TESTING SYMMETRY OF TEMPORAL WINDOW OF INTEGRATION IN VIBROTACTILE AND AUDITORY SPEECH INFORMATION ON PHONEME PERCEPTION

Tzu Hsu Chu* [1,2], David Marino † [1,2], Hannah Elbaggari‡ [1,2], Bryan Gick § [1], and Karon MacLean ¶ [2]
[1]University of British Columbia Department of Linguistics
[2]University of British Columbia Department of Computer Science

1 Introduction

We show that temporal asynchronies between vibrotactile and auditory speech information follow a symmetrical distribution. Speech perception is more than a unimodal process: it requires concurrent integration of multiple sensory modalities. Auditory and visual modalities have been the main focus of many on multimodal speech perception studies, as seen in the well-known McGurk effect [1], where incongruent presentation of auditory and visual speech stimuli resulted in an integrated illusion. Multi-modal sensory signals don't need to be synchronous for effective integration [2]. Testing temporal constraints show that the illusion is still maintained when visual stimuli precedes audio stimuli for up to 180ms, but when audio stimuli precedes visual stimuli the window for integration shrinks to 60ms [2]. This explanation for asymmetry is resulting from a difference in signal speeds in the natural world: light travels faster than sound, so humans are more experienced in perceiving events where visual information is received before the audio information.

The tactile modality also contributes information for enhanced speech perception–syllables heard alongside puffs of air on the skin were more likely to be perceived as aspirated (e.g. hearing /ba/ as /pa/) [3]. A follow up study [4] demonstrated a temporal window of integration between -50ms and 200ms. This asymmetry can again be explained by a difference in relative signal speeds: sound travels faster than airflow, so integration has a larger window of opportunity in events where audio information precedes tactile information.

However, not all cross-modal integration can be explained with a difference in signal speed. Vibrotactile cues often accompany acoustic cues as experienced in the laryngeal vibrations felt during voiced speech, with little to no difference in relative signal speed (i.e. no concomitant perceptual asymmetry).

Through establishing temporal window for such vibrotactile devices, implications of cross-modal integration can be used in practical implementation to aid speech intelligibility in real time application.

*sophychu@cs.ubc.ca
†damarino@cs.ubc.ca
‡hre@cs.ubc.ca
§gick@mail.ubc.ca
¶maclean@cs.ubc.ca

2 Methods

A pilot study [5] was conducted to establish effect of vibrotactile stimulation on speech intelligibility. Participants discriminated the content of speech in noise, while receiving vibrotactile stimuli through voice-coiled transducer. Device placement (neck vs hand), phonological contrast (fricatives, stops, vowel heights), and vibration styles were manipulated. Results showed greatest enhancement in speech perception when the amplitude of the vibrations were coupled to the amplitude envelope of voiced fricatives. The present study uses the ground truth established by the pilot as a framework to investigate the effects of temporal offsets on participant accuracy scores when discerning speech in noise.

2.1 Participants

We recruited 26 students from the University of British Columbia of normal-hearing, normal or corrected eye-vision, and have no previous experience with the tactile devices used in this experiment. All participants were compensated with course credit or $10 for their time.

2.2 Stimuli Delivery

Vibrations were administered through a Tectonic Element Audio (TEAX12C02-8/RH) linear resonant acutator (LRA) that was held between the index finger and thumb. The vibrotactile waveforms were procedurally generated from speech during the experiment–the vibrations were designed to mimic the laryngeal vibrations normally felt during voiced speech: they were only present on voicing, and the amplitude of the vibrations were coupled to the amplitude envelope of the speech. Auditory speech and noise were delivered through AKG over-the-ear headphones.

2.3 Stimuli

Each participant underwent 218 trials, where different phonemic contrasts and temporal offsets were randomly administered. Speech was recorded by a female identifying native speaker of English with a DR40 TASCAM hand held linear PCM digital recorder, audio file volumes were normalized before generating vibrations.

Phonemic contrasts: For consistency purposes and prevention of stimuli-related confounds, pre-vocalic voiced and unvoiced minimal pairs of fricatives (/va/ vs. /fa/, /za/ vs. /sa/) and stops (/pa/ vs /ba/) were used.There were 8 recordings for each minimal pair.

Temporal offsets: Temporal offsets of 0ms (synchronous), $\pm 50\ ms$, $\pm 100\ ms$, $\pm 200\ ms$, $\pm 300\ ms$ were administered. Offsets were given positive value when audio signals precede vibrotactile signals; and negative value when vibrotactile signals precede audio signals.

