Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 75 (2015) 1–13

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Examining visible articulatory features in clear and plain speech

Lisa Y.W. Tang [a], Beverly Hannah [b], Allard Jongman [c,*], Joan Sereno [c], Yue Wang [b], Ghassan Hamarneh [a]

[a] *Medical Image Analysis Lab, School of Computing Science, Simon Fraser University, 9400 TASC1, 8888 University Drive, Burnaby, BC V5A 1S6, Canada*

[b] *Language and Brain Lab, Department of Linguistics, Simon Fraser University, 6203 Robert C. Brown Hall, 8888 University Drive, Burnaby, BC V5A 1S6, Canada*

[c] *KU Phonetics and Psycholinguistics Lab, Department of Linguistics, University of Kansas, 15 Blake Hall, Lawrence, KS 66045-3129, USA*

Received 11 May 2015; received in revised form 13 September 2015; accepted 16 September 2015
Available online 28 September 2015

Abstract

This study investigated the relationship between clearly produced and plain citation form speech styles and motion of visible articulators. Using state-of-the-art computer-vision and image processing techniques, we examined both front and side view videos of speakers' faces while they recited six English words (keyed, kid, cod, cud, cooed, could) containing various vowels differing in visible articulatory features (e.g., lip spreading, lip rounding, jaw displacement), and extracted measurements corresponding to the lip and jaw movements. We compared these measurements in clear and plain speech produced by 18 native English speakers. Based on statistical analyses, we found significant effects of speech style as well as speaker gender and saliency of visual speech cues. Compared to plain speech, we found in clear speech longer duration, greater vertical lip stretch and jaw displacement across vowels, greater horizontal lip stretch for front unrounded vowels, and greater degree of lip rounding and protrusion for rounded vowels. Additionally, greater plain-to-clear speech modifications were found for male speakers than female speakers. These articulatory movement data demonstrate that speakers modify their speech productions in response to communicative needs in different speech contexts. These results also establish the feasibility of utilizing novel computerized facial detection techniques to measure articulatory movements.

Keywords: Articulation; Clear speech; English vowels; Computational methods; Facial landmark detection

1. Introduction

Previous research has established that the movements of facial articulatory features contribute to the myriad of cues generated during speech (Gagné et al., 2002; Mixdorff et al., 2005; Smith and Burnham, 2012; Tasko and Greilick, 2010). The current study explores how visual cues generated by the visible articulatory movements of the lips and facial muscles are deployed by speakers during production of different speech styles (clearly produced and plain citation form), by utilizing novel computerized facial detection techniques to measure differences in articulatory movements during clear versus plain speech tokens of English tense and lax vowels embedded in /kVd/ contexts.

1.1. Audio-visual speech perception

Research has demonstrated that bimodal (auditory and visual, AV) perception is superior to auditory-only (AO) perception of speech (Massaro, 1987; Sumby and Pollack,

* Corresponding author at: 422 Blake Hall, Lawrence, KS 66045-3129, USA. Tel.: +1 785 864 2384.
E-mail addresses: lisat@sfu.ca (L.Y.W. Tang), beverly_hannah@sfu.ca (B. Hannah), jongman@ku.edu (A. Jongman), sereno@ku.edu (J. Sereno), yuew@sfu.ca (Y. Wang), hamarneh@sfu.ca (G. Hamarneh).

http://dx.doi.org/10.1016/j.specom.2015.09.008
0167-6393/

1954; Summerfield, 1979, 1992). This is presumably due to the additional stream of linguistic information available to the perceiver in the visible articulatory movements of the speaker's lips, teeth, and tongue as useful sources for segmental perception (Kim and Davis, 2014b; Tasko and Greilick, 2010; Traunmüller and Öhrström, 2007). Additionally, visual cues from movements of facial features including the eyebrows, neck, and head may contribute to the perception of prosodic information such as lexical tone, stress, and focus (Chen and Massaro, 2008; Cvejic et al., 2012; Krahmer and Swerts, 2007; Smith and Burnham, 2012).

Further findings reveal that the weight granted to visual cues depends on the relative availability and accessibility of the visual (relative to auditory) information, which is affected by factors such as the visual saliency of articulatory input, the quality of auditory input, and the condition of perceivers. For example, perceivers are found to put more weight on the visual input for rounded vowels than for open vowels, as lip-rounding is more visually salient to uniquely characterize rounded segments than the generic mouth opening gesture (Traunmüller and Öhrström, 2007). Likewise, perceivers are more accurate in identifying speech contrasts with more visible articulatory gestures (e.g., labial/labio-dental /p-f/) compared to those with less visible ones (e.g., alveolar /l-ɹ/) (Hazan et al., 2006). Moreover, research has shown that visual information enhances speech perception when auditory environment is degraded, such as in a noisy environment (Bernstein et al., 2004; Hazan et al., 2010; Sumby and Pollack, 1954; Summerfield, 1979). Visual input has been found to particularly benefit special populations for whom the auditory speech distinctiveness is challenging or unfamiliar, such as hearing-impaired or non-native perceivers (Grant and Seitz, 1998; Sekiyama and Tohkura, 1993; Smith and Burnham, 2012; Wang et al., 2009, 2008). These findings clearly demonstrate that visible articulatory information can provide reliable cues to facilitate speech perception.

1.2. Clear speech

With the goal of increasing their intelligibility, speakers may alter their speech productions in response to the communicative needs of perceivers (Hazan and Baker, 2011; Kim et al., 2011; Smiljanić and Bradlow, 2009; Tasko and Greilick, 2010), such as when speaking in the presence of background noise (Sumby and Pollack, 1954), competing with other talkers (Lu and Cooke, 2008), or communicating with the hearing-impaired or non-native perceivers (Ferguson, 2012; Maniwa et al., 2009; Payton et al., 1994; Picheny et al., 1986). Such accommodations typically involve clear speech, a clarified, hyperarticulated speech style, relative to the natural plain, conversational speech style.

Acoustic measures show that plain-to-clear speech modifications of English vowels may involve increased duration, intensity, fundamental frequency value and range, formant frequency range and distance, and expanded vowel space (Bradlow et al., 1996; Ferguson, 2012; Ferguson and Kewley-Port, 2007, 2002; Ferguson and Quené, 2014; Hazan and Baker, 2011; Lam et al., 2012); as well as more dynamic spectral and temporal changes (Ferguson and Kewley-Port, 2007; Tasko and Greilick, 2010. The clear speech strategies found to be most effective in contributing to intelligibility are the expansion of the vowel space (and corresponding formant changes) and increased duration of vowels (Bond and Moore, 1994; Bradlow, 2002; Ferguson and Kewley-Port, 2007, 2002; Picheny et al., 1986). More specifically, compared to conversational speech, clear speech involves lower second formant for back vowels and higher second formant for front vowels, as well as higher first formant for all vowels, which presumably could be attributed to more extreme articulatory movements and longer articulatory excursions involving a higher degree of mouth opening and jaw lowering (Ferguson and Kewley-Port, 2007, 2002).

Moreover, there is evidence that clear speech vowel characteristics may interact with vowel tensity, with more expanded vowel space and longer duration for tense vowels than for lax vowels in clear speech (Picheny et al., 1986; Smiljanić and Bradlow, 2009). However, such evidence either lacks statistical power (Picheny et al., 1986), or is only restricted to the temporal domain (Smiljanić and Bradlow, 2008). Additionally, despite the fact that both tense and clear vowels bear similar acoustic correlates, the two factors are not cumulative to further enhance intelligibility (Ferguson and Quené, 2014). Further research is needed to examine the extent to which such acoustic effects, if any, are salient in articulation.

1.3. Articulatory features in clear speech

Given that acoustic variations in clear speech may be triggered by alterations in articulatory features, it is conceivable that such articulatory variations are measurable and can be perceived to aid intelligibility. It has been shown that the clear speech strategies that speakers adopt when conversing with normal as well as hearing-impaired persons in noisy settings may further enhance intelligibility when presented in both auditory and visual modalities as compared to audio-only presentation (Gagné et al., 2002; Sumby and Pollack, 1954). Furthermore, research has demonstrated that the benefits accrued from the availability of both visual information and a clear speaking style are complementary and not merely redundant sources of additional information in improving intelligibility over the auditory-only conversational style condition (Helfer, 1997).

