

## Audio-visual synthesis of a talking head and McGurk effect

Shinji Maeda

In this tutorial, we first describe a modeling of face deformations during speech. The deformations are measured using a motion capture system that provides temporal variations of 3D coordinates of about 60 markers glued on the face. A factor analysis on the measured markers' coordinates has indicated that the first six factors can describe marker positions with a reasonable accuracy, because they explain about 90% of the variance. Data derived values of the six factors allow us to reconstruct observed face deformations. At present, we simply connect markers by straight lines to obtain a gross representation of the human face. Also, it can be used as a face synthesizer for a talking head, because we can arbitrarily manipulate factor values to synthesize some intended face deformations along speech sounds. We have therefore a more flexible face synthesizer than that based on concatenation of visemes (the visual equivalent of phonemes) with interpolations.

Second, we describe our recent progress in the evaluation of the face synthesizer exploiting McGurk effect as a paradigm. In an informal test, reconstructed face deformations during [da] and [ga] combined with the sound of [ba] elicit the perception of /*da*/, i.e. the McGurk effect, at least for some listeners. In another test, face deformations are synthesized by temporal interpolations in 6 steps from the measured deformations of [ba] to those of [ga], creating a visual continuum. Each of interpolated face deformations is again combined with the same [ba] sound. Responses of those listeners indicate an abrupt change in the perception from /*ba*/ to /*da*/ (but never to /*ga*/). Although not conclusive, these results suggest that our rudimentary face model can convey visual phonetic cues to listeners. Moreover, such visual cues seem to behave in a way similar to the auditory place cue.