2.4 General Procedure

(1) Calibration Phase: Signal-to-noise ratio of target audio stimuli and background babble track were adjusted until above chance accuracy. Calibration responses were not recorded, except for volume and calibration accuracy.

(2) Testing Phase: Participants sat in front of computer monitor with over-the-ear headphones, while holding the vibrotactile device between their fingers. The target audio stimuli were played through the headphones, while the babble track in the background simulated a noisy environment. After audio and vibrotactile stimuli of different temporal offsets were presented,visual prompts would appear on the computer screen to elicit a correct selection in a forced-choice task Participants continued the process to the end of the experiment, response data was collected.

3 Results

Figure 1 compares the mean percentage of participant accuracy scores of temporal offsets overall, and in respect to phonemic contrasts (fricative vs stops). A 2-way ANOVA shows that there is no amount of the variation in accuracy scores can be attributable to the factor of temporal offset ($F_{(1,25)}$ = 0.0061, p = 0.94),however, there is significant variation for the factor of phonetic contrast ($F_{(1,25)}$ = 66.0102, p = 1.7704*10^{-8}). Since no variance can be accounted for in terms of temporal offset, skewedness was not tested. The distribution of the overall curve is not skewed in any discernible direction, and is symmetrical around an unknown mean [6].

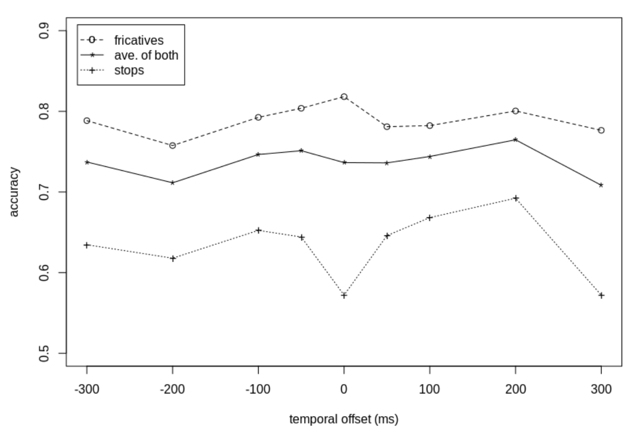

Figure 1: Mean average of participant accuracy scores

4 Discussion

Temporal offset did not effect accuracy scores as anticipated. However, there was a significant effect of phonemic contrasts on participant accuracy. Qualitative data from participants collected during an unstructured post-study interview reinforced the data that stops (pa/ba) were more difficult to perceive in general compared to the fricatives. There was no clear temporal window of integration between the vibrotactile feedback and audio signals. Possible reasons could be that instead of supplementing additional linguistic information–the vibrotactile feedback aided in directing the participant's attention to the audio signal. If that were the case, latency issues in the practical real-time application of such vibrotactile devices for speech intelligibility may not be of a big concern, as the perceiver could depend on the vibrations to guide their attention to what is being said.

Acknowledgments

This research project was funded by UBC Language Sciences.

References

[1] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746, 1976.

[2] Kevin G Munhall, P Gribble, L Sacco, and M Ward. Temporal constraints on the mcgurk effect. *Perception & psychophysics*, 58(3):351–362, 1996.

[3] Bryan Gick and Donald Derrick. Aero-tactile integration in speech perception. *Nature*, 462(7272):502, 2009.

[4] Bryan Gick, Yoko Ikegami, and Donald Derrick. The temporal window of audio-tactile integration in speech perception. *The Journal of the Acoustical Society of America*, 128(5):EL342–EL346, 2010.

[5] David Marino, Hannah Elbaggari, Tzu Hsu Chu, Bryan Gick, and Karon MacLean. Single-channel vibrotactile feedback for voicing enhancement in trained and untrained perceivers. *in press*, 2018.

[6] Weiwen Miao, Yulia R Gel, and Joseph L Gastwirth. A new test of symmetry about an unknown median. In *Random Walk, Sequential Analysis And Related Topics: A Festschrift in Honor of Yuan-Shih Chow*, pages 199–214. World Scientific, 2006.

[7] Katie Bicevskis, Donald Derrick, and Bryan Gick. Visual-tactile integration in speech perception: Evidence for modality neutral speech primitives. *The Journal of the Acoustical Society of America*, 140(5):3531–3539, 2016.