool for teaching children the speech sounds they have failed to acquire naturally. Only one study has attempted to investigate this issue. Fagel and Madany (2008) use a Talking Head to treat interdentalized /s/ and /z/ in German children. Six out of eight children lessened their degree of lisping after just one learning session but the authors were unable to demonstrate that the improvement was a direct result of the Talking Head intervention. Most speech therapists, especially in the UK, will question the necessity of a mid-sagittal Talking Head for remediation of interdental sibilants since the incorrect production would be easily viewed on the face of another speaker or in a mirror if visual feedback is required. Moreover, dentalizations are normally considered a minor distortion (Shriberg, Austin, Lewis, & McSweeny, 1997) and at least in the UK are not usually considered for treatment.

In clinical phonetics, researchers and clinicians will be more familiar with instrumental techniques that provide visual feedback of the speakers' own articulations. Electropalatography (EPG), for example, is a technique for displaying the timing and location of tongue–palate contact (Hardcastle & Gibbon, 1997). The speaker with disordered speech sees an abstract representation of linguo-palatal consonants (and some information for high vowels) in real time and is encouraged to use this to modify their own erroneous articulations. Clinically, this computer-based therapy tool has been used widely to provide visual feedback to remediate speech sound disorders (Bernhardt, Gick, Bacsfalvi, & Ashdown, 2003) with positive results reported in a large number of case and small group studies. The understanding of this visual display is described as relatively intuitive (Gibbon & Wood, 2010), even for those with cognitive impairment (Cleland, Timmins, Wood, Hardcastle, & Wishart, 2009). Although its therapeutic success has been reported, there has been little exploration of why it might be useful for the speaker to view their own articulation and precisely how the presence of a real-time visual image of tongue–palate contact is able to help after disordered productions become habitualized.

There is no existing evidence supporting the hypothesis that there may be a natural capacity to interpret, or tongue-read, EPG. Clinical application of EPG usually follows training and demonstration, in conjunction with instruction-based direct therapy provided by a specialist Speech and Language Therapist (Gibbon & Wood, 2010), and it is entirely possible that the power of EPG lies more in its diagnostic value (since it can be used to create a fine grained analysis of speech and to suggest underlying causes of speech disorders) than exploitation of the putative mirror neurone system whereby actually showing the child a correct articulation would lead to improvement. Even more likely is that a combination of the above factors is at play, with improved accuracy of diagnosis leading to more efficacious therapy, which may or may not exploit the mirror neurone system, but certainly gives the patient access to an additional visual modality for learning new speech sounds.

EPG differs from Talking Heads in two important ways; firstly, the display is an abstract representation of one aspect of speech production, rather than a more anatomically correct representation of a speaker's mouth. Secondly, it is almost exclusively used as a real-time feedback tool by an SLT, providing a dynamic representation of the speaker's own speech production, rather than a model alone.

No studies report on whether there is a capacity to tongue-read from pre-recorded EPG. If there is a capacity to tongue-read from EPG, then it might be used like a Talking Head, avoiding the need to make expensive dental plates for each speaker. Alternatively, if the feedback of the speaker's own speech is required, then the ability to tongue-read might speed up the therapeutic process, or make indirect therapy (where the child uses the EPG equipment exclusively at home) a viable approach. Furthermore, Badin et al. (2010) report that some participants in their study were "good tongue-readers," whereas others were "poor tongue-readers." Whilst it is possible that poor tongue-readers might be receptive to training, if this is not the case then having such an evaluation before offering EPG therapy might be a useful way of screening out those who are not likely to benefit. Moreover, it might give some clues as to why not all children have benefited from EPG in the past.

Another visual feedback technique that is gaining popularity is ultrasound tongue imaging (UTI). With this technique, most of the surface of the tongue can be made visible in a mid-sagittal view in real time. This view can be used for visual feedback of the tongue and interpreting such images is thought to be relatively "intuitive" (Bernhardt, Gick, Bacsfalvi, & Adler-Bock, 2005). However, all studies using UTI involve multiple intervention sessions with a specially trained therapist, again making this assertion unfounded. Unlike EPG, the image is an anatomically correct representation of a slice of the tongue, as in a Talking Head. However, other relevant anatomical information, such as the lateral margins of the tongue in the sagittal plane and the relation of the tongue to the hard palate, are not visible in UTI. Also, during speech the tip of the tongue may be in shadow from the speaker's jaw or invisible due to a sublingual airpocket. However, since the tongue itself is imaged, rather than tongue–palate contact, ultrasound shows fuller information for a variety of segments, especially perhaps for vowels. The viewer can see the shape and location of the tongue change from one sound to another, based on ultrasonic echoes from structures within the tongue, and, more obviously, from the tongue's surface. Figure 1 (above) compares ultrasound and EPG with a typical mid-sagittal diagram of [t].

