# EXAMINING VISIBLE ARTICULATORY FEATURES IN CLEAR AND CONVERSATIONAL SPEECH

Lisa Tang[1], Beverly Hannah[2], Allard Jongman[3], Joan Sereno[3], Yue Wang[2], Ghassan Hamarneh[1]

1. Medical Image Analysis Lab, School of Computing Science, Simon Fraser University, Canada
2. Language and Brain Lab, Department of Linguistics, Simon Fraser University, Canada
3. KU Phonetics and Psycholinguistics Lab, Department of Linguistics, University of Kansas
lisat@sfu.ca, beverlyw@sfu.ca, jongman@ku.edu, sereno@ku.edu, yuew@sfu.ca, hamarneh@sfu.ca

## ABSTRACT

This study investigated the relationship between clear and conversational speech styles and motion of visible articulators. Using state-of-the-art computer-vision and image processing techniques, we examined front and side view videos of 18 native English speakers' faces while they recited six English words containing various vowels (keyed, kid, cod, cud, cooed, could) and extracted measurements corresponding to the lip and jaw movements. Significant effects were found for style, gender, and saliency of visual speech cues. Clear speech exhibited longer vowel duration and more vertical lip stretching and jaw movement for all vowels, more horizontal lip stretching for front vowels, and a greater degree of lip protrusion for rounded vowels. Additionally, greater articulatory movements were found for male than female speakers in clear speech. These articulatory movement data demonstrate that speakers modify their speech productions in response to communicative needs in different speech contexts.

**Keywords**: articulation, clear speech, English vowels, computational methods, landmark detection

## 1. INTRODUCTION

The movements of facial articulatory features contribute to the myriad cues generated during speech [5, 10, 21]. Indeed, auditory and visual (AV) speech perception has been demonstrated to be superior to auditory-only perception, presumably due to the additional stream of linguistic information available in the visible articulatory movements of the speaker's lips, teeth, tongue and facial movements [16, 19]. These AV cues are efficiently integrated and perceptually weighted based on the ambiguity and reliability of the information from each stream [14, 22].

In addition to visual cues, to increase their intelligibility, speakers may produce clear speech, a hyperarticulated speech style, relative to the natural conversational speech style, in response to the communicative needs of perceivers [12, 13, 18, 21]. Acoustic and perceptual studies of English vowels in clear speech show that expanded vowel space and increased duration, presumably due to more extreme articulatory movements involving a higher degree of mouth opening and jaw lowering, are the factors that contribute the most to increased intelligibility over conversational speech [1, 2, 3, 4].

Research has also demonstrated that visual information and clear speech are complementary and not merely redundant sources of additional information in improving intelligibility [9]. The few existing studies examining visual articulatory effects in clear speech use kinematic measures, and show positive correlations of articulation, acoustics, and intelligibility in clear speech effects [10, 11, 21], where increased movement of the tongue, jaw, and mouth (lip rounding) during clear speech translated to increased intelligibility.

There are few reports on visible articulatory movements of vowels in clear speech. Previous studies focused on visual clear speech effects at the level of sentences or a single vowel (diphthong). The present study aims to characterize the differences in visible articulatory features of a representative set of English vowels (/i, I, ɑ, ʌ, u, ʊ/) in /kVd/ word tokens produced in clear and conversational speech styles. This research uses advanced computerized facial detection and image processing techniques to extract the articulatory movement information from the speaker faces as captured by video recordings. This technique employs training data (videos with human-identified landmark annotations) to build a mathematical model that predicts landmark locations in new, unseen images [24]. The technique thus differs from the previous studies in that no physical markers were placed on the speakers, thus facilitating more natural speech productions as well as concurrent research on the perceptual correlates of the articulatory measurements. We hypothesize that, compared to conversational speech, vowels produced in clear speech involve greater motion of visible articulators. In particular, we expect a greater degree of lip spreading (for unrounded front vowels), lip rounding (for rounded vowels), and jaw lowering (for low vowels) [10, 21]. In addition, we expect the difference between conversational and clear speech to be greater for tense vowels which involve greater articulatory movement than lax vowels. While lip-tracking and face-detection algorithms have been applied to various computer-vision problems, the present study is the first to apply them to speech production [7].

