Investigating visual correlates produced during voicing of Canadian English stops

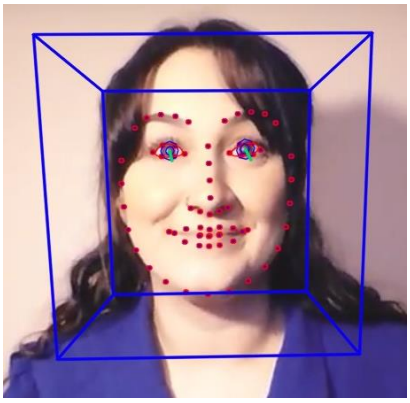
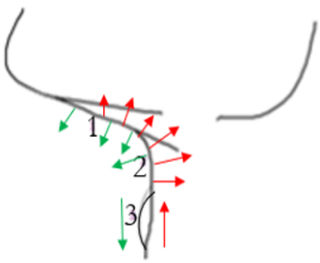
Theresa Rabideau[1], and Suzy Ahn[1]
[1]*University of Ottawa (Canada)*

Background. Decades of research on visual information for speech has demonstrated its informativeness on speech perception [1, 2, 3] However, it remains unclear which specific visual cues are spatially and temporally correlated to certain features of speech segments, such as obstruent voicing. Research primarily concentrates on the jaw and lips, leaving unclear the visual cues produced during speech beyond this limited facial region. Although the jaw and lips are most often the focus of visual speech research, stronger McGurk effects and improved performance on other speech reading tasks has been shown when a broader facial area is visible [4]. Previous research has also suggested that visual cues from neck [5], eyebrows [6], and head movements [7, 8], may also contribute to perception of different speech segments or suprasegments. This raises the question of what visual correlates are presents during production and what cues speakers rely on for perception.

Research question. English voiced stops in initial position often lack phonetic voicing during closure, while intervocalic voiced stops exhibit voicing [9]. Despite the lack of consistency in the amount of 'voicing' in English voiced stops, many studies have shown that articulatory adjustments such as larynx lowering [10], pharyngeal cavity expansion [11, 10, 12] and tongue root advancement [13, 12] are employed during the production of voiced stops compared to voiceless stops. These differences in articulation may have an impact on the visible anatomy during production; thus, potentially producing different visual cues for stop categorization. In the current study, we explore the visual correlates of Canadian voiced stops, paying special attention to the throat, chin, and neck areas which have been understudied by the previous literature, to determine which facial movements correlate to the production of voiced (whether produced with actual voicing or not) and voiceless stops in Canadian English.

Method. Video and audio recordings were taken of six Canadian English speakers (1 M; 5 F) during their productions of real words with voiced and voiceless stops in utterance-initial, phrase medial post-vocalic (e.g., a beak) and sentence-medial post-vocalic positions. Two videos were collected simultaneously: one for the front view and another for the side view. Stimuli consisted of 24 CVC words, consisting of oral and nasal stops across three POA: labial (/p/, /b/, /m/), coronal (/t/, /d/, /n/) and velar (/k/, /g/). Acoustic measures included f0, voicing during closure, and release burst duration. We automatically tracked 30 facial action units using facial recognition technology [14] and manually tracked (coded in ELAN [15]) 4 areas of interest (see Figure 1).

Results and discussion. The preliminary data shows expanding movement in the submental triangle (Figure 2) and throat during the production of voiced stops compared to voiceless stops. This finding supports previous literature that shows tongue body lowering and larynx lowering in the production of English voiced stops [12]. Where visible, coding of the Adam's apple showed the most movement during the production of nasals, then voiced stops and the least movement during voiceless stops, which may be correlated to f0. The comparison between utterance initial stops and post-vocalic stops shows that certain visual movement correlates may be related to phonological voicing categorization irrespective of actual voicing during closure while others reflect phonetic voicing reality. The preliminary results indicate potential visual correlates distinguishing voiced stops from voiceless stops in Canadian English, which should be confirmed by further analysis of data from more speakers. Additionally, a follow-up perception study should be conducted to ascertain how these visual correlates are utilized in speech perception.

AOI	View	Area of Interest
1	Side	Under chin
2	Side	Throat
3	Side	Thyroid cartilage
4	Front	Throat

Figure 1. Facial action units tracked by facial recognition technology (left) and areas of interest manually coded in ELAN (right).

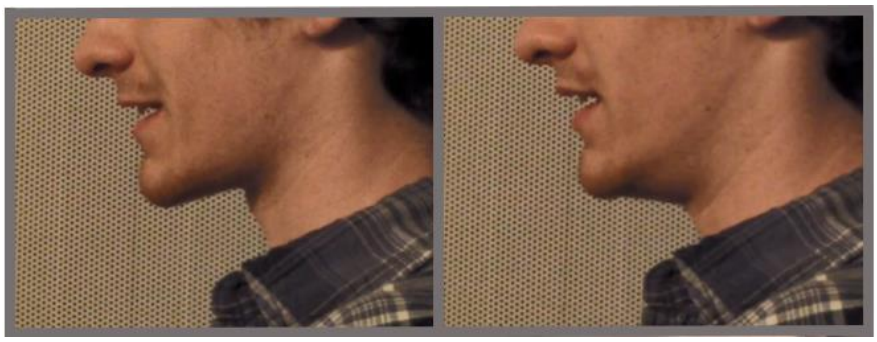


Figure 2. *An example of voiceless-voiced token pair 'keep' (left) and 'geek' (right)*

References.

[1] Cho, S. et al. (2020). Multi-modal cross-linguistic perception of fricatives in clear speech. *JASA, 147*(4), 2609-2624.

[2] Kawase, S. et al. (2014). The influence of visual speech information on the intelligibility of English consonants produced by non-native speakers. *JASA, 136*(3), 1352-1362.

[3] MacDonald, J., and McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics, 24(3), 253-257.*

[4] Smeele, P. et al. (1995). Investigating the role of specific facial information in audio-visual speech perception. *JASA 98*, 2983.

[5] Chen, T., & Massaro, D. (2008). Seeing pitch: Visual information for lexical tones of Mandarin-Chinese. *JASA 123*(4), 2356-2366.

[6] Granström, B. et al. (1999). Prosodic cues in multimodal speech perception. *Proc. of 14th ICPhS*, San Francisco.

[7] Graf, H. P. et al. (2002). Visual prosody: Facial movements accompanying speech. *IEEE International Conference on Automatic Face and Gesture Recognition, 2002*, 396-401.

[8] Munhall, K. G. et al. (2004). Visual Prosody and Speech Intelligibility: Head Movement Improves Auditory Speech Perception. *Psychological Science, 15*(2), 133–137.

[9] Keating, P. (1984). Phonetic and phonological representation of stop consonant voicing. *Language, 60*(2), 286-319.

[10] Perkell, J. S. (1969). *Physiology of speech production: Results and implications of a quantitative cineradiographic study*. Cambridge, MA: MIT Press.

[11] Kent, R. D. & Moll, K. L. (1969). Vocal tract characteristics of the stop consonants. *JASA 46*, 1549-1555.

[12] Westbury, J. R. (1983). Enlargement of the supraglottal cavity and its relation to stop consonant voicing. *JASA 73*, 1322-1336

[13] Ahn, S. (2018). The role of tongue position in laryngeal contrasts: An ultrasound study of English and Brazilian Portuguese. *J. Phon, 71*(4), 451-467.

[14] Baltrušaitis, T. et al. (2018). OpenFace 2.0: Facial Behavior Analysis Toolkit. *IEEE International Conference on Automatic Face and Gesture Recognition.*

[15] ELAN (Version 6.4) [Computer software]. (2022). Nijmegen: Max Planck Institute for Psycholinguistics.