

Miki Inoue
Max-Planck-Institute for Human Cognitive and Brain Sciences,
Sensorimotor Coordination Junior Research Group, Munich

inoue@phonetik.uni-muenchen.de

inoue@cbs.mpg.de

1. Modelling of physical properties in speech production:

In former studies (Inoue 2003), an acoustic model for lip production was developed. In this simple 2D model lip shapes of static vowels are predicted by using acoustic features (F1, F2, F3) as inputs for 4 direction parameters. A perception experiment of the output shapes and its results will be presented. As we know, kinematic (and dynamic) data show us that there are large inter- and interspeaker variability which is also dependent from the context, language, and from prosodic properties like speech rate, loudness, or emphasis. Nevertheless, in non-pathological case, for each speaker the same set of articulators and muscles should be available. Also, the really basic mechanism and higher-level representations of action (and perception) could be the same in each speaker. Several theories, like in the Theory of Common Coding (Hommel et al. 2001), the Motor Theory of Speech Perception (Liberman 1989), or in the Realist Direct Theory of Speech Perception (Fowler 1993) provide the assumption that there is a common amodal (motor) space where auditory and visual inputs are projected during the perception process. For motor production there is an access to that common amodal representation. In my doctoral thesis I attempt to find some evidences for common action and perception representation using Optotrak recordings of lip movements during speech.

2. Audiovisual perception and speech production:

In the recent literature, there are several models on audiovisual integration which tries to explain the relationship between physical signals and human speech perception. As known from Sumbly and Pollack (1954) speech intelligibility is improved when synchronous visual information is added to noisy signal. In elder models like the Direct Identification Model (Klatt 1979) it is assumed that the audiovisual input signal is transmitted directly to a phonemic classifier where the most similar(?) prototypic item is selected from a bimodal lexicon. The parameters for acoustic (spectral) and visual (face) similarities to the prototype are not clearly defined. According to the Separated Identification Model there are parallel identification processes for each modality which yield two phonemic classifications (visual properties for articulation place and auditory for articulation mode) and output the most weighted category. This model fails in explaining the McGurk-effect where concurrent inputs lead to a new phoneme category. Since speech is intelligible in common communication with closed eyes but not with ear-plugs the auditory channel is seen as the dominant modality. It is supposed in the Dominant Modality Recoding Model that the visual information is recoded in auditory dimension as a cepstrum. Then both VT transfer-functions are integrated after a separate evaluation and weighting. If the visual information allow more than two different VT functions (e.g. visible lip closure: /m/, /p/ or /b/?) the ambiguous function to the auditory input will be selected. It could be shown in one of my visual only experiments where productions of VCV-sequences spoken by one male speaker were presented as movies, that the subjects tend to identify voiced bilabial consonants (/m, b/) by chance which confirms the DMR model. But what happens when in presented audiovisual movies a short medial sequence is visual only? There would be no any dominant auditory information which could be improved by vision. Do subjects then switch in a lipreading mode for a short period? If yes, does the change cost time? Which kind of artefacts could arise? If we interpolate the missing acoustics using some transition information and coarticulatory effects, which information are crucial? Or do humans have other systematical strategies to perceive a visual sound?