## 2. METHODS

### 2.1. Experimental setup & data acquisition

Eighteen native speakers of Western Canadian English (10 females) aged 17-30 were recruited. The speakers reported no hearing or speech impairments.

Three vowel pairs /i-ɪ/, /ɑ-ʌ/, /u-ʊ/ differing in articulatory features (involving lip spreading, jaw lowering, and lip rounding) were the target vowels, with the tense vowels /i, ɑ, u/ corresponding to a higher degree of visual salience than the lax vowels /ɪ, ʌ, ʊ/. These vowels were embedded in /kVd/ contexts resulting in English target words "keyed", "kid", "cod", "cud", "cooed", and "could".

The elicitation of clear and conversational speech followed the procedures developed by [13], where a simulated interactive computer seemingly attempted to perceive and recognize the tokens produced by a speaker. The software would systematically make wrong guesses due to 'perception errors'. 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). Recordings were made in a sound-attenuated booth. Front view video recordings were captured with one camera and left side views of the speaker face were captured with a second camera. Each token was evaluated by two native English speakers to ensure that the speakers produced the intended vowels.

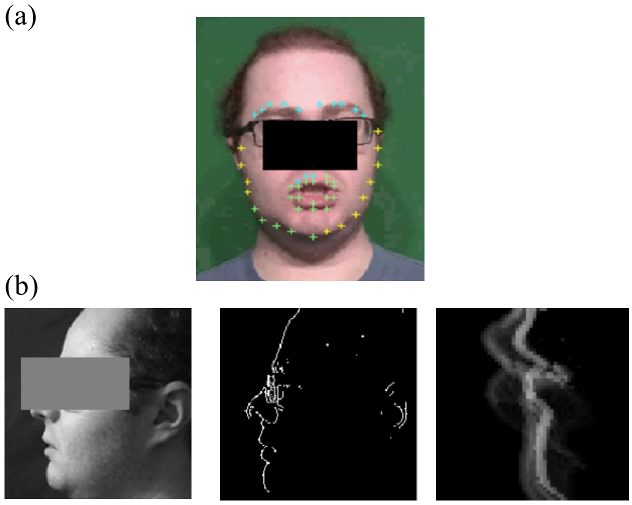
### 2.2. Video analysis

The raw videos were first semi-automatically processed into annotated segments (at token-level) with MATLAB [15], using the audio channel of the recorded videos [6]. 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 (e.g. peak horizontal lip stretch, amount of lip-protrusions) were computed based on the detected facial landmark positions.

For front-view videos (Figure 1a), which we used to examine lip and jaw movements, frontal view facial landmarks were extracted as follows: 1) the face detector of [24] was used to localize facial landmarks in each frame; 2) estimates of the lip landmarks from step 1 were further refined based on color intensity gradients; and 3) articulatory measurements characterizing facial movements were computed using the detected landmark positions. More specifically, in step 2, the final lip landmark positions were computed as the location of maximal change in 1-dimensional intensity profile drawn along each lip landmark. To account for differences in head size of speakers, we also estimated horizontal and vertical scale factors of the speaker's head by computing, respectively, the interpupillary distance (IPD, or horizontal distance, HD) and eye-to-nose-tip distance (vertical distance, VD), which is defined as the distance from the nose-tip to its perpendicular projection on the pupil-line. We then extracted the following articulatory measurements using the positions of the extracted facial landmarks and the estimated factors VD and HD: peak horizontal lip stretch, peak vertical lip stretch, peak lip rounding, and peak vertical displacement of jawline. These articulatory measurements are not expressed in physical units[i] but rather expressed as a fraction of speaker's head, thereby facilitating scale normalization across speakers.