The few studies that have performed such kinematic measurements showed positive correlations among articulation, acoustics, and intelligibility measures in clear speech effects (Kim and Davis, 2014a; Kim et al., 2014, 2011; Tasko and Greilick, 2010). Kim et al. (2011) used an Optotrak system to track the articulatory movements of clear speech produced in the presence of background noise

(Lombard speech) by measuring the motion of face markers as speakers produced English sentences either in quiet or in noise. They also tested the audio-visual intelligibility of Lombard speech embedded in noise. The tracking results revealed a greater degree of articulatory movement in speech in noise (clear speech) than in quiet (conversational speech), with the differences correlated with speech acoustics. Moreover, they found a visual speech benefit, where increased movement of the jaw and mouth (lip rounding) during clear speech translated to increased intelligibility. A follow-up study by Kim and Davis (2014a) investigated whether properties of clear speech were more distinct and less variable (i.e., more consistent across productions) than conversational speech. Consistent with the previous findings, this study also revealed that mouth and jaw motion was larger for clear than conversational speech, indicating that clear speech is more visually distinct. However, the degree of variability was comparable for clear and conversational speech. Tasko and Greilick (2010) used a midsagittal X-ray microbeam system to track tongue and jaw movements in clear versus conversational productions of the word-internal diphthong /aɪ/. The tongue fleshpoint tracking results show that, in clear relative to conversational speech, the tongue began in a lower position at the onset of diphthong transition and ended in a higher position at transition offset. This indicates that clear speech results in significantly larger and longer movements of the tongue and jaw, accompanied by the associated larger first and second formant changes.

1.4. Facial landmark detection

According to the survey of Çeliktutan et al. (2013), methods for facial landmark detection date back to classic methods known as Active Appearance models (AAMs) (Cootes et al., 1995) and elastic graph matching (Wiskott et al., 1997). Countless extensions followed, including Göcke and Asthana (2008) and Milborrow and Nicolls (2008). A recent research trend (Uricar et al., 2012; Zhu and Ramanan, 2012) for improving the performance of facial landmark detectors is to adopt a supervised approach, where one employs training data (videos with human-identified landmark annotations) to build a mathematical model that would predict landmark locations in new, unseen images (i.e. video frames not previously seen by the learned model). Models learned from training data have the advantage that they are trained to be less sensitive to algorithm parameters and scene variations (e.g. photographic changes due to illumination differences, changes in camera viewpoints, etc.). On the other hand, as learned models are trained to be as generalizable as possible to new unseen videos, they tend to compromise in terms of the precision of the detected landmark locations. Accordingly, based on the conclusions from the comparative analysis of Çeliktutan et al. (2013) and our own preliminary experiments, we chose to employ the state-of-the-art face detector of Zhu and Ramanan (2012) to first localize the facial landmarks and then develop image analysis processing algorithms to refine the positions of the landmarks identified by this detector to further improve localization precision. Further details on our video-analysis approach will be described in Section 2.

1.5. The present study

The findings from the earlier kinematic and acoustic studies of clear speech motivate the present analysis of the visual articulatory features of vowels in clear versus plain citation form speech. Vowels have been well documented in acoustic studies, but there are few reports on vowels in audio-visual analyses, especially concerning measurements of vowels differing in visible articulatory movements as a function of visual saliency. The previous studies primarily focused on general visual clear speech effects at the sentential level or specific effects in a single vowel. Moreover, the kinematic measures involve placing physical markers inside the oral cavity or on the speaker's face and head during optical motion capture or X-ray recording, and thus may be physically intrusive to the speaker as well as distracting to perceivers of the speech.

The present study aims at characterizing the differences in visible articulatory features (e.g., lip spreading, lip rounding, jaw position) of English vowels (/i, ɪ, ɑ, ʌ, u, ʊ/) in /kVd/ word tokens produced in clear and plain speech styles. This research uses advanced computerized facial detection and image processing techniques to extract the articulatory movement information from the speaker's face captured by video recordings. This video-analysis based technique thus differs from the previous studies in that no physical markers were placed on the speakers, a decision that facilitates more natural speech production as well as concurrent research on the perceptual correlates of the articulatory measurements reported here. We hypothesize that, compared to plain speech, vowels produced in clear speech involve greater motion of visible articulators. In particular, we expect a greater degree of lip spreading (for unrounded vowels such as /i/), lip rounding (for rounded vowels such as /u/), and jaw lowering (for low vowels such as /ɑ/) (Kim and Davis, 2014a; Tasko and Greilick, 2010). In addition, we expect the difference between clear and plain speech to be greater for the visually more salient tense vowels /i, ɑ, u/ than the less salient lax vowels /ɪ, ʌ, ʊ/. To the best of our knowledge, the specific articulatory contrasts for vowels in clear and plain speech in terms of the degree and direction of articulatory movements, as well as the more subtle distinctions between tense and lax tokens and their effects on visual saliency have not been explored in previous studies.

The present study employs the state-of-the-art face detector of Zhu and Ramanan (2012) to first localize the facial landmarks and then develop image analysis algorithms that further refine the landmarks identified by this detector to improve localization precision. To the best of our knowledge, while lip-tracking and face-detection

algorithms have been applied to various computer-vision problems (Çeliktutan et al., 2013; Göcke and Asthana, 2008; Uricar et al., 2012), e.g. for the development of security and intelligent car systems, the present study is the first to apply them to speech production research.

2. Methods

2.1. Experimental setup and data acquisition

2.1.1. Participants

Eighteen native speakers of English (10 females, 8 males) aged 17–30 were recruited as talkers for this study. All were born and raised in Canada with Western Canadian English as their native and dominant language. Since one of the target vowels was /ɑ/, the speakers were further asked to confirm that they pronounce the vowel in "cod" as /ɑ/ rather than /ɔ/ (i.e., their dialects do not contain an /ɑ-ɔ/ split). The speakers further reported no hearing, speech, or neurological impairments, and had normal or corrected-to-normal vision. As speakers' facial activities were video-captured for visual analysis of articulatory features, they were asked to wear a non-collared shirt and the visibility of their facial contours was ensured such that their eyebrows, jawline and lips should be clearly visible.

2.1.2. Stimuli

Three English vowel pairs /i-ɪ/, /ɑ-ʌ/, /u-ʊ/ differing in articulatory features (involving lip spreading, jaw lowering, and lip rounding, respectively) were the target vowels for examination in this study, with the tense vowels /i, ɑ, u/ having higher visual salience than the lax vowels /ɪ, ʌ, ʊ/. These six vowels were embedded in monosyllabic /kVd/ contexts, resulting in the corresponding English words "keyed", "kid", "cod", "cud", "cooed", and "could" for the elicitation procedure.

2.1.3. Elicitation of clear and plain speech

The elicitation of clearly produced and plain citation form speech followed the procedures developed by Maniwa et al. (2009). Using MATLAB (The Mathworks, R2013, Natick, MA, USA), a simulated interactive computer software was developed for stimulus presentation that seemingly attempted to perceive and recognize the tokens produced by a speaker. The speaker was instructed by the computer to read each token that was shown on the screen naturally (to elicit plain style productions), after which the program would show its 'guess' of the recited token. The software would systematically make wrong guesses due to 'perception errors' that involved tensity (e.g., "keyed" vs. "kid"), mouth position/opening (e.g., "keyed" vs. "cod"), lip-rounding (e.g., "keyed" vs. "cooed") errors, or an open error in which the participants were told that the computer program did not understand what they said. In response, the speaker was requested to repeat the token more clearly, as if to help the software disambiguate the confused tokens (to elicit clear style productions). The tokens were shown in a fixed random order. In total, each speaker produced 90 plain productions [6 words × 5 response types (1 correct + 4 error types) × 3 repetitions] and 72 clear productions (6 words × 4 error types × 3 repetitions). Prior to the elicitation session, all speakers underwent a warm-up session where they practiced producing the six words using the elicitation software and were familiarized with the on-screen prompts and widgets.

