Development and Assessment of Visual Articulatory Biofeedback Systems for Speech Therapy

Diandra Fabre1,2, Thomas Hueber1,2, Mélanie Canault1, Nathalie Bedoin1, Pierre Badin1,2

1 Univ. Grenoble Alpes, GIPSA-Lab, F38000 Grenoble, France
2 CNRS, GIPSA-Lab, F38000 Grenoble, France
3 Laboratoire Dynamique du Langage, UMR 5596, CNRS, Université Lumière Lyon 2, France
diandra.fabre@gipsa-lab.grenoble-inp.fr

Abstract
Rehabilitating an articulation or swallowing disorder may require correcting the placement of the tongue. In this context, visual biofeedback system aims at providing a patient with a visual representation of his/her own tongue movements. This biofeedback is intended to help the patient to better understand the origin of the trouble, and should therefore improve rehabilitation. This PhD thesis aims at improving visual biofeedback systems based on ultrasound imaging. Two approaches are investigated: 1) augmenting the raw ultrasound image by enlightening the tongue contour as well as the articulatory targets, and 2) replacing the raw ultrasound image by a so-called articulatory talking head revealing all the internal structure of the vocal tract (such as the palate, the pharyngeal wall, etc.).

Index Terms: speech technology, machine learning, ultrasound imaging, biofeedback, tongue, speech therapy

1. Introduction
In the context of speech therapy, some troubles are related to a wrong placement of the tongue, like the articulatory and phonological troubles, and swallowing disorders. To help the rehabilitation process, speech therapists can use drawings, puppets or a mirror to help their patients understand the correct articulation to perform. We make the hypothesis that a visual biofeedback on the patient’s tongue movements could help both the speech therapist and the patient to understand the trouble and improve the articulation (we assume here that this visual information will be efficiently combined with auditory and kinesthetic biofeedback).

Several systems of so-called visual articulatory biofeedback have been proposed in the literature. One of the main techniques is the Electropalatography (EPG) that consists on an artificial palate placed against the patient’s palate, as described in [1]. Contacts between the tongue and the palate can then be visualized on a screen. Unfortunately, this technique is relatively invasive, and required to build an artificial palate for each of the patient. Ultrasound imaging is another way to capture the movements of the tongue (as shown in Figure 1). It is minimum invasive and harmless. This is the technique considered in this PhD thesis.

In [2], Cleