ARTICULATORY-TO-ACOUSTIC MAPPING: APPLICATION TO SILENT SPEECH INTERFACE AND VISUAL ARTICULATORY FEEDBACK

Thomas Hueber, Atef Ben-Youssef, Pierre Badin, Gérard Bailly, Frédéric Elisei
GIPSA-lab UMR 5216/CNRS/INP/UJF/U.Stendhal, Grenoble, France

We present here two research projects developed in the Speech and Cognition department of GIPSA-lab: (1) a *silent speech interface* which converts tongue and lip motions, captured by ultrasound and video imaging, into audible speech, and (2) a *visual articulatory feedback system*, which automatically animates, from the speech sound, a 3D orofacial clone. Both systems are based on the modeling of "parallel" articulatory-acoustic data (speech sounds recorded simultaneously with articulatory movements) using supervised machine learning techniques (a branch of artificial intelligence).

A "silent speech interface" (SSI) is a device that allows speech communication without the necessity of vocalizing. SSI could be used in situations where silence is required (as a silent cell phone), or for communication in very noisy environments. SSI could also be used by laryngectomized patients as an alternative to electrolarynx; to oesophageal speech; or to tracheo-oesoephageal speech. The design of a SSI has recently received considerable attention from the speech research community. In our approach [1], articulatory movements are captured during silent articulation by a non-invasive multimodal imaging system composed of an ultrasound transducer placed beneath the chin and a video camera in front of the lips (see figure 1).

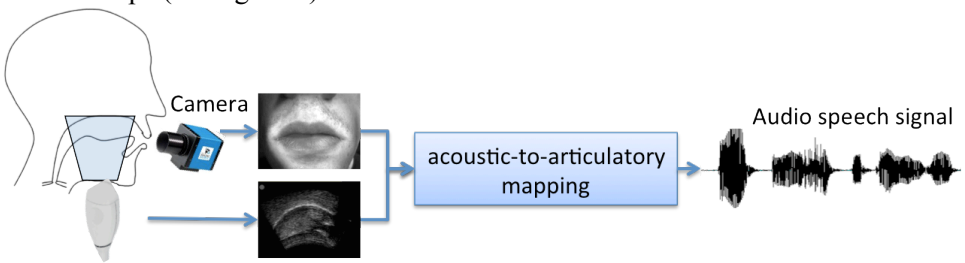

Figure 1: Ultrasound-based silent speech interface

A "visual articulatory feedback system" aims at providing the speaker with visual information about his/her own articulation. Several studies show that this kind of system can be useful for both speech therapy and Computer Aided Pronunciation Training (CAPT). The use of different types of sensors, such as electro-palatography (EPG) and ultrasound imaging, has been proposed in the literature. Our system [2] is based on a 3D talking head used in an "augmented speech scenario", *i.e.* displaying all speech articulators including the tongue (see figure 2). The talking head consists of three-dimensional models of various speech organs of the same speaker, built from MRI, video, and EMA (electromagnetic articulography) data. The talking head is animated automatically from the audio speech signal, using acoustic-to-articulatory inversion. The inversion method is based on the modeling of the relationship between the speech sound and the corresponding articulatory trajectories, using statistical mapping techniques.

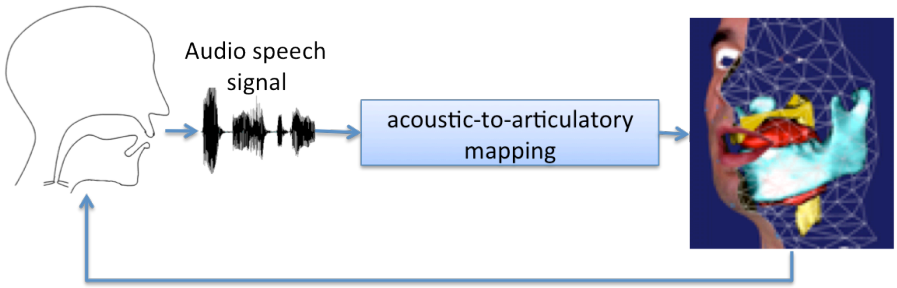

Figure 2: Visual articulatory feedback system

[1] Hueber, T., Benaroya, E. L., Chollet, G., Stone, M., "Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips," Speech Communication, vol. 52, no. 4, pp. 288-300, 2010.

[2] Badin, P., Ben Youssef, A., Bailly, G., Elisei, F., and Hueber, T., "Visual articulatory feedback for phonetic correction in second language learning," L2SW, Tokyo, Japan, 2010.