Ultrasound might have an application as a Talking-Head-like model, but this has not been investigated. Models derived from ultrasound might have an advantage over models derived from EMA or MRI since data can be acquired quickly and easily at a high sample rate (Wrench & Scobbie, 2011) and from multiple speakers. The suitability of this technique for capturing the articulation of children is particularly useful since at present Talking Heads are based on adult speech. A model based on ultrasound of children's speech might be more realistic, especially since children's speech is likely to differ from that of adults, while children are a key target group for speech therapy.

Aims

As a first step to determine whether tongue-reading can be observed with EPG and UTI, we sought to determine whether naïve adults, without disorders of speech, can identify a single segment from silent videos of EPG and/or ultrasound. The research questions were:

1. Are participants able to tongue-read (a) EPG and (b) ultrasound displays above the level of chance?
2. Is the effect, if any, stronger with either technique?
3. Are vowels or consonants easier to tongue-read in (a) EPG and (b) ultrasound displays?

We predicted that, like the studies of tongue-reading with Talking Heads, there would be some capacity to identify segments above chance level in both the techniques. Furthermore, we hypothesized that the percentage of lingual–palatal consonants correctly identified will be higher for EPG, whereas vowels will be better discriminated in the ultrasound condition.

Method

Participants

Ten male and ten female speakers of Standard Scottish English (see Scobbie, Gordeeva, & Matthews, 2007, for a description of this variety of English) aged 20–22 ($M = 21$, SD = 0.71) were recruited. Participants were born and raised in Scotland and were currently attending Scottish universities in the central belt. Participants were excluded if they had any disorders of speech, or any related disorders, such as dyslexia. Participants were final year undergraduate students and none had previous experience viewing EPG or lingual ultrasound displays. None were students of linguistics or Speech and Language Therapy. Seven were students of professions allied to medicine; seven were students of the arts such as English and journalism, four were students of engineering or technical subjects and two were training to become school teachers.

Stimuli

Simultaneous EPG and ultrasound video recordings of a female Scottish speaker were made using Articulate Assistant Advanced™ (Articulate Instruments Ltd, 2011). Ultrasound recordings were of the mid-sagittal view only. Whilst it is also possible to use ultrasound to visualize the coronal view, it was not possible to record simultaneous EPG, mid-sagittal and coronal ultrasound. Moreover, it was desirable to have one view only for each condition. Each target segment was recorded three times. Consonants were placed between open vowels to highlight the lingual gestures required. Vowels were prolonged. Table 1 shows the stimuli.

The consonants chosen were all present in, or specific to, Scottish English and allowed for a variety of place and manner of articulation. Only consonants typically classified as lingual were selected, as these can be imaged using either EPG or ultrasound. Voicing was not assessed, as this cannot be observed using these visual feedback tools. It was not expected that participants would be able to intuitively distinguish between consonants sharing the same main place of articulation, for example, [t] and [n]; these were, therefore, not offered alongside each other in the forced-choice task (see below). It was anticipated that participants would, however, observe a difference between consonants such as [t] and [tʃ], as dynamic information was available. Four Scottish vowels were chosen representing a range of tongue height and position.

Training materials

To avoid the need for a large number of practice items, each participant was orientated to both the EPG and ultrasound displays using a scripted presentation with silent videos of practice

Table 1. Test segments.

Consonants	Vowels
[ata]	[iː]
[ana]	[ɛː]
[asa]	[aː]
[aʃa]	[ɔː]
[atʃa]	
[aɹa]	
[aça]	
[aja]	
[aka]	
[axa]	

segments. A tutorial was designed to briefly describe both EPG and ultrasound, demonstrating how to read each display without revealing any specific information regarding the test segments. As [l] and [ŋ] were used as examples of "front" and "back" sounds, these segments were therefore not used in the main experiment.

Procedure

Each of the 14 test sounds were assessed four times resulting in 56 tokens per condition (EPG or UTI), 112 in total. The order of the tokens was randomized within each condition.

Following the tutorial participants individually viewed silent movies of each condition. The order of presentation (EPG or UTI first) was counterbalanced. Participants first viewed the test item in real time and then in slow motion (25% real time) and identified the segment from a four-option forced choice. Segments were presented in the context of words that demonstrated their usual phonetic realization to avoid confusion from orthography. To illustrate, [ç] was presented as "huge" and [x] as "loch." Example words were taken from Hewlett and Beck (2006, p. 48). For each consonant, forced-choice options consisted of the correct target segment and three distractor segments, one differing in place, one in manner (and perhaps place) and one non-lingual consonant. There was, therefore, only one possible correct answer for each test item. Since the four test vowels differed in tongue height and position, they were all available for selection in the forced-choice options. Figure 2 shows an example test item. On completion of both the conditions, participants were asked which style of visual feedback they preferred and why and gave a prediction as to which condition they performed better in. The full procedure took approximately 60 min.

Analysis

Two levels of analysis were carried out, broadly similar to Kröger et al. (2005). Firstly, a percentage segment correct criterion was applied. Since the forced choice was carefully designed, it was

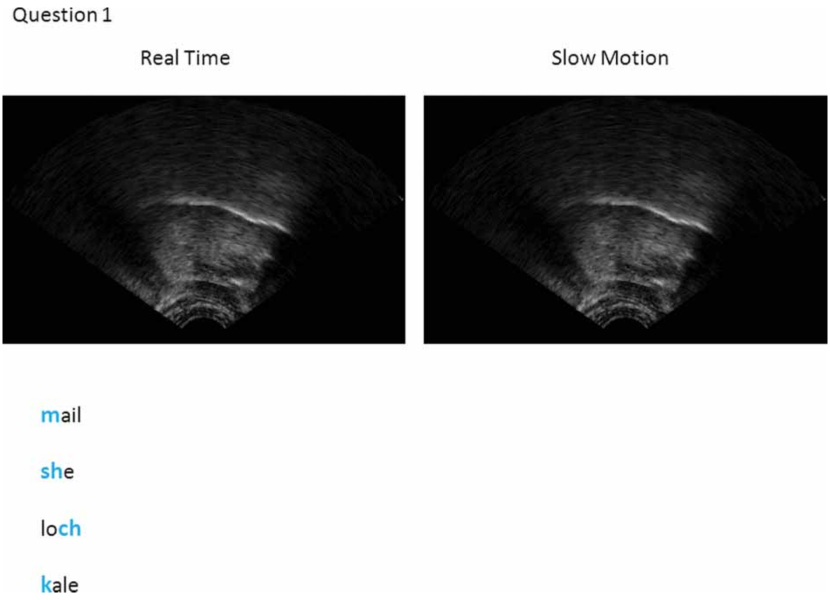

Figure 2. Example test item from the ultrasound condition. Videos were clickable.

possible for participants to score 100%. Secondly, a place, voice and manner feature analysis was applied on a four-point scale. Correct selections received a score of 3. The selection of segments produced at the same place of articulation, for example, an alveolar plosive for an alveolar fricative, received a score of 2. The selections of segments produced in an adjacent place of articulation, for example, a post-alveolar fricative rather than an alveolar fricative, received a score of 1. Since in UTI the hard palate is not visible (unless the speaker is swallowing or similar), it may be difficult to discern for example the difference between [s] and [ʃ]. All other selections received a score of 0. This four-point scoring method was not used for the vowel tokens, as they varied noticeably in tongue height and position. Percentage consonants correct results were compared with the level of chance (25%), as in Kröger, Graf-Borttscheller, and Lowit's (2008) study of the natural ability to interpret 2D and 3D models.

Results

Participants were able to identify which segment had been uttered from a silent movie of EPG 52% of the time (SD = 15.75) and from a silent movie of UTI 41% of the time (SD = 11.28). Like Badin et al. (2010), there were both good and poor tongue-readers with scores ranging from 18% to 82% for EPG and 23% to 61% for ultrasound. The majority of participants scored significantly above chance (Chi-square: EPG condition, χ^2 (19, $n = 20$) $= 403.24, p \leq 0.001$; ultrasound condition, χ^2 (19, $n = 20$) $= 144.17, p \leq 0.001$), showing that most people display a natural capacity to tongue-read from these techniques. Figure 3 shows the individual results for participants, with chance level (25%) indicated.

Overall, there was a highly significant difference between correct answers achieved in the EPG and ultrasound conditions, $p \leq 0.001$, suggesting that EPG is more conducive to tongue-reading for the segments tested here. There was a correlation evident ($r = 0.55, p = 0.01$) between performance in the EPG and UTI conditions.

Comparing consonants and vowels

Figure 4 shows the group results for consonants (a) and vowels (b) in each condition. Percentage consonants correct was 55% in the EPG condition (SD = 18.52) and 46% in the ultrasound condition (SD = 13.30). The large standard deviations (especially in the EPG condition) reflect the

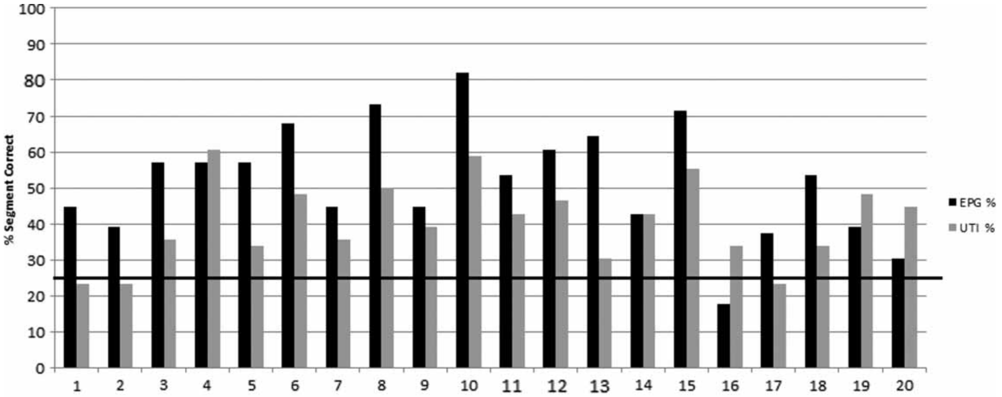

Figure 3. Individual results with chance (25%) represented by a line.

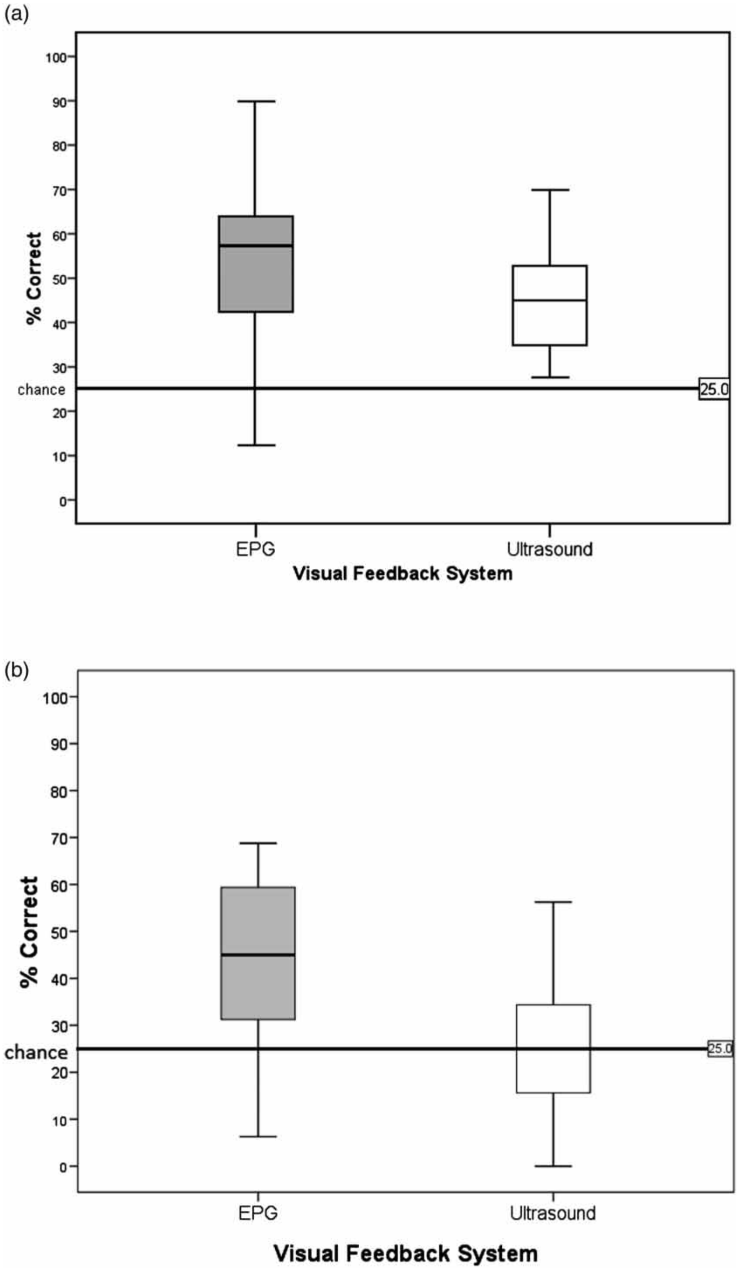

Figure 4. (a) Group results of consonants in EPG and ultrasound and (b) group results of vowels in EPG and ultrasound.

heterogeneity in individual performance. Again, in both the conditions performance was above chance (Chi-square: EPG condition, χ^2 (19, $n = 20$) $= 387.21$, $p \leq 0.001$; ultrasound condition, χ^2 (19, $n = 20$) $= 192.67$, $p \leq 0.001$). Consonants were more easily tongue-read with EPG than ultrasound ($p \leq 0.001$).

For vowels, correct identification was lower: 44% for EPG (SD = 18.65) and only 26% for ultrasound (SD = 14.98). This was found to be at chance level for the ultrasound condition

Table 2. Comparison of two-level scoring.

EPG mean % correct		UTI mean % correct	
Strict correct segment	4-point scale	Strict correct segment	4-point scale
51.96	58.07	40.54	46.31

(2 (19, $n = 20$) = 0.27, $p > 0.20$). Surprisingly, this was not the case for the EPG condition, where the percentage of correct vowels was significantly above chance level, 2(19, $n = 20$) = 62.02, $p \leq 0.001$. Contrary to our hypothesis, vowels were more easily tongue-read with EPG than ultrasound ($p \leq 0.001$).

As the number of vowels and consonants tested were not equal, it is difficult to assess the difference in performance between these. However, this is assumed to be significant in the ultrasound condition, as the identification of consonants above chance level was found to be highly significant whilst vowels were not. Participants appear to be more successful in tongue-reading consonants than vowels in both conditions.

Feature analysis

Table 2 compares the scores from the strict scoring criteria with the place, voice and manner feature analysis. As expected, when a feature analysis was applied rather than strict right/wrong criteria the % correct increased, suggesting that participants sometimes made errors involving the same or adjacent place of articulation. As expected, the scores obtained with both methods were very strongly correlated: EPG condition, $r = 0.98$, $p \leq 0.001$, ultrasound condition, $r = 0.99$, $p \leq 0.001$.

Participants' preferences

In a qualitative debrief, 60% of female participants reported a preference for EPG whereas 80% of males specified that they preferred EPG. Therefore, overall, 70% (14) of participants preferred the EPG display, whereas the remaining 30% (6) reported that they found ultrasound easier to understand. There was no correlation between which condition the participants were exposed to first and their stated preference, with identical proportions in each (70% of those who viewed EPG first preferred it as did 70% of those who viewed it after the ultrasound condition). It is interesting to note that participants' preferences did not always correspond to the condition they were most successful in. Participants 4 and 16 preferred EPG, however, they were more successful in the ultrasound condition. Participants 2, 9, 12 and 17 all preferred ultrasound yet performed better in the EPG condition (see Figure 3 above).

Discussion

Previous research has found some natural capacity to tongue-read from mid-sagittal animations of the vocal tract, despite the fact that speakers will have little or no opportunity to observe such tongue motions naturally. Our experiment extends this to instrumental methods commonly used in phonetic research and speech and language therapy.

Overall, consonants were easier to tongue-read than vowels, supporting the view of Speech and Language Therapists that EPG is most useful for remediation of consonant errors (Gibbon & Paterson, 2006). It was surprising that participants performed at chance level in the ultrasound vowel condition since previous research has highlighted the value of this visual feedback tool in the treatment of vowels due to the anatomically correct visualization of the configuration and position of the tongue (Bernhardt et al., 2005). Difficulty tongue-reading from vowels may be due to a speaker's lack of awareness of their own tongue during vowel production or because vowel quality is highly dependent on the shape and width of the whole vocal tract, not just tongue location and shape. Clinically, this might suggest that remediation of vowel disorders with ultrasound may be highly dependent on training from a specialist speech and language therapist. Alternatively, it may be that the client groups involved in the studies of vowel remediation have somehow superior ability for interpreting vowels. This might be the case in the hearing impaired population where visual skills are often strong. Theoretically, it is difficult to reconcile why participants had such difficulty with vowels if we subscribe to Motor Theory and/or Mirror Neurone Theory, especially since most speakers acquire vowels easily and early in development. Other researchers have also suggested that mirror neurones do not play as central a role in speech as first hypothesized. Motor Theory would predict that since there is a direct link between perception and production, damage to Broca's area (if it contains echo neurones in humans) would result in parallel difficulties in speech production and perception. Studies of people with lesions in this area do not support this (Lotto et al., 2008). If mirror neurones are not at play, then perhaps our participants were using a much more conscious strategy to complete the tongue-reading task, perhaps watching the silent movie, then silently articulating each of the forced-choice options to find a match. It would be interesting to investigate this using a paradigm, where the participants were recorded using ultrasound or EPG while they undertook the perception task.

Despite ultrasound showing an anatomically correct representation of the central tongue slice similar to Talking Heads, and despite EPG using an abstract representation, participants were more successful at tongue-reading from EPG. It is known that speakers make use of tactile feedback provided by tongue–palate contact in order to detect lingual position and movement in consonant production (Hewlett & Beck, 2006). It may, therefore, be the case that participants had more success in intuitively reading these silent videos, as EPG provides a visual representation of a tactile event, tongue–palate contact. The same might be true for vowels, with high vowels being much easier to interpret with EPG since tongue–palate contact, and hence tactile feedback, is available. Moreover, the EPG display is normalized across speakers, perhaps making it easier to tongue-read when, as in this experiment, viewing the tongue movements of an unfamiliar speaker. In contrast, ultrasound is individualized for each speaker; therefore, an experiment which tests how well a speaker can interpret their own pre-recorded ultrasound tongue movements may have been more successful.

We asked participants which instrumental method they preferred and most (70%) had a preference for EPG. Qualitatively, participants commented on the benefit of the precise contact points and enjoyed the layout of the EPG display. Some said they found the mid-sagittal view provided by the ultrasound display confusing. They reported that they could ascertain which patterns would be produced by each sound and found it easier to locate the place of articulation using EPG. Again, these comments support the idea that some participants may have been using a strategy to complete the task, rather than unconsciously making use of a mirror neurone system. Participants also appreciated EPG's use of color, despite this being arbitrary. Even those that reportedly preferred ultrasound often described this feedback tool as "unclear" or "fuzzy." However, participants did not mention any negative impact caused by the shadowed tip of the tongue in ultrasound. Those who preferred ultrasound reported that they benefited from viewing an "actual tongue." These participants felt that this made it easier to appreciate the range and duration of movement.

Since EPG has a clear advantage over ultrasound, it may have some potential as a Talking-Head-like model where pre-recorded EPG is used to demonstrate speech sounds to either second language learners or people with speech sound disorders. The possible advantage of this over existing Talking Heads is that normative data exist for a small number of children (Timmins, Hardcastle, Wood, & Cleland, 2011) and many more adults (e.g. McLeod & Roberts, 2005). It is also relatively straightforward to average data across speakers, since the EPG display is already normalized. However, if visual feedback is required then ultrasound should still be considered as it is cheaper and more flexible than EPG, since speakers do not require a custom-made artificial palate. Moreover, a current research project, Ultrax (2011), aims to make the ultrasound image more accessible by adding anatomical information, essentially making it more like a Talking Head and also allowing speakers to view tongue–palate contact. It is possible that this would enhance the tongue-reading potential of ultrasound.

Is tongue-reading essential for visual feedback success?

Although tongue-reading appears to be possible with both Talking Heads and instrumental phonetic techniques, it is unclear whether it is a necessary step in using EPG or ultrasound for visual feedback. With both these techniques, therapy will usually involve either demonstration of the speech sound to be taught by the Speech and Language Therapist and/or drawing the speaker's attention to a static target pattern. However, therapy mostly focuses on directed feedback, with the therapist acting as a crucial mediator, interpreting articulatory information and then instructing the speaker how to move their tongue in order to achieve the correct articulation. It seems likely that a combination of the visual biofeedback coupled with teaching by a therapist with expert phonetic knowledge leads to the therapeutic success of these techniques. It is, therefore, possible that tongue-reading by clients may not be essential for these techniques to be useful, suggesting that even those who are "poor tongue-readers" (Badin et al., 2010), performing at chance, will still benefit from visual feedback therapy.

Studies that investigate the use of an articulatory model only to teach new speech sounds are few. Massaro, Bigler, Chen, Perlman, and Ouni (2008) used a Talking Head to teach native English speakers a new vowel [y] and a new consonant [q]. While the view of the lips was successful for teaching the high-front rounded vowel [y], learners who had access to a mid-sagittal Talking Head for learning a contrast between /k/ and a uvular stop [q] had no advantage over those who used audio only. Similarly, the study by Fagel and Madany (2008) that used a Talking Head to teach [s] and [z] to children with interdental lisps was unable to show an effect. In their experiment at least it seemed a visual model was not sufficient for the success. However, since the above studies did not give the learners any feedback (e.g. a Speech and Language Therapist telling the learner how close their production was to the target), a further study is required that compares an articulatory model with a visual biofeedback system using both the same type of display and with the same amount of support from a Speech and Language Therapist.

Summary and conclusions

This study sought to establish whether naïve participants are able to determine which speech sound is being produced in silent videos of the dynamic aspects of speech production, using EPG and ultrasound. Most participants performed above chance, confirming some capacity for tongue-reading. It is still unknown how participants completed the task. How did they know which speech sound the speaker was making? While there is most certainly some kind of perception–production link, our experiment does not offer explicit evidence for Motor Theory or for mirror neurones, since it was probably possible to complete the task offline by silently articulating each of the forced choice answers to find a plausible match for the articulation shown in the silent movie. This would also account for the fact that no participant achieved a ceiling score and some achieved a floor score, despite no history of any difficulty in learning to speak.

In sum, our findings support the notion that EPG and ultrasound are relatively intuitive techniques (Bernhardt et al., 2005; Gibbon & Wood, 2010). Both techniques seem suitable for indirect therapy, since little training in interpreting the images would be required prior to devising home programmes. Thus, rather than the SLT being present for every intervention session the client could work on their own, either at home (Portable training units are available for EPG) or unsupervised in the clinic, if given a structured programme. The ability to tongue-read from EPG and ultrasound varied hugely among participants with some individuals performing at chance level. However, most participants were able to tongue-read, perhaps giving some clues as to the mechanisms that underlie the success of EPG and ultrasound as therapeutic tools. For those who are "poor tongue-readers" extra training by an SLT prior to commencing therapy may be required. The leap between tongue-reading native phonemes and using the displays to learn speech sounds which are not in the speaker's phonetic inventories still need further investigation. Moreover, tongue-reading in different populations, such as developmental or acquired speech sound disorders, and second language learners require further investigation as does the contribution of the therapist to the process in terms of how much direct training is required for successful visual biofeedback therapy.

Acknowledgments

We thank all the participants who gave up their time to complete the experiment and Steve Cowen for technical assistance.

***Declaration of Interest*:** The first author is supported by an EPSRC grant (Grant No.: EP/I027696/1). The authors report no conflict of interest.

References

Articulate Instruments Ltd. (2011). *Articulate Assistant Advanced User Guide: Version 2.13*. Edinburgh: Articulate Instruments Ltd.

Badin, P., & Serrurier, A. (2006). Three-dimensional linear modelling of tongue: Articulatory data and models. In H. C. Yehia, D. Demolin, & R. Laboissière (Eds.), *Seventh International Seminar on Speech Production, ISSP7* (pp. 395–402). Ubatuba, SP, Brazil: UFMG, Belo Horizonte, Brazil.

Badin, P., Tarabalka, Y., Elisei, F., & Bailly, G. (2010). Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding. *Speech Communication*, *52*, 493–503.

Benoît, C., & Le Goff, B. (1998). Audio-visual speech synthesis from French test: Eight years of models, designs and evaluation at the ICP. *Speech Communication*, *26*, 117–129.

Bernhardt, B., Gick, B., Bacsfalvi, P., & Adler-Bock, M. (2005). Ultrasound in speech therapy with adolescents and adults. *Clinical Linguistics and Phonetics*, *19*, 605–617.

Bernhardt, B., Gick, B., Bacsfalvi, P., & Ashdown, J. (2003). Speech habilitation of hard of hearing adolescents using electropalatography and ultrasound as evaluated by trained listeners. *Clinical Linguistics and Phonetics*, *17*, 199–216.

Cleland, J., Timmins, C., Wood, S. E., Hardcastle, W. J., & Wishart, J. G. (2009). Electropalatographic therapy for children and young people with Down's syndrome. *Clinical Linguistics and Phonetics*, *23*, 926–939.

Fagel, S., & Madany, K. (2008, September 22–26). A 3-D virtual head as a tool for speech therapy for children. *Paper presented at Interspeech 2008*.

Gibbon, F., & Paterson, L. (2006). A survey of speech and language therapists' views on electropalatography therapy outcomes in Scotland. *Child Language Teaching and Therapy*, *23*, 275–292.

Gibbon, F. E., & Wood, S. E. (2010). Visual feedback therapy with electropalatography. In A. L. Williams, S. McLeod, & R. J. McCauley (Eds.). *Interventions for speech sound disorders in children* (pp. 509–532). Baltimore: Paul H. Brookes Pub.

Hardcastle, W., & Gibbon, F. (1997). Electropalatography and its clinical applications. In M. Ball and C. Code (Eds.), *Instrumental clinical phonetics* (pp. 149–193). London: Whurr.

Hewlett, N., & Beck, J. (2006). *An introduction to the science of phonetics.* London: Lawrence Erlbaum.

Kohler, E., Keysers, C., Umilta, A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science*, *297*, 846–848.

Kröger, B. (2003). Ein visuelles Modell der Artikulation. *Laryngo-Rhino-Otologie*, *82*, 402–407.

Kröger, B., Gotto, J., Albert, S., & Neuschaefer-Rube, C. (2005). A visual articulatory model and its application to therapy of speech disorders: A pilot study. *ZAS Papers in Linguistics*, *40*, 79–94.

Kröger, B., Graf-Borttscheller, V., & Lowit, A. (2008, September 22–26). Two- and three-dimensional visual articulatory models for pronunciation training and for treatment of speech disorders. *Paper presented at Interspeech 2008*.

Liberman, A., & Mattingly, I. (1985). The motor theory of speech perception revised. *Cognition*, *21*, 1–36.

Lotto, A., Hickok, G., & Holt, L. (2008). Reflections on mirror neurones and speech perception. *Trends in Cognitive Sciences*, *13*, 110–114.

Massaro, D., Bigler, S., Chen, T., Perlman, M., & Ouni, S. (2008, September 22–26). Pronunciation training: The role of ear and eye. *Paper presented at Interspeech 2008*.

McLeod, S., & Roberts, A. (2005). Templates of tongue/palate contact for speech sound intervention. In C. Heine & L. Brown (Eds.), *Proceedings of the 2005 Speech Pathology Australia National Conference* (pp. 104–112). Melbourne: Speech Pathology Australia.

Scobbie, J. M., Gordeeva, O. B., & Matthews, B. (2007). Scottish English speech acquisition. In S. McLeod (Ed.), *The international guide to speech acquisition* (pp. 221–240). Clifton Park, NY: Thomson Delmar Learning.

Shriberg, L., Austin, D., Lewis, B., & McSweeny, J. (1997). The Speech Disorders Classification System (SDCS): Extensions and lifespan reference data. *Journal of Speech, Language, and Hearing Research*, *40*, 723–740.

Timmins, C., Hardcastle, W. J., Wood, S., & Cleland, J. (2011). An EPG analysis of /t/ in young people with Down's syndrome. *Clinical Linguistics and Phonetics*, *25*, 1022–1027.

Ultrax. (2011). Overview of research for Ultrax [online]. Available at: http://www.ultrax-speech.org/research [Accessed November 17 2011].

Wrench, A., & Scobbie, J. M. (2011). Very high frame rate ultrasound tongue imaging. *Proceedings of the 9th International Seminar on Speech Production* (pp. 155–162), Montreal, Canada.