Audiovisual enhancement in clear speech production of English laterals

Jonathan Havenhill, Ming Liu, Shuang Zheng & Jonah Lack

The University of Hong Kong
jhavenhill@hku.hk, {mingliu_7, ivyzheng, jlack}@connect.hku.hk

The role of auditory perceptibility in the maintenance and enhancement of phonological contrast is well established [1]. In clear speech styles, speakers are observed to increase the acoustic distance between contrastive phones to improve auditory perceptibility [2]. Non-auditory perceptual cues (notably vision) also influence speech perception [3] and have long been known to improve perceptibility under noisy conditions [4]. As such, clear speech may also involve enhancement of visible articulations, e.g., lip rounding [5]. The finding that blind speakers show less rounding than sighted speakers in clear speech [6] suggests that such modifications are at least partly mediated by visual factors. Yet while increasing the degree of lip protrusion may improve visual perceptibility, doing so simultaneously increases acoustic distance by lowering F2. In many cases it is therefore difficult to determine the extent to which speakers are optimizing their speech for auditory-acoustic and/or visual-articulatory cues.

To address this question, we investigate the production of laterals and other coronal consonants in normal and listener-oriented clear speech. In English, visibly articulated variants of /l/ have been noted to occur in lip syncing [7] and can also be observed in other clear or emphatic speech styles. Such variants have not been systematically investigated, however, so their frequency, phonetic properties, and phonological distribution, as well as their communicative function, remain unknown. The goal of this study is to examine how speakers use visible articulatory gestures in producing English laterals during clear speech, to test the hypothesis that some articulatory gestures serve a visuoperceptual rather than auditory enhancement function.

18 adult native English speakers (8 men, 10 women, mean age 28.2) participated in a two-part speech production experiment. Speakers were asked to produce a pseudo-randomly ordered list of 96 words containing /l n d θ/ in syllables with primary stress. Target syllables were balanced for vowel height (/i/ vs. /æ/) and position of the target consonant (onset, coda, monomorphemic intervocalic, pre-boundary intervocalic) positions. In the first block, speakers produced three repetitions of each word in citation form while seated alone in a sound-attenuated booth. Audio was recorded using an Earthworks Ethos condenser microphone, while high-speed (120 fps) video was recorded using a Sony DSC RX10-IV camera. In the second block, speakers repeated the wordlist in a cooperative game with a native English-speaking listener. Audio and video were recorded and simultaneously transmitted to the listener (seated in another room) over Zoom. Ambient recordings from a noisy bar were overlaid at −9dB SNR, such that the speaker's speech was effectively unintelligible. The speaker was instructed to produce three repetitions of each word as clearly as possible, while the listener attempted to guess the word after each repetition. The speaker clearly heard the listener's guesses with the same ambient noise at +3dB SNR.

Each token was visually coded according to its articulatory configuration. Non-visible (NV), dental (D), visible alveolar (V), interdental (ID), and linguolabial (LL) configurations were observed, as shown in Figure 1. Frequency of each variant is provided in Figure 2a. The typical realization of /θ/ in normal speech was interdental (69.2%) or dental (19.9%). For /d/ and /n/, over 98% of tokens were realized with non-visible articulations in normal speech, with only a handful of dental and visible alveolar tokens in the clear speech task. In contrast, /l/ was produced with both dental (11.9%) and interdental (5.4%) articulations, even in normal speech. Multinomial logistic regression analysis indicates significantly higher rates in clear speech of dental and interdental variants for /l/ ($p < 0.001$), dental and visible variants of /n/ ($p < 0.001$) and interdental variants of /θ/ ($p < 0.001$). Linguolabial variants of both /θ/ and /l/ also occur in clear speech, albeit rarely, but were never observed in normal speech. Visibly articulated variants of /l/ occur in all syllable positions, as in Figure 2b, but with some variability by vowel height.

Acoustic analysis indicates that visibly articulated variants of /l/ (particularly ID and LL) do not consistently preserve /l/-like acoustic features. As seen in Figure 3, interdental [l̪] not only lacks the characteristic resonance patterns of alveolar [l] (3a), but also exhibits frication, particularly toward the vowel onset (3b), yielding an audibly [ð]-like sound. This finding suggests that listener-oriented speech does not necessarily prioritize auditory perceptibility.