Using side-view videos (Figure 1b), we also examined articulatory features relating to lip protrusions for "cooed" and "could" that involve the rounded vowels. As the side-videos show a limited set of facial features[ii], a different computerized analysis procedure was developed. First, for *scale-normalization*, we computed for each video token a *feature image* (FI), defined as the pixel-wise average edge-strength of each video frame, which, at a high level, summarizes lip movement across the duration of each token. Then, the FI of a randomly selected video token was chosen as reference, denoted as VRef. Next, to spatially normalize the scale differences between video tokens, linear image registration (LIR) was performed to resolve the similarity transform that would align the FI of each token to VRef [20]. The spatially aligned FIs were then trimmed to center around the lip. Next, we measure the amount of lip protrusion by extracting summary statistics about the FIs that would quantify the visual differences between the aligned FIs. Specifically, we computed an image dissimilarity measure between the registered FI and VRef that would quantify differences due to lip deformations (as LIR would have removed global misalignments). We explored various dissimilarity measures [20] and based on empirical experiments, we chose the mean absolute difference (MAD) measure to characterize lip protrusions as this measure generally yielded higher values in tokens with amplified lip movements than those without such amplifications.

**Figure 1**. (a) example front-view video frame and corresponding facial landmarks; (b) 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). The black/grey bars are to protect the speaker’s privacy.



## 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.

For each of the front-view measurements, a series of 2x2x2 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. Firstly, as hypothesized, there is a significant main effect of Style: in horizontal [F(1,765)=21.5, p<.0005] and vertical [F(1,765)=51.0, p<.0005] stretches for "keyed/kid", in vertical stretch [F(1,765)=24.2, p<.0005] and jaw displacement [F(1,765)=8.6, p<.0034] for "cod/cud", and in roundedness [F(1,655)=4.82, p<.0284], vertical stretch [F(1,655)=21.7, p<.0005] and jaw displacement [F(1,655)=6.6, p<.0104] for "cooed/could". As shown in Table 1, for each of these significant differences in style, the degree of movements in clear speech is greater than in conversational speech. Additionally, for each word, the duration in clear speech was longer than in conversational speech, as expected [p<.05]. For the main effect of Tensity, tense vowels show longer duration and greater degrees 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”, in all of the measurements of “cod/cud” and “cooed/could”, with overall greater degree of movements in male than female productions (p<.05).

**Table 1.** Comparisons between conversational (Cn) and clear (Cl) speech styles for /i-ɪ/: keyed/kid; /ɑ-ʌ/: cod/cud; /u-ʊ/: cooed/could by male (M) and female (F) speakers. The values are the mean amount of vertical lip stretch, horizontal lip stretch, jaw displacement, and lip rounding (with standard deviations in parentheses). Shaded cells indicate statistically significant style effects (p<.05).

| | | Vertical | | Horizontal | | Jaw | | Rounding | |
|---|---|---|---|---|---|---|---|---|---|
| | | M | F | M | F | M | F | M | F |
| /i-ɪ/ | Cn | **1.17 (.19)** | **0.98 (.17)** | **0.93 (.10)** | 0.82 (.09) | **0.10 (.09)** | 0.08 (.07) | 0.76 (.10) | 0.75 (.10) |
| | Cl | **1.35 (.33)** | **1.05 (.21)** | **0.99 (.12)** | 0.84 (.09) | **0.13 (.10)** | 0.12 (.36) | 0.75 (.10) | 0.76 (.09) |
| /ɑ-ʌ/ | Cn | **1.23 (.29)** | **1.05 (.17)** | 0.89 (.09) | 0.81 (.08) | **0.14 (.10)** | 0.11 (.08) | 0.77 (.07) | 0.75 (.08) |
| | Cl | **1.36 (.34)** | **1.10 (.18)** | 0.89 (.08) | 0.80 (.07) | **0.18 (.12)** | 0.12 (.08) | 0.77 (.06) | 0.75 (.06) |
| /u-ʊ/ | Cn | **1.10 (.21)** | **0.94 (.17)** | 0.89 (.10) | 0.79 (.07) | **0.09 (.08)** | 0.07 (.06) | **0.75 (.08)** | 0.76 (.09) |
| | Cl | **1.21 (.28)** | **1.00 (.20)** | 0.88 (.08) | 0.80 (.07) | **0.12 (.10)** | 0.08 (.07) | **0.72 (.08)** | 0.76 (.08) |

The statistically significant interactions mostly involved style and gender. Post-hoc analyses were further done to examine the effects of style per gender group for each pair of words using a series of one-way ANOVAs. For keyed/kid, the vertical lip stretch is greater in clear (M=1.35) than conversational speech (M=1.17) [F(1, 352)=36.5, p<0.001] in males. To a lesser degree, in females, the vertical lip stretch is also greater in clear speech (M=1.05) than conversational speech (M=0.98) [F(1, 410)=15.1, p<.0.001]. Horizontal lip stretch is also greater in clear speech (M=0.99) than in conversational (M=0.93) [F(1, 352)=19.7, p<0.001] for males, but the difference is not significant for females (M=0.84 vs. M=0.82) [F(1, 410)=3.71, p=0.055]. For cod/cud, the vertical lip stretch is greater in clear (M=1.36) than conversational speech (M=1.23) [F(1, 360)=15.5, p<0.001] in males. To a lesser degree, in females, the vertical lip stretch is also greater in clear speech (M=1.05) than conversational speech (M=1.01) [F(1, 402)=8.36, p=0.004]. In addition, for males, the jaw movement was greater in clear than in conversational speech (M=0.18 vs. M=0.14) [F(1,360)=9.21, p=0.003], but no such difference was observed in females. For cooed/could, the vertical lip stretch is greater in clear (M=1.21) than in conversational speech (M=1.1) [F(1,298)=13.7, p<0.001] in males. To a lesser degree, in females, the vertical lip stretch is also greater in clear (M=1.0) than in conversational speech (M=0.94) [F(1, 354)=7.48, p=0.007]. Additionally, males employed greater jaw movement in clear than in conversational speech (M=0.12 vs. M=0.09) [F(1,298)=7.15, p<0.001], but no such difference was observed in females.

To test the hypothesis that differences in style can be observed in terms of lip protrusions for the rounded vowels "cooed" and "could", a 2x2x2 ANOVA was performed on the extracted side-view measurements. The results show a significant main effect of Style and a significant Style and Gender interaction. Subsequent one-way ANOVAs for each gender with Style as the within-subject factor revealed a greater lip protrusion in clear speech (M=0.105) than in conversational speech for males (M=0.084) [F(1,309)=40.64, p<0.0001]. To a lesser degree, a greater degree of lip protrusion for clear (M=0.065) versus conversational style (M=0.052) in the female speakers was also observed [F(1,402)=26.22, p<0.0001].

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

## 4. DISCUSSION

This study makes use of dual-view video sequences to examine articulatory features between clear and conversational speech, involving a representative set of vowels embedded in English monosyllabic words. Our overall results support and may be positively correlated with previous findings of the acoustic features of vowels in clear speech [e.g., 1, 2, 4] in that expanded acoustic vowel space and more peripheral formant frequencies in clear speech may be attributed to more extreme and greater degrees of articulatory movements, in terms of vertical lip and jaw lowering, horizontal lip stretches, and lip-protrusion. The finding of clear speech effects attributable to vertical lip and jaw movements across words is consistent with the previous claim that the chin 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 (which gives rise to horizontal lip movement) [23]. However, the horizontal lip movement does show clear speech effects in the production of the vowels that involve horizontal lip spreading, i.e., [i-ɪ] in keyed/kid. Finally, the side-view video captured the greater degree of lip-protrusion for the rounded vowels [u-ʊ] in cooed/could.

The current results from video imaging analyses are also in line with previous findings using kinematic measures reporting increased movement of the tongue, jaw, and mouth during clear speech [10, 11, 21]. With the exception of a study on a single diphthong [21], most of the kinematic studies conducted articulatory analyses based on longer utterances than individual segments [e.g., 10]. The present results show that comparable and even more subtle visible articulatory movements can be captured using video imaging processing, 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 the computerized lip-tracking and face-detection algorithms [7] to the study of speech production, acoustics, and perception.

