Speech-driven facial animation improve speech-in-noise comprehension in humans

Enrico Varano*[1] and Tobias Reichenbach†[2]

[1]Imperial College London – United Kingdom
[2]Friedrich-Alexander Universität Erlangen-Nürnberg – Germany

Abstract

Comprehension of speech in noise can be improved by looking at the speaker's face. This effect is even more pronounced in people with hearing impairments and is thought to be linked to both the temporal and categorical cues carried by the visual component of speech. For instance, rhythms such as the amplitude modulations of a speech signal, which are known to play an important role in speech processing, correlate with the opening and closing of the mouth. Categorical cues are involved in audiovisual speech perception through the non-injective surjective mapping between phonemes and visemes and information about a speech signal can be obtained from other aspects of lip movement as well. However, the precise contributions of the different aspects of lip motion to speech comprehension, as well as the neural mechanisms behind the audio-visual integration, still remain unclear.
We considered both the natural video of a speaker as well as a variety of synthesized visual signals. The synthesized videos were designed to capture features of lip movements of increasing complexity, from the amplitude modulations of speech to realistic facial animations generated by deep neural networks. We then assessed the speech comprehension of participants for the different types of videos. We also recorded their brain activity through EEG while they listened to the audiovisual speech stimuli.

We found that simple visual features such as the size of the mouth opening, related to the speech envelope, modulated the neural response to the speech envelope. However, they failed to enhance speech comprehension. More complex videos including the realistic synthesised facial animations did improve the comprehension of speech in noise significantly, albeit not as much as the natural videos.
Taken together, our results suggest that categorical cues in the texture of realistic facial animations drive the audiovisual gain in speech-in-noise comprehension. Although the amplitude modulation of speech matters for speech processing, and although simplified visual signals that track these amplitude modulations influence the neural response, these signals do not aid in understanding speech in background noise.

*Speaker
†Corresponding author: tobias.j.reichenbach@fau.de