Rather, speakers may choose to provide listeners with a direct visual cue of a segment's articulatory/gestural properties, e.g., tongue lengthening and narrowing for /l/. However, this strategy is not available for all segments, suggesting that speakers recruit visually exaggerated gestures when such gestures a) also occur in normal speech and b) are visually distinct from similar, contrasting segments. These results call into question the hypothesis that the objects of speech perception are primarily auditory and suggest that speakers may prioritize visual perceptibility, consistent with the inherently multimodal nature of speech perception [8]. Theories of contrast and representation must therefore incorporate both auditory and non-auditory perceptibility.

Figure 1: From left to right: dental (D), visible alveolar (V), interdental (ID), and linguolabial (LL) productions of /l/ in clear speech task.

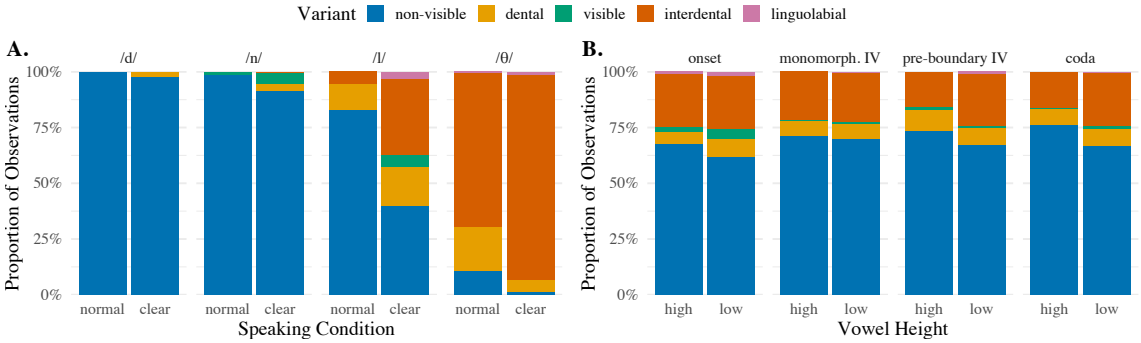

Figure 2: (A) Variants of /d n l θ/ observed in normal and clear speaking tasks. (B) Variants of /l/ by syllable position and vowel quality.

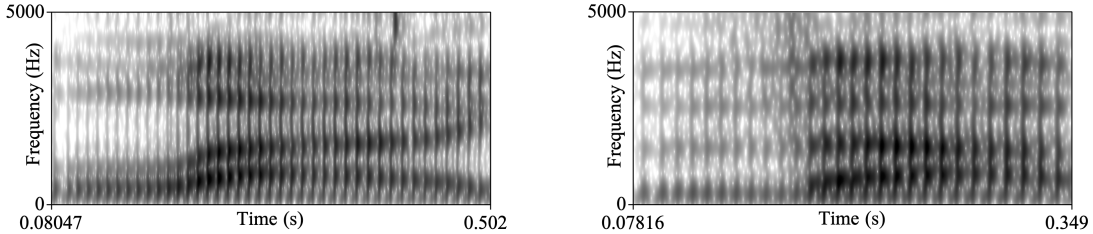

Figure 3: Spectrograms for a) non-visible [læg] (left) and b) interdental [l̪æp] in clear speech.

[1] R. L. Diehl and K. R. Kluender. "On the objects of speech perception". In: *Ecol. Psychol.* 1.2 (1989), pp. 121–144. **[2]** R. Smiljanić and A. R. Bradlow. "Speaking and hearing clearly: Talker and listener factors in speaking style changes". In: *Lang. Linguist. Compass* 3.1 (2009), pp. 236–264. **[3]** J.-P. Gagné, A.-J. Rochette, and M. Charest. "Auditory, visual and audiovisual clear speech". In: *Speech Commun.* 37.3-4 (2002), pp. 213–230. **[4]** W. H. Sumby and I. Pollack. "Visual Contribution to Speech Intelligibility in Noise". In: *JASA* 26.2 (1954), pp. 212–215. **[5]** J. Havenhill. "Constraints on Articulatory Variability: Audiovisual Perception of Lip Rounding". PhD thesis. Georgetown Univ., 2018. **[6]** L. Ménard et al. "Speaking Clearly for the Blind: Acoustic and Articulatory Correlates of Speaking Conditions in Sighted and Congenitally Blind Speakers". In: *PLOS ONE* 11.9 (2016), e0160088. **[7]** M. Liberman. *Apico-labials in English*. Language Log, Accessed: 2023-02-10. 2010. **[8]** L. D. Rosenblum. "Primacy of Multimodal Speech Perception". In: *The Handbook of Speech Perception*. Ed. by D. B. Pisoni and R. E. Remez. Oxford: Blackwell, 2005, pp. 51–78.