The current results also reveal a 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 interaction between gender and clear speech effects both from acoustic and articulatory measures [10, 21, 22]. It would be interesting for further research to evaluate if these gender differences can be captured in perception to affect the intelligibility of clear versus conversational speech. Additionally, although the current study revealed an overall greater articulatory movement for tense compared to lax vowels, it did not find any interaction between vowel tensity and clear speech effects. Acoustically, tense vowels demonstrate greater conversational-to-clear speech modifications than lax vowels [17]. The factor of visual saliency in audio-visual speech processing has been investigated with consonants [8]. Further research may be conducted to examine the effects of visual saliency and vowel articulation in clear speech.

## 5. ACKNOWLEDGMENTS

We thank Anthony Chor and Keith Leung from the Language and Brain Lab at Simon Fraser University for their assistance. We also thank Dr. Kazumi Maniwa for sharing the elicitation method used in Maniwa et al. (2009). This research has been funded by a SSHRC Insight Grant (435-2012-1641).

## 5. REFERENCES

[1] Bond, Z. S., Moore, T. J. 1994. A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech Commun.* 14, 325–337.

[2] 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.

[3] 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.

[4] 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.

[5] Gagné, J. P., Rochette, A. J., Charest, M. 2002. Auditory, visual and audiovisual clear speech. *Speech Commun.* 37, 213-230.

[6] Giannakopoulos, T., Petridis, S., Perantonis, S. 2010. User-driven recognition of audio events in news videos. *Proc. 5th Intl. Workshop on Semantic Media Adaptation and Personalization (SMAP)* Limassol, 44-49.

[7] Göcke, R., Asthana, A. (2008). A comparative study of 2D and 3D lip tracking methods for AV ASR. *Proc. AVSP* Moreton Island, 235-240.

[8] 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.

[9] Helfer, K. S. 1997. Auditory and auditory-visual perception of clear and conversational speech. *J. Speech Lang. Hear. Res.* 40, 432-443.

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

[11] Kim, J., Sironic, A., Davis, C. 2011. Hearing speech in noise: Seeing a loud talker is better. *Perception* 40, 853-862.

[12] Lu, Y., Cooke, M. 2008. Speech production modifications produced by competing talkers, babble, and stationary noise. *J. Acoust. Soc. Am*. 124, 3261-3275.

[13] Maniwa, K., Jongman, A., Wade, T. 2009. Acoustic characteristics of clearly spoken English fricatives. *J. Acoust. Soc. Am.* 125, 3962-3973.

[14] Massaro, D. W. 2004. From Multisensory Integration to Talking Heads and Language Learning. In: Calvert, G. A., Spence, C., Stein, B. E. (eds), *The handbook of multisensory processes*. Cambridge: MIT Press, 153-176.

[15] MATLAB Release 2013b, The MathWorks, Inc., Natick, MA.

[16] Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., Vatikiotis-Bateson, E. 2004. Visual prosody and speech intelligibility head movement improves auditory speech perception. *Psychol. Sci.* 15, 133-137.

[17] 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.

[18] Smiljanić, R., Bradlow, A. R. 2009. Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Lang. Linguist. Compass* 3, 236-264.

[19] Sumby, W. H., Pollack, I. 1954. Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 2, 212-215.

[20] Tang, L. Y., Hamarneh, G. 2013. Medical Image Registration: A Review. In: Farncombe, T., Iniewski, K. (eds), *Medical Imaging: Technology and Applications*, Boca Raton: CRC Press, 619-660.

[21] 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.

[22] Traunmüller, H., Öhrström, N. 2007. Audiovisual perception of openness and lip rounding in front vowels. *J. Phon.* 35, 244–258.

[23] Yehia, H. C., Kuratate, T., Vatikiotis-Bateson, E. 2002. Linking facial animation, head motion and speech acoustics. *J. Phon.* 30, 555-568.

[24] Zhu, X., Ramanan, D. 2012. Face detection, pose estimation, and landmark localization in the wild. *Proc. 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* Providence, 2879-2886.

[i] Nor can they be, as the physical size of each video frame pixel was not measured at the time of acquisition (which would require careful camera calibration procedure).

[ii] As only one eye is visible on the side, we could not estimate HD and VD for spatial normalization that would adjust for spatial scale differences across tokens.