Recognising Speech Patterns through Hearing, Seeing and Touching

Brandon Maria*

Department of Otolaryngology, University of Washington, USA

Editorial

When there is only a visual speech signal available, people can lip-read themselves more accurately than they can lip-read others. The common-coding hypothesis, which contends that watching an action activates the same motor plan representation as actually performing it, and that watching one's own actions activates motor plan representations more than watching those of others because of greater congruity between precepts and corresponding motor plans, is consistent with the self-advantage for vision-only speech recognition. By evaluating whether there is a self-advantage when the visual signal is added to the auditory signal in difficult listening situations, the current study expands this line of research to audio-visual speech recognition [1]. For round-robin testing, participants were divided into subgroups. Each person was paired with every other participant in their subgroup, including themselves, acting as both a talker and a listener/observer. On average, participants benefited more from the visual cue when they were speaking than when someone else was, and they also benefited more from the visual cue than others did from watching and listening to them.

Furthermore, after statistically correcting for individual differences in the two participants' capacities to gain from a visual speech signal and the degree to which their own visual speech signal helped others, the self-advantage in audio-visual speech recognition remained substantial. These results support the concept of a shared code for action perception and motor plan representation as well as our earlier discovery of a self-advantage in lip reading. Individuals watching a video clip may identify prior movements as being self- rather than other-generated, even when little to no identifying information is supplied other than the peculiarities of the movements themselves, which is consistent with this idea. It's significant to note that common coding's support does not just come from its use in agency identification [2]. Participants, for instance, may more accurately predict an action's outcome from movies of themselves than from videos of others, which raise the possibility that people can directly understand certain acts by mapping perceived actions onto their own action repertoire.

According to some researchers, the perception of action, and speech perception in particular, involves the activation of the motor representations in the brain that are essential for producing the action. Auditory speech stimuli also produce this activation, suggesting that similar processes may be involved in both visual and auditory speech recognition. There should be more agreement between visually perceived speech gestures and the corresponding motor plans and associated kinaesthetic experiences when the talker and the observer are the same person because individuals have distinctive motor signatures for speech gestures, just as they do for other actions [3].

This should result in individuals' speech motor plans being more activated when they observe themselves speaking compared to when they hear others talk, and this increased activation may be the reason why people are better at lip-reading themselves. We hypothesise that people's capacity to utilise visual speech information in audio-visual environments when listening is challenging should be impacted by the stronger activation of suitable speech motor plans. The present study's goal was to test our common coding account by determining whether participants would experience a greater benefit from adding the visual signal to the auditory signal when they were shown recordings of their own visual and auditory speech signals than when the talker was not present [4].

The level of each talker in the participant's subgroup was altered while the level of background babbling was maintained constant as part of a staircase technique that was programmed in to calculate the required. In order for the experimenter to assess whether the participant was obtaining about accuracy and, if not, change the levels, the programme proceeded to play back that talker's recordings after three reversals. In order to acquaint participants with the other talkers and lessen the novelty of hearing them talk, practise sessions with each talker in each condition including participants themselves were conducted prior to testing. Participants in the data collection stage were shown stimuli from each talker in their group, including themselves [5].

Conflict of Interest

None

References

1. De Gelder, Beatrice and Jean Vroomen. "The perception of emotions by ear and by eye." *Cogn Emot* 14 (2000): 289-311.

2. Klatzky, Roberta L., Susan J. Lederman and Victoria A. Metzger. "Identifying objects by touch: An "expert system"." *Perc Psyc* 37 (1985): 299-302.

3. Haxby, James V., Elizabeth A. Hoffman and M. Ida Gobbini. "The distributed human neural system for face perception." *Tre Cogn Sci* 4 (2000): 223-233.

4. Pantic, Maja and Leon J.M. Rothkrantz. "Toward an affect-sensitive multimodal human-computer interaction." *Proce IEEE* 91 (2003): 1370-1390.

5. Potamianos, Gerasimos, Chalapathy Neti, Guillaume Gravier and Ashutosh Garg. "Recent advances in the automatic recognition of audiovisual speech." *Proc IEEE* 91 (2003): 1306-1326.

***Address for Correspondence**: Brandon Maria, Department of Otolaryngology, University of Washington, USA; E-mail: brandonmaria@gmail.com*