Automatic Animation of an Articulatory Tongue Model from Ultrasound Images of the Vocal Tract

Diandra Fabre, Thomas Hueber, Laurent Girin,
Xavier Alameda-Pineda, Pierre Badin

PII: S0167-6393(16)30369-7
DOI: 10.1016/j.specom.2017.08.002
Reference: SPECOM 2479

To appear in: *Speech Communication*

Received date: 15 December 2016
Revised date: 26 July 2017
Accepted date: 1 August 2017

Please cite this article as: Diandra Fabre, Thomas Hueber, Laurent Girin, Xavier Alameda-Pineda, Pierre Badin, Automatic Animation of an Articulatory Tongue Model from Ultrasound Images of the Vocal Tract, *Speech Communication* (2017), doi: 10.1016/j.specom.2017.08.002

Automatic Animation of an Articulatory Tongue Model from Ultrasound Images of the Vocal Tract

Diandra Fabre[1], Thomas Hueber[1], Laurent Girin[1,2], Xavier Alameda-Pineda[2,3], Pierre Badin[1]

1 Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
2 INRIA, Perception Team, 38334, Montbonnot, France
3 Univ. of Trento, DISI, 38123, Trento, Italy

Abstract

Visual biofeedback is the process of gaining awareness of physiological functions through the display of visual information. As speech is concerned, visual biofeedback usually consists in showing a speaker his/her own articulatory movements, which has proven useful in applications such as speech therapy or second language learning. This article presents a novel method for automatically animating an articulatory tongue model from ultrasound images. Integrating this model into a virtual talking head enables to overcome the limitations of displaying raw ultrasound images, and provides a more complete and user-friendly feedback by showing not only the tongue, but also the palate, teeth, pharynx, etc. Altogether, these cues are expected to lead to an easier understanding of the tongue movements. Our approach is based on a probabilistic model which converts raw ultrasound images of the vocal tract into control parameters of the articulatory tongue model. We investigated several mapping techniques such as the Gaussian Mixture Regression (GMR), and in particular the Cascaded Gaussian Mixture Regression (C-GMR) techniques, recently proposed in the context of acoustic-articulatory inversion. Both techniques are evaluated on a multispeaker database. The C-GMR consists in the adaptation of a GMR reference model, trained with a large dataset of multimodal articulatory data from a reference speaker, to a new source speaker using a small set of adaptation data recorded during a preliminary enrollment session (system calibration). By using prior information from the reference model, the C-GMR approach is able (i) to maintain good mapping performance while minimizing the amount of adaptation data (and thus limiting the duration of the enrollment session), and (ii) to generalize to articulatory configurations not seen during enrollment better than the GMR approach. As a result, the C-GMR appears to be a good mapping technique for a practical system of visual biofeedback.