2.1.4. Recording

All recordings were made in a sound-attenuated booth in the Language and Brain Lab at Simon Fraser University, where each speaker was recorded individually. The stimuli for elicitation were displayed on a 15 in LCD monitor three feet directly in front of the speaker at or slightly above eye-level to facilitate the placement of a front-view video camera immediately below the monitor on a desktop tripod. Speakers were seated with their back against a monochromatic green backdrop. High definition front-view .mts (AVCHD) video recordings were made with a Canon Vixia HF30 camera at a recording rate of 29 fps. A second standard-definition camera (Sony Handycam) captured a left side view of the speaker's face. For interaction with the computer display, speakers were instructed in the usage of an Xbox controller, which offered a comfortable and quiet way to interact with the display with minimal movement required from the speaker or interference with the video and audio recordings. The alternative of a computer mouse and mousepad held on their lap was also offered for those who preferred such. After reading each word, speakers were instructed to return their face to a neutral position and keep their mouth closed. They were also asked to reset and hold their posture for two seconds in case of a disruption to the recording, such as after a cough or sneeze. Each of the productions was evaluated by two phonetically-trained native speakers of English to ensure that the speakers produced the intended vowels. All the productions were judged as correct productions of the target vowels.

2.2. Video analysis

The raw videos were first semi-automatically processed into annotated segments (at token-level) with MATLAB, using the audio channel of the recorded videos. Next, facial landmarks were extracted from each video frame of each video token using a fully automatic procedure and image analysis methods. Lastly, articulatory measurements were computed based on the detected facial landmark positions, including peak horizontal lip stretch, peak vertical lip stretch, peak eccentricity of lip rounding, peak vertical jaw displacement from the front-view analysis, as well as degree of lip protrusion from the side-view analysis.

2.2.1. Video-segmentation

To segment the video sequence into individual word tokens and trim out non-interest frames (e.g., when

speakers were not speaking for the requested tokens), speakers' word productions were automatically detected based on the audio signal using the algorithm of Giannakopoulos et al. (2010). Briefly, this algorithm extracts two sets of audio features (signal energy and spectral centroid) from the raw audio signal and dynamically estimates appropriate thresholds for defining when an audio frame contains human speech. The onset and offset of spoken words were then computed using the dynamically estimated thresholds. After some parameter-tuning, we found this algorithm to be fairly robust against ambient and other non-speech noises (mouse-clicks, etc.). Nevertheless, it was sensitive to other human-made sounds (e.g., coughs, dialog with the experimenters) and consequently, some of the generated detections had to be discarded through an operator-guided verification process.[1] Next, one-second onset and offset times were appended to each detected token such that the motions of the mouth made before and after the audio would also be included in the subsequent video analysis step. Lastly, each video segment was annotated with pairs of labels that identify the prompting token (e.g. 'cod', 'caught', etc.) and the nature of the token produced (i.e. 'plain' or 'clear', which identifies the token as being a response to a first elicitation, or a second one, spoken in response to a wrong guess, respectively). This annotation procedure was done using a semi-automatic procedure.

2.2.2. Analysis of front-view video tokens

In summary, for the front-view video analysis, facial landmark annotations of the segmented video tokens were performed using the following four steps. First, we ran a face detector for each frame in the segmented video to localize landmark candidates. Second, the horizontal and vertical scales of each subject's head were estimated. Third, we employed image analysis algorithms to examine the local intensity profiles of the landmarks to refine the candidate positions to obtain more accurate position estimates. Fourthly, the measurements of interest were computed using the detected facial landmarks. Fig. 1 displays an example front-view video frame and corresponding facial landmarks. We now describe the details of each step.

First, for landmark estimation, we employed the state-of-the-art face detector of Zhu and Ramanan (2012) over each frame of each segmented video sequence. This approach was employed as it was shown to be one of the top-performing face detectors in a recent comparative analysis (Çeliktutan et al., 2013) that evaluated a total of seven state-of-the-art methods for facial annotations using four large-scale public datasets. This face detector leverages machine learning techniques to build a mathematical model that was trained on public annotated videos (i.e., manually created facial landmark annotations). As several public video databases were employed, we found the detector robust to the variations in subjects' poses and lighting conditions, as well as local photographic changes (e.g., from shadows). Additionally, due to the "parts-based" approach of (Zhu and Ramanan, 2012), it generally detected the eye landmarks and nose tip in a robust and accurate manner. However, it was sensitive to differences in facial configurations (e.g., scale differences between parts) and thus gave less accurate estimates of lip landmarks, which thus motivated our refinement procedure as explained in step 4 to be described below.

Fig. 1. Example front-view video frame and corresponding facial landmarks. A black bar has been superimposed to protect the speaker's privacy.

Second, to account for head-size differences across subjects, we needed to estimate the scale factors needed to normalize these scale variations. We do so by estimating the horizontal and vertical scale factors of each speaker by computing, respectively, the inter-pupillary distance (IPD) and eye-to-nose-tip distance, which is defined as the distance from the nose-tip to its perpendicular projection on the pupil-line. For brevity, we denote these distances simply as horizontal distance (HD) and vertical distance (VD), respectively. More specifically, to calculate HD, we approximated each pupil's location as the mean of the four detected eye landmarks and computed HD as the distance between the approximated pupil locations. Similarly, we computed VD by computing the distance between the nose-tip and the point projected perpendicularly from the nose-tip onto the pupil-line.

Third, for lip landmark refinement, having obtained the lip landmarks estimated by the detector, we then computed a box that enclosed all detected lip landmarks. Intensity profiles at the middle of each side of the bounding box for all four sides were then examined. In order to avoid false detections due to shadows from the lips, the bounding box was drawn to be within a margin of $0.05 \times$ IPD to the closest lip landmarks. This strategy proved adequate in all of our videos due to the well-controlled recording setup. The final landmark positions were then computed as the location where the maximal intensity change was detected. Prior to drawing the intensity profiles, we converted each color frame to the HSV color-space that transforms the RGB color channel to 3 channels representing hue, saturation, and value of a color. Following Zhang and Wang

[1] This procedure involved having an operator correcting the annotations that were generated automatically based on the predefined order of the test tokens.

(2000), we discarded the hue and value channel, since hue information is unreliable for regions with low color saturation like the lips. Instead, we employed the saturation channel in which the lip is most distinguishable from the skin color. Further, we applied a low-pass spatial smoothing filter to each frame to remove speckle noise produced during data acquisition.

Finally, with the automatically detected and refined landmarks computed for all frames within each video token, we computed the following four articulatory measurements for the front-view video tokens:

- Peak of vertical lip stretch:
  $F1 = 1/VD \quad \text{quantile}(\{1^1_{\text{lipT}} - 1^1_{\text{lipB}}, \ldots, 1^f{}_{\text{lipT}} - l^f{}_{\text{lipB}}, \ldots, 1^n{}_{\text{lipT}} - l^n{}_{\text{lipB}}\}, 0.95)$,
  where 1^f_{lipT} and $1^f{}_{\text{lipB}}$ denote the top and bottom lip landmarks, respectively; f is the frame index, n is the video token length, and quantile denotes the quantile function that computes the 95th quantile of the input arguments;
- Peak of horizontal lip stretch:
  $F2 = 1/HD\,\text{quantile}(\{1^1_{\text{lipR}} - 1^1_{\text{lipL}}, \ldots, 1^f{}_{\text{lipR}} - l^f_{\text{lipL}}, \ldots, 1^n{}_{\text{lipR}} - l^n_{\text{lipL}}\}, 0.95)$,
  where 1^f_{lipL} and $1^f{}_{\text{lipR}}$ denote the left and right lip landmark, respectively;
- Peak of eccentricity of lip rounding:
  $F3 = \text{quantile}(\{r^1, \ldots, r^f, \ldots, r^n\}, 0.95)$,
  where $r^f = \sqrt{1 - b^2/(a^2+1)}$, with a and b being the constants of the fitted ellipses (note that a low eccentricity value indicates a greater extent of lip rounding);
- Peak of vertical jaw displacement:
  $F4 = 1/VD \quad \text{quantile}(\{I^1_{\text{jaw}} - I^{1+1}_{\text{jaw}}, \ldots, 1^f{}_{\text{jaw}} - 1^{f+1}{}_{\text{jaw}}, \ldots, 1^{n-1}{}_{\text{jaw}} - 1^n{}_{\text{jaw}}\}, 0.95)$,
  where 1^f_{jaw} is the landmark with the lowest y-coordinate as determined by the face detector.

Note that these articulatory measurements are not expressed in physical units but rather as normalized values, thereby facilitating scale normalization across speakers. For example, the vertical lip stretch and jaw displacement are expressed in reference to distance from nose-tip to pupil-line of each subject, while horizontal lip stretches are expressed in reference to IPD.

2.2.3. Analysis of side-view video tokens

Another novel methodological contribution of the present study is the way in which we examined the articulatory features relating to lip protrusion, which we hypothesize to be salient for "cooed" and "could" tokens that involved the rounded /u/ and /ʊ/ vowels, respectively. For this purpose, we also captured side-view videos of the speakers, in addition to the front-view videos. As the side-view videos show a limited set of facial features (nose tip, one eye, one ear, and the shape of the lips), computer-aided analyses of these videos required a different procedure. Specifically, as only one eye is visible on the side, we could not employ intraocular distances and mid-point-to-nose-tip distances for spatial normalization to adjust for the scale changes of the facial features across tokens. Accordingly, we employed linear image registration to normalize the scale changes in these videos. Below is a detailed description of the side-view video analysis procedure, which at a high level involved scale normalization and lip-protrusion quantification.

To initialize, for each speaker, a video token was selected as a reference video (V_{Ref}). Then, the first step involves scale normalization. For each video token V_i, the following steps were performed: (1) Using the "profile-face" detector of MATLAB (The Mathworks, R2013, Natick, MA, USA), a region belonging to the complete side-view of the face was semi-automatically[2] detected. (2) For each video frame, the facial outlines were detected to generate a rough *outline* of the face that would represent the contours of the face at each time instance. This was achieved[3] by using a Sobel edge detector. Based on empirical experiments, we found the edge detector robust against shadows, illumination changes, and image noises, thanks to the homogeneity of the background and the saliency of the face contour. (3) Next, a *feature image* (FI) was computed that summarizes the face contours across time for each video token. This was done by summing the intensity values of corresponding pixels of all outline images in a token and subsequently dividing each pixel by token size (number of frames in each token). (4) To spatially normalize the scale differences between video tokens, linear image registration was performed to resolve the similarity transform (i.e. 4 degrees of freedom: vertical and horizontal translation, clockwise rotation, and scaling/resizing) that would align each FI of V_i to the FI of the reference video V_{Ref}. (5) Finally, the spatially normalized FIs were trimmed so they all center around the lip region. Fig. 2 displays an example output of steps in the above algorithm.

The second step was lip-protrusion quantification, involving extracting some summary statistics about the FIs that would quantify the visual differences between the degree of lip protrusion involved in each token. These measurements should be sensitive to capture the degree of lip protrusion across time but remain insensitive to token length. Accordingly, to quantify the relative difference between the registered FIs and the FI of V_{Ref}, which is mainly due to lip deformations, we computed an image dissimilarity measure. The image dissimilarity measures examined are: mean of absolute difference (MAD), mutual information (MI) (Maes et al., 1997), and sum of conditional variance (SCV) (Pickering et al., 2009). These are standard image measures that examine how two images

[2] The face region is manually redrawn when the detected region is deemed incorrect as judged by the experimenter after quick visual inspection.

[3] We also explored the options of using a Canny or a Prewitt's edge detector. We found responses to the Sobel edge detector gave the right amount of contour details even without parameter-tuning, as required by the explored alternatives.

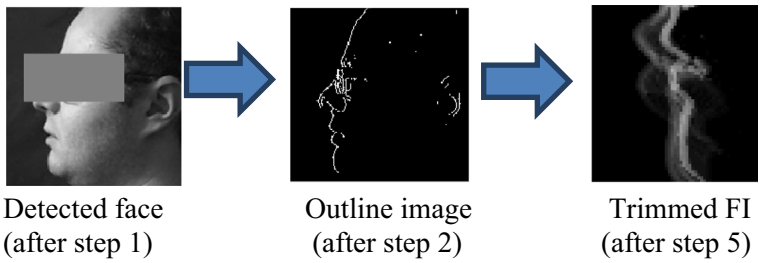

Fig. 2. Example side-view video frame (left) and corresponding edge features (middle) which are used to compute the feature image of an entire video token (right). A grey bar has been superimposed to protect the speaker's privacy.

differ. Specifically, MAD is a computationally efficient measure that assumes that the intensity relationship is preserved across the FIs compared, while the other two measures are statistical measures that do not require this assumption and thus may be more robust to illumination changes, albeit slightly more computationally intensive to calculate. Generally, we find the tokens with amplified lip protrusions yielded higher dissimilarity between the compared FIs, thus yielding higher values in the dissimilarity measure. Fig. 3 illustrates the relative differences between the registered FIs and the FI of V_{Ref}.

3. Results

The extracted measurements from the front and side videos, including horizontal and vertical lip stretch, jaw displacement, lip rounding and lip protrusion, as well as duration, were submitted to statistical analyses. For conciseness, only the significant effects and interactions involving style are reported.

3.1. Front-view analysis

For each of the front-view measurements, a series of $2 \times 2 \times 2$ repeated measures analysis of variance (ANOVAs) was conducted with Style, Gender, and Tensity as factors. The ANOVAs show significant differences for the main effects of Style, Gender, and Tensity for the various measurements. Since no significant interactions of Style and Tensity were observed, subsequent style comparisons pooled data across Tensity for each vowel pair. Firstly, as hypothesized, there is a significant main effect of Style: in horizontal [$F(1,765) = 21.5$, $p < .001$] and vertical [$F(1,765) = 51.0$, $p < .001$] stretches for "keyed/kid", in vertical stretch [$F(1,765) = 24.2$, $p < .001$] and jaw displacement [$F(1,765) = 8.6$, $p = .003$] for "cod/cud", and in rounding [$F(1,655) = 4.8$, $p = .028$], vertical stretch [$F(1,655) = 21.7$, $p < .001$] and jaw displacement [$F(1,655) = 6.6$, $p = .010$] for "cooed/could". Fig. 4(a)–(d) displays the measurement comparisons between plain and clear speech styles for each of the three word pairs (the measurements per word, style and gender are displayed in Table A1). As shown in the figure, for each of these significant differences in style, the extent of movement in clear speech is greater than in plain speech. Additionally, for each word pair, the duration in clear speech was longer than in plain speech, as expected ($p < .05$). For the main effect of Tensity, tense vowels show longer duration and a greater degree of displacement than lax vowels, involving greater horizontal lip stretches for "keyed" than "kid", greater vertical lip stretches for "cod" than "cud", and greater lip stretches in both directions for "cooed" than "could" ($p < .05$). Moreover, a significant main effect of Gender was observed in the horizontal and vertical stretch for "keyed/kid", and in all of the measurements of "cod/cud" and "cooed/could", with overall greater extent of movement in male than female productions ($p < .05$).

The statistically significant interactions mostly involved Style and Gender. Post-hoc analyses were further conducted to examine the effects of Style per Gender group for each pair of words using a series of one-way ANOVAs. As shown in Fig. 4, for keyed/kid, the vertical lip stretch is greater in clear ($M = 1.35$) than plain speech ($M = 1.17$) [$F(1,352) = 36.5$, $p < .001$] in males. To a lesser degree, in

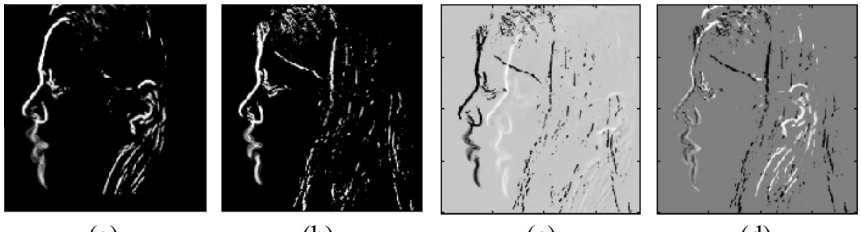

Fig. 3. (a) Feature image (FI) of reference video token; (b) FI of another token; (c) intensity difference between these FIs; (d) intensity difference between the registered FI pair. Visual inspection of (d) is a common and effective way to examine the quality of spatial alignment of images (Tang et al., 2008; Tang and Hamarneh, 2013).

