ACOUSTICS2008/1146
An ultrasound-based silent speech interface

Thomas Hueber[a], Gerard Chollet[b], Bruce Denby[c], Gerard Dreyfus[c] and Maureen Stone[d]
[a]ESPCI - Telecom Paris, 10 rue Vauquelin, 75005 Paris, France
[b]Telecom Paris Tech, 46 rue Barrault, 75013 Paris, France
[c]Université Paris VI, ESPCI - Laboratoire d'Electronique, 10 rue Vauquelin, 75005 Paris, France
[d]Vocal Tract Visualization Lab, Depts of Biomedical Sciences and Orthodontics, University of Maryland Dental School, 650 W. Baltimore St., Baltimore, MD 21201, USA

The paper proposes the use of ultrasound scans of tongue movement and video sequences of the lips to synthesize speech. A speech synthesizer driven only by video acquisitions may be qualified as a "silent speech interface," which could be used by laryngectomyzed patient as an alternative to tracheo-esophageal speech, for voice communication where silence must be maintained, or in very noisy environments. Our system is based on the building of a one-hour audiovisual corpus of phonetic units, which associates visual features extracted from video with acoustic observations. The ultrasound and optical images are interpreted as a linear combination of standard configurations obtained by Principal Components Analysis (PCA) from a phonetically balanced subset of typical frames. HMM-based stochastic models trained on these visual features sequences are subsequently used to predict phonetic targets from video-only data. Finally, a Viterbi unit selection algorithm is used to find the optimal sequence of acoustic units given both this phonetic prediction and the sequence of visual features. The system is able to perform phonetic transcription from video-only speech data with over 55% correct recognition, on continuous speech, using neither phonotactic nor linguistic constraints.