ACOUSTICS2008/1354
Multimodal control of talking heads

Gerard Bailly[a], Oxana Govokhina[a] and Gaspard Breton[b]
[a]GIPSA-lab. Dept Speech & Cognition, INPG, 46, av. Félix Viallet, 38031 Grenoble, France
[b]Orange R&D, 4 rue du Clos Courtel, 35512 Cesson-Sévigné, France

Multimodal speech synthesis has been devoted for years to the rendering of linguistic or paralinguistic content - i.e. parameterized but discrete information - by continuous audible and visible consequences of speech articulation, eventually complemented by facial expressions, gaze and other body gestures including head, hand and arm movements. Articulatory synthesizers (producing sounds from gestures) intrinsically compute coherent audiovisual signals but do not presently compete with data-driven techniques: most talking heads are nowadays controlled by models built using human audiovisual data. These control models should replicate the laws governing the coherence of observed multimodal signals and the correct phasing relations between salient events of the multimodal stream. We will report on two comparative evaluations of various lip-sync models (dealing with post-synchronization between speech sounds and articulatory movements) and present a trainable control model that learns automatically phasing relations between acoustic and gestural events. This model can be further extended to capture the fine temporal structure of multimodal scores and a first application to the synchronization between speech and head, face and hand movements during cued speech production will be presented.