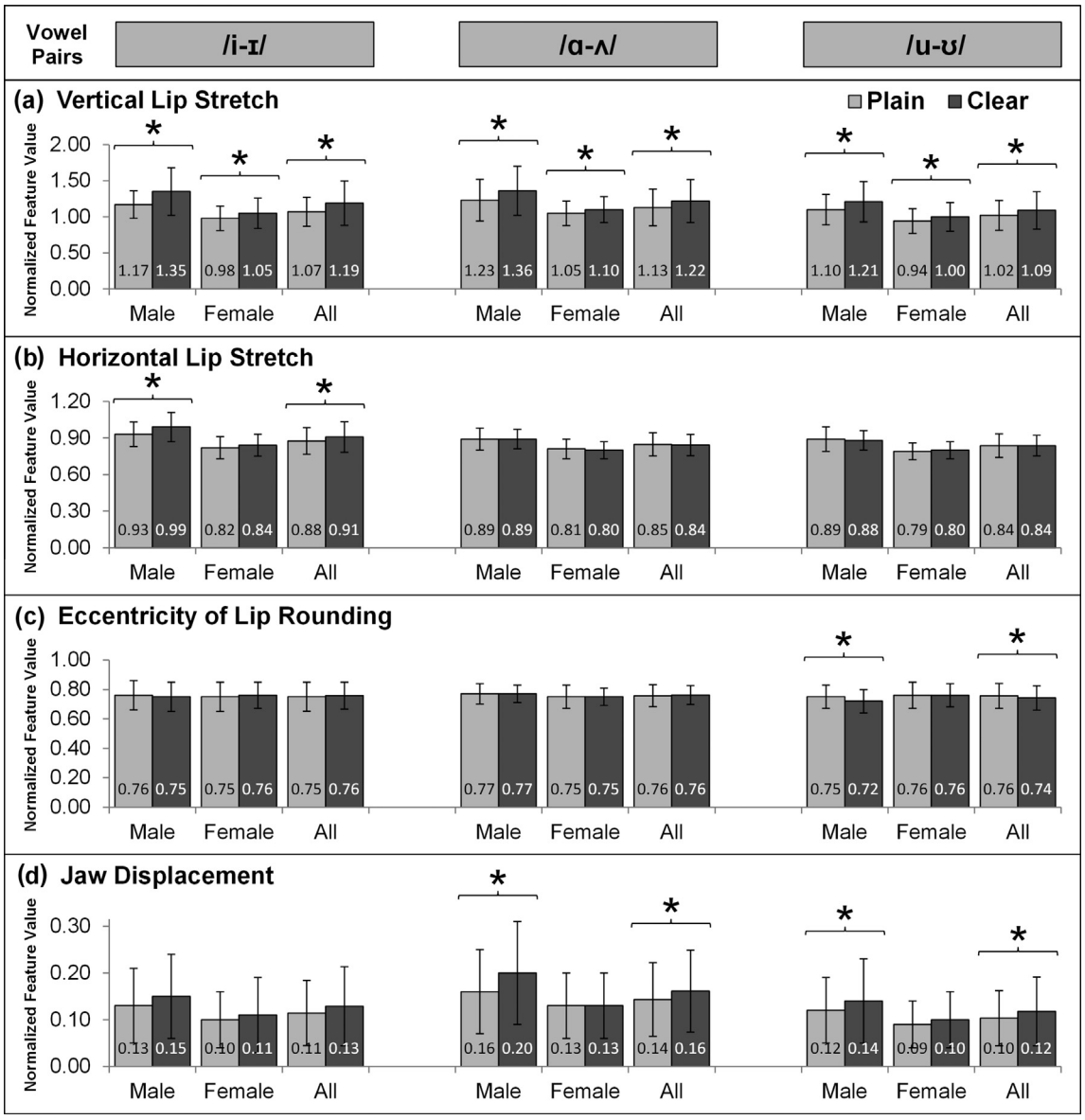

Fig. 4. Comparisons between plain and clear speech styles for each pair of vowels by male ($n = 8$) and female ($n = 10$) speakers, as well as the same speech style comparisons with gender-weighted mean values across all speakers (All). The values are the mean peak of (a) vertical lip stretch, (b) horizontal lip stretch, (c) eccentricity of lip rounding, and (d) jaw displacement. Error bars indicate ±1 standard deviation. These measurements are not expressed in physical units but as a fraction of the speaker's head to facilitate scale normalization across speakers. Note that while a larger value generally indicates greater displacement, a low eccentricity value indicates a greater extent of lip rounding. "*" indicates statistically significant style effects ($p < .05$).

females, the vertical lip stretch is also greater in clear speech ($M = 1.05$) than plain speech ($M = 0.98$) [$F(1,410) = 15.1$, $p < .001$]. Horizontal lip stretch is also greater in clear speech ($M = 0.99$) than in plain ($M = 0.93$) [$F(1,352) = 19.7$, $p < .001$] for males, but the difference is not significant for females ($M = 0.84$ vs. $M = 0.82$) [$F(1,410) = 3.7$, $p = .055$]. For cod/cud, the vertical lip stretch is greater in clear ($M = 1.36$) than plain speech ($M = 1.23$) [$F(1,360) = 15.5$, $p < .001$] in males. To a lesser degree, in females, the vertical lip stretch is also greater in clear speech ($M = 1.10$) than plain speech ($M = 1.05$) [$F(1,402) = 8.4$, $p = .004$]. In addition, for males, the jaw movement was greater in clear than in plain speech ($M = 0.20$ vs. $M = 0.16$) [$F(1,360) = 9.2$, $p = .003$], but no such difference was observed in females. For cooed/could, the vertical lip stretch is greater in clear ($M = 1.21$) than in plain speech ($M = 1.10$) [$F(1,298) = 13.7$, $p < .001$] in males. To a lesser degree, in females, the vertical lip stretch is also greater in clear ($M = 1.00$) than in plain speech ($M = 0.94$) [$F(1,354) = 7.5$, $p = .007$]. Additionally, males employed greater jaw movement in clear than in plain speech ($M = 0.14$ vs. $M = 0.12$) [$F(1,298) = 7.2$, $p < .001$], but no such difference was observed in females.

3.2. Side-view analysis

To test the hypothesis that differences in style can be observed in terms of lip protrusion for the rounded vowels "cooed" and "could", a 3-way ANOVA was performed on the extracted side-view measurements with Style, Gender, and Tensity as factors. The results show a significant main effect of Style [$F(1,847) = 52.5$, $p < .001$], with the extent of lip protrusion being greater in clear speech ($M = 0.08$) than in plain speech ($M = 0.07$). Additionally, a significant main effect of Gender was observed, with male speakers ($M = 0.10$) showing greater lip protrusion than female speakers ($M = 0.06$) [$F(1,847) = 291.5$, $p < .001$]. Although there was no significant interaction of Style and Tensity or of Style, Tensity, and Gender, a significant Style and Gender interaction was observed [$F(1,847) = 4.4$, $p = .036$]. Subsequent one-way ANOVAs for each gender with Style as the within-subject factor revealed that for males, the extent of lip protrusion was greater in clear speech ($M = 0.11$) than in plain speech ($M = 0.08$) [$F(1,309) = 40.6$, $p < .001$]. To a lesser degree, a greater degree of lip protrusion for clear ($M = 0.07$) versus plain style ($M = 0.05$) in the female speakers was also observed [$F(1,402) = 26.2$, $p < 0.001$]. Fig. 5 displays the lip protrusion comparisons between plain and clear speech styles for the word pair cooed/could containing rounded vowels.

In sum, plain-to-clear speech modifications involve longer duration, greater vertical lip stretch and jaw movement in all three pairs of words, as well as a greater degree of lip rounding and lip protrusion for the words involving rounded vowels. Additionally, relative to female speakers, male speakers exhibited greater speech style differences, particularly greater degrees of horizontal lip stretch (for keyed/kid) and jaw movement (for cod/cud, cooed/could) in clear than plain speech.

4. Discussion

This study makes use of dual-view video sequences to examine articulatory features between clearly produced and plain citation form speech, involving a representative set of vowels embedded in English monosyllabic words. The overall results support our hypothesis of greater articulatory movements for clear relative to plain speech. Specifically, the finding of longer durations in clear than plain speech across vowels agrees with both the acoustic and articulatory data in previous research (Ferguson and Kewley-Port, 2002; Kim and Davis, 2014a; Lu and Cooke, 2008; Picheny et al., 1986). The articulatory motion results are consistent with previous findings of the acoustic features of vowels in clear speech (e.g., Bond and Moore, 1994; Bradlow et al., 1996; Ferguson and Kewley-Port, 2007, 2002) in that expanded acoustic vowel space and more peripheral formant frequencies in clear speech may be attributed to more extreme and greater extent of articulatory movements, in terms of vertical lip movement, jaw lowering, horizontal lip stretches, and lip protrusion.

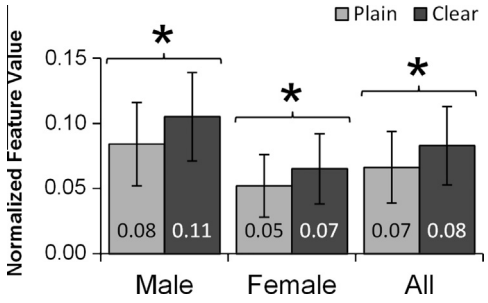

Fig. 5. Comparisons of degree of lip protrusion between plain and clear speech styles for the word pair cooed/could containing rounded vowels produced by male ($n = 8$) and female ($n = 10$) speakers, as well as the same speech style comparisons with gender-weighted mean values across all speakers (All). Error bars indicate ±1 standard deviation. "*" indicates statistically significant style effects ($p < .05$).

In particular, the present finding of clear speech effects attributable to greater vertical lip and jaw movements across words is consistent with the previous acoustic studies showing an increase in first formant values across vowels in clear speech as compared to plain speech (Ferguson and Kewley-Port, 2002; Ferguson and Quené, 2014; Kim and Davis, 2014a; Lu and Cooke, 2008). Additionally, clear speech typically involves increased vocal effort (consequently, increased intensity), which also requires a larger jaw opening (Huber et al., 1999; Kim and Davis, 2014a; Schulman, 1989). In terms of articulation, these findings are also in line with the claim that the jaw and lower lip, which give rise to vertical displacement, are more relevant to active speech articulation and can be better tracked than the upper lip (as a passive articulator in this process) (Yehia et al., 2002). Moreover, the horizontal lip movement also shows clear speech effects in the production of the front vowels (/i-ɪ/ in keyed/kid) that involve horizontal lip spreading. Acoustically, it has been found that the second formant of these front vowels generally increases in clear speech relative to plain speech (Ferguson and Kewley-Port, 2002; Ferguson and Quené, 2014; Lu and Cooke, 2008). Thus, the increase in the second formant is conceivably due (in part) to the shortening of the vocal tract resulting from lip-spreading. Finally, both the front- and side-view videos captured the (slightly but significantly) greater degree of lip-rounding and lip-protrusion for the rounded vowels /u-ʊ/ in cooed/could in clear versus plain speech. The observation that the second formant of these rounded vowels is lower in clear than in plain speech (Ferguson and Kewley-Port, 2002) is consistent with the present articulatory finding that speakers lengthen their vocal tract by rounding and protruding their lips to a greater degree in clear speech. Previous research has shown that perceivers rely more on the visual feature of lip rounding than the auditory information to perceive rounded vowels (Traunmüller and Öhrström, 2007). However, no articulatory kinematic research has examined the visible articulatory features of lip rounding and lip protrusion in clear versus plain speech. The present finding

thus offers new evidence of clear speech effects in the articulation of rounded vowels, and opens the door to the investigation of additional rounded segments (such as /w/ and /ɹ/).

The current study revealed an overall greater and longer articulatory movement for tense vowels compared to lax vowels, consistent with previous acoustic findings (Ferguson and Quené, 2014; Hillenbrand et al., 1995). However, the results show no interaction between vowel tensity and speech style, contrary to the predictions of greater clear speech effects for tense than lax vowels. The lack of greater plain-to-clear speech modifications in articulation for tense vowels may be due to articulatory constraints. The productions of tense vowels and clear-speech vowels both involve longer articulatory excursions, such as greater lip-spreading, jaw displacement, and lip-rounding. These extreme articulatory features that are intrinsic to tense vowels may have limited the room for further modifications that are more "deliberate" such as in clear speech (cf. Hazan and Markham, 2004). Acoustically, there has been supporting evidence showing that the spectral distance between tense and lax vowels is smaller in clear speech as compared to that in plain speech, indicating constraints to variability in speech style for tense vowels involving more peripheral formant frequencies (e.g., /i/) (Granlund et al., 2012).

The results reveal an unexpected gender effect in that male speakers often show greater clear speech effects than female speakers, particularly involving greater degrees of horizontal lip stretch and jaw movement. These patterns are not consistent with some of the previous findings showing no or less clear-cut gender effects on clear speech production both from acoustic and articulatory measures (Hazan and Markham, 2004; Kim and Davis, 2014a; Tasko and Greilick, 2010; Traunmüller and Öhrström, 2007). The observed gender differences may be due to anatomical factors. Past research has indicated that the size of the vocal tract and articulators may be positively correlated with movement displacement (Kuehn and Moll, 1976; Perkell et al., 2002). It may thus be speculated that the current male speakers' greater articulatory movements in clear speech could be attributed to their larger-size articulators relative to females', which allow more room for variability and more extreme speech articulation. However, it is not clear whether such differences are idiosyncratic in nature or can serve as effective cues to perception. Future research is needed to evaluate if these gender differences can be captured in perception to affect the intelligibility of clear versus plain speech.

The current results from video image analyses also extend the previous articulatory findings based on kinematic measures reporting increased movement of the tongue, jaw, and mouth during clear speech (Kim and Davis, 2014a; Kim et al., 2011; Perkell et al., 2002; Tasko and Greilick, 2010). In line with Kim and Davis' (2014) findings of greater vertical jaw movements and mouth opening in clear versus plain speech, the current study additionally shows clear speech modifications in terms of vertical and horizontal lip movements as well as lip rounding and protrusion. Moreover, most of the vowel kinematic studies focus on one vowel (e.g., a single diphthong, Tasko and Greilick, 2010; or a single vowel embedded in different consonantal contexts, Perkell et al., 2002), or conduct articulatory analyses based on measurements across vowels in an utterance (e.g., Kim and Davis, 2014a). The present study systematically examined a representative set of vowels with specific predictions on the basis of their articulatory and acoustic features, demonstrating that subtle visible articulatory movements in vowel productions can be captured in dual-view video recording and extracted using advanced video image processing techniques, without the need for placing physical markers on the speakers. This method not only allows more natural speech production, but also enables concurrent speech intelligibility research on the perceptual correlates of the articulatory measurements. The current research thus points to promising directions to apply computerized lip-tracking and face-detection algorithms (Göcke and Asthana, 2008) to the study of speech production, acoustics, and perception.

5. Concluding remarks

Using image processing analysis techniques with data collected from video captured from two complementary views, this study revealed a positive relationship between speech style and motion of visible articulators, indicating that clear speech relative to plain speech involves greater lip and jaw movements. These findings demonstrate that the novel computer-assisted face annotations and tracking techniques can precisely quantify and characterize the differences in articulatory features of speech segments produced in clear and plain speech style. Future video image analyses may involve real-time tracking to capture the dynamic features of vowels (including those that have been examined using kinematic measures, e.g., movement velocity, Perkell et al., 2002). Additionally, the side-view data acquired from the second camera points to the promising potential of using multiple cameras simultaneously, as well as the potential of reconstructing a multi-dimensional surface of the face and extracting more elaborate shape descriptors.

The present articulatory results contribute to a three-pronged approach to encompass articulatory, acoustic, and intelligibility analyses of clear and plain speech to explore the extent to which the measured articulatory differences correlate to differences in acoustic features between clear and plain speech, and the extent to which the differences in articulatory features are perceptible such that these observed clear speech features can enhance speech intelligibility.

Acknowledgments

We would like to thank Anthony Chor, Mathieu Dovan, Katelyn Eng, Parveen Kaila, Alyssa Lee, Keith

Table A1
Comparisons between plain and clear speech styles for each word by male and female speakers. The values are the mean amount of vertical lip stretch, horizontal lip stretch, eccentricity of lip rounding and jaw displacement (with standard deviations in parentheses).

		Vertical lip stretch		Horizontal lip stretch		Eccentricity of lip rounding		Jaw displacement	
		Male	Female	Male	Female	Male	Female	Male	Female
Keyed	Plain	1.186	0.978	0.933	0.836	0.745	0.752	0.127	0.101
		(0.194)	(0.157)	(0.100)	(0.083)	(0.101)	(0.102)	(0.074)	(0.064)
	Clear	1.355	1.055	0.994	0.854	0.748	0.770	0.129	0.109
		(0.317)	(0.225)	(0.119)	(0.087)	(0.102)	(0.080)	(0.087)	(0.079)
Kid	Plain	1.156	0.979	0.937	0.805	0.773	0.739	0.133	0.103
		(0.186)	(0.181)	(0.098)	(0.094)	(0.091)	(0.096)	(0.084)	(0.068)
	Clear	1.330	1.049	0.975	0.820	0.765	0.747	0.178	0.170
		(0.340)	(0.197)	(0.112)	(0.081)	(0.088)	(0.096)	(0.079)	(0.561)
Cod	Plain	1.278	1.075	0.893	0.802	0.770	0.742	0.159	0.123
		(0.328)	(0.174)	(0.083)	(0.073)	(0.060)	(0.080)	(0.095)	(0.077)
	Clear	1.402	1.129	0.878	0.790	0.774	0.757	0.187	0.128
		(0.376)	(0.191)	(0.078)	(0.059)	(0.061)	(0.058)	(0.108)	(0.076)
Cud	Plain	1.171	1.010	0.896	0.809	0.768	0.751	0.158	0.129
		(0.209)	(0.170)	(0.090)	(0.094)	(0.078)	(0.081)	(0.080)	(0.068)
	Clear	1.309	1.057	0.906	0.807	0.766	0.751	0.207	0.130
		(0.271)	(0.158)	(0.087)	(0.071)	(0.065)	(0.072)	(0.101)	(0.067)
Cooed	Plain	1.072	0.915	0.893	0.805	0.748	0.780	0.111	0.091
		(0.186)	(0.158)	(0.089)	(0.083)	(0.089)	(0.088)	(0.072)	(0.050)
	Clear	1.151	0.960	0.883	0.810	0.715	0.768	0.137	0.090
		(0.257)	(0.191)	(0.076)	(0.086)	(0.090)	(0.090)	(0.095)	(0.056)
Could	Plain	1.132	0.970	0.880	0.783	0.756	0.742	0.130	0.093
		(0.236)	(0.177)	(0.103)	(0.065)	(0.071)	(0.087)	(0.073)	(0.050)
	Clear	1.264	1.031	0.873	0.795	0.734	0.747	0.145	0.100
		(0.289)	(0.195)	(0.078)	(0.063)	(0.071)	(0.074)	(0.080)	(0.067)

Leung, and Elysia Saundry from the Language and Brain Lab at Simon Fraser University for their assistance in data collection. We also thank Dr. Kazumi Maniwa for sharing the elicitation method used in Maniwa et al. (2009). This research has in part been funded by a research grant from the Social Sciences and Humanities Research Council of Canada (SSHRC Insight Grant 435-2012-1641). Portions of this study were presented at the 18th International Congress of Phonetic Sciences, Glasgow, UK, August, 2015.

Appendix A

See Table A1.

References

Bernstein, L.E., Auer, E., Takayanagi, S., 2004. Auditory speech detection in noise enhanced by lipreading. Speech Commun. 44, 5–18. http://dx.doi.org/10.1016/j.specom.2004.10.011.

Bond, Z.S., Moore, T.J., 1994. A note on the acoustic–phonetic characteristics of inadvertently clear speech. Speech Commun. 14, 325–337. http://dx.doi.org/10.1016/0167-6393(94)90026-4.

Bradlow, A.R., 2002. Confluent talker- and listener-oriented forces in clear speech production. In: Gussenhoven, C., Warner, N. (Eds.), Laboratory Phonology 7. Mouton de Gruyter, Berlin, pp. 241–273.

Bradlow, A.R., Torretta, G.M., Pisoni, D.B., 1996. Intelligibility of normal speech I: global and fine-grained acoustic–phonetic talker characteristics. Speech Commun. 20, 255–272. http://dx.doi.org/10.1016/S0167-6393(96)00063-5.

Çeliktutan, O., Ulukaya, S., Sankur, B., 2013. A comparative study of face land-marking techniques. EURASIP J. Image Vide. 2013 (13), 1–27. http://dx.doi.org/10.1186/1687-5281-2013-13.

Chen, T.H., Massaro, D.W., 2008. Seeing pitch: visual information for lexical tones of Mandarin–Chinese. J. Acoust. Soc. Am. 123, 2356–2366. http://dx.doi.org/10.1121/1.2839004.

Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J., 1995. Active shape models – their training and application. Comput. Vis. Image Understand. 61, 38–59.

Cvejic, E., Kim, J., Davis, C., 2012. Recognizing prosody across modalities, face areas and speakers: examining perceivers' sensitivity to variable realizations of visual prosody. Cognition 122, 442–453. http://dx.doi.org/10.1016/j.cognition.2011.11.013.

Ferguson, S.H., 2012. Talker differences in clear and conversational speech: vowel intelligibility for older adults with hearing loss. J. Speech Lang. Hear. Res. 55, 779–790. http://dx.doi.org/10.1121/1.1788730.

Ferguson, S.H., Kewley-Port, D., 2002. Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. J. Acoust. Soc. Am. 112, 259–271. http://dx.doi.org/10.1121/1.1482078.

Ferguson, S.H., Kewley-Port, D., 2007. Talker differences in clear and conversational speech: acoustic characteristics of vowels. J. Speech Lang. Hear. Res. 50, 1241–1255. http://dx.doi.org/10.1044/1092-4388(2007/087).

Ferguson, S.H., Quené, H., 2014. Acoustic correlates of vowel intelligibility in clear and conversational speech for young normal-hearing and elderly hearing-impaired listeners. J. Acoust. Soc. Am. 135, 3570–3584, doi:http://dx.doi.org/10.1121/1.4874596.

Gagné, J.P., Rochette, A.J., Charest, M., 2002. Auditory, visual and audiovisual clear speech. Speech Commun. 37, 213–230. http://dx.doi.org/10.1016/S0167-6393(01)00012-7.

Giannakopoulos, T., Petridis, S., Perantonis, S., 2010. User-driven recognition of audio events in news videos. In: Proceedings of the 5th International Workshop on Semantic Media Adaptation and Personalization (SMAP), pp. 44–49, doi:http://dx.doi.org/10.1109/SMAP.2010.5706867.

Göcke, R., Asthana, A., 2008. A comparative study of 2D and 3D lip tracking methods for AV ASR. Proc. Intl. Conf. Audit.-Visual Speech Process. 2008, 235–240.

Granlund, S., Hazan, V., Baker, R., 2012. An acoustic–phonetic comparison of the clear speaking styles of Finnish–English late bilinguals. J. Phon. 40, 509–520. http://dx.doi.org/10.1016/j.wocn.2012.02.006.

Grant, K.W., Seitz, P.F., 1998. Measures of auditory-visual integration in nonsense syllables and sentences. J. Acoust. Soc. Am. 104, 2438–2450. http://dx.doi.org/10.1121/1.423751.

Hazan, V., Baker, R., 2011. Acoustic–phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. J. Acoust. Soc. Am. 130, 2139–2152. http://dx.doi.org/10.1121/1.3623753.

Hazan, V., Markham, D., 2004. Acoustic–phonetic correlates of talker intelligibility for adults and children. J. Acoust. Soc. Am. 116, 3108–3118. http://dx.doi.org/10.1121/1.1806826.

Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., Chung, H., 2006. The use of visual cues in the perception of non-native consonant contrasts. J. Acoust. Soc. Am. 119, 1740–1751. http://dx.doi.org/10.1121/1.2166611.

Hazan, V., Kim, J., Chen, Y., 2010. Audiovisual perception in adverse conditions: language, speaker and listener effects. Speech Commun. 52, 996–1009. http://dx.doi.org/10.1016/j.specom.2010.05.003.

Helfer, K.S., 1997. Auditory and auditory-visual perception of clear and conversational speech. J. Speech Lang. Hear. Res. 40, 432–443. http://dx.doi.org/10.1044/jslhr.4002.432.

Hillenbrand, J., Getty, L.A., Clark, M.J., Wheeler, K., 1995. Acoustic characteristics of American English vowels. J. Acoust. Soc. Am. 97, 3099–3111. http://dx.doi.org/10.1121/1.411872.

Huber, J.E., Stathopoulos, E.T., Curione, G.M., Ash, T.A., Johnson, K., 1999. Formants of children, women, and men: the effects of vocal intensity variation. J. Acoust. Soc. Am. 106, 1532–1542. http://dx.doi.org/10.1121/1.427150.

Kim, J., Davis, C., 2014a. Comparing the consistency and distinctiveness of speech produced in quiet and in noise. Comput. Speech Lang. 28, 598–606. http://dx.doi.org/10.1016/j.csl.2013.02.002.

Kim, J., Davis, C., 2014b. How visual timing and form information affect speech and non-speech processing. Brain Lang. 137, 86–90. http://dx.doi.org/10.1016/j.bandl.2014.07.012.

Kim, J., Sironic, A., Davis, C., 2011. Hearing speech in noise: seeing a loud talker is better. Perception 40, 853–862. http://dx.doi.org/10.1068/p6941.

Kim, J., Cvejic, E., Davis, C., 2014. Tracking eyebrows and head gestures associated with spoken prosody. Speech Commun. 57, 317–330. http://dx.doi.org/10.1016/j.specom.2013.06.003.

Krahmer, E., Swerts, M., 2007. The effects of visual beats on prosodic prominence: acoustic analyses, auditory perception and visual perception. J. Mem. Lang. 57, 396–414. http://dx.doi.org/10.1016/j.jml.2007.06.005.

Kuehn, D.P., Moll, K.L., 1976. A cineradiographic study of VC and CV articulatory velocities. J. Phon. 4, 303–320.

Lam, J., Tjaden, K., Wilding, G., 2012. Acoustics of clear speech: effect of instruction. J. Speech Lang. Hear. Res. 55, 1807–1821. http://dx.doi.org/10.1044/1092-4388.

Lu, Y., Cooke, M., 2008. Speech production modifications produced by competing talkers, babble, and stationary noise. J. Acoust. Soc. Am. 124, 3261–3275. http://dx.doi.org/10.1121/1.2990705.

Maes, F., Collignon, A., Vandermeulen, D., Suetens, M.G.P., 1997. Multimodality image registration by maximization of mutual information. IEEE Trans. Med. Imag. 16, 187–198.

Maniwa, K., Jongman, A., Wade, T., 2009. Acoustic characteristics of clearly spoken English fricatives. J. Acoust. Soc. Am. 125, 3962–3973. http://dx.doi.org/10.1121/1.2990715.

Massaro, D.W., 1987. Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry. Erlbaum, Hillsdale, NJ.

Milborrow, S., Nicolls, F., 2008. Locating facial features with an extended active shape model. In: Forsyth, D., Torr, P., Zisserman, A. (Eds.),

Computer Vision–ECCV 2008. Springer, Berlin Heidelberg, pp. 504–513, http://dx.doi.org/10.1007/978-3-540-88693-8_37.

Mixdorff, H., Hu, Y., Burnham, D., 2005. Visual cues in Mandarin tone perception. Proc. Interspeech, 405–408.

Payton, K.L., Uchanski, R.M., Braida, L.D., 1994. Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. J. Acoust. Soc. Am. 95, 1581–1592. http://dx.doi.org/10.1121/1.408545.

Perkell, J.S., Zandipour, M., Matthies, M.L., Lane, H., 2002. Economy of effort in different speaking conditions. I. A preliminary study of intersubject differences and modeling issues. J. Acoust. Soc. Am. 112, 1627–1641. http://dx.doi.org/10.1121/1.1506369.

Picheny, M.A., Durlach, N.I., Braida, L.D., 1986. Speaking clearly for the hard of hearing II: acoustic characteristics of clear and conversational speech. J. Speech Lang. Hear. Res. 29, 434–446. http://dx.doi.org/10.1044/jshr.2904.434.

Pickering, M.R., Muhit, A.A., Scarvell, J.M., Smith, P.N., 2009. A new multi-modal similarity measure for fast gradient-based 2d–3d image registration. Proc. IEEE Eng. Med. Biol. Soc., 5821–5824, http://dxdoi.org/10.1109/IEMBS.2009.5335172.

Schulman, R., 1989. Articulatory dynamics of loud and normal speech. J. Acoust. Soc. Am. 85, 295–312. http://dx.doi.org/10.1121/1.397737.

Sekiyama, K., Tohkura, Y., 1993. Inter-language differences in the influence of visual cues in speech perception. J. Phon. 21, 427–444.

Smiljanić, R., Bradlow, A.R., 2008. Stability of temporal contrasts across speaking styles in English and Croatian. J. Phon. 36, 91–113. http://dx.doi.org/10.1016/j.wocn.2007.02.002.

Smiljanić, R., Bradlow, A.R., 2009. Speaking and hearing clearly: talker and listener factors in speaking style changes. Lang. Linguist. Compass 3, 236–264. http://dx.doi.org/10.1111/j.1749-818X.2008.00112.x.

Smith, D., Burnham, D., 2012. Faciliation of Mandarin tone perception by visual speech in clear and degraded audio: implications for cochlear implants. J. Acoust. Soc. Am. 131, 1480–1489. http://dx.doi.org/10.1121/1.3672703.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 2, 212–215. http://dx.doi.org/10.1121/1.1907309.

Summerfield, Q., 1979. Use of visual information for phonetic perception. Phonetica 36, 314–331, doi:http://dx.doi.org/10.1159/000259969.

Summerfield, Q., 1992. Lip-reading and audio-visual speech perception. Philos. Trans. R. Soc. B 335, 71–78. http://dx.doi.org/10.1098/rstb.1992.0009.

Tang, L., Hamarneh, G., 2013. Medical image registration: a review. In: Farncombe, T., Iniewski, K. (Eds.), Medical Imaging: Technology and Applications. CRC Press, Boca Raton, pp. 619–660.

Tang, L., Hamarneh, G., Celler, A., 2008. Validation of mutual information-based registration of CT and bone SPECT images in dual-isotope studies. Comput. Meth. Prog. Biol. 92, 173–185. http://dx.doi.org/10.1016/j.cmpb.2008.06.003.

Tasko, S.M., Greilick, K., 2010. Acoustic and articulatory features of diphthong production: a speech clarity study. J. Speech Lang. Hear. Res. 53, 84–99. http://dx.doi.org/10.1044/1092-4388(2009/08-0124).

Traunmüller, H., Öhrström, N., 2007. Audiovisual perception of openness and lip rounding in front vowels. J. Phon. 35, 244–258. http://dx.doi.org/10.1016/j.wocn.2006.03.002.

Uricar, M., Franc, V., Hlavac, V. 2012. Detector of facial landmarks learned by the structured output SVM. In: VISAPP '12: Proceedings of the 7th International Conference on Computer Vision Theory and Applications.

Wang, Y., Behne, D.M., Jiang, H., 2008. Linguistic experience and audio-visual perception of non-native fricatives. J. Acoust. Soc. Am. 124, 1716–1726. http://dx.doi.org/10.1121/1.2956483.

Wang, Y., Behne, D.M., Jiang, H., 2009. Influence of native language phonetic system on audio-visual speech perception. J. Phon. 37, 344–356. http://dx.doi.org/10.1016/j.wocn.2009.04.002.

Wiskott, L., Fellous, J.M., Kruger, N., von der Malsburg, C., 1997. Face recognition by elastic bunch graph. IEEE Trans. Pattern Anal. Mach. Intell. 7, 775–779.

Yehia, H.C., Kuratate, T., Vatikiotis-Bateson, E., 2002. Linking facial animation, head motion and speech acoustics. J. Phon. 30, 555–568. http://dx.doi.org/10.1006/jpho.2002.0165.

Zhang, C., Wang, P., 2000. A new method of color image segmentation based on intensity and hue clustering. In: Proceedings of the 15th International Conference on Pattern Recognition, pp. 613–616, http://dx.doi.org/10.1109/ICPR.2000.903620.

Zhu, X., Ramanan, D., 2012. Face detection, pose estimation, and landmark localization in the wild. in: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2879–2886, http://dx.doi.org/10.1109/CVPR.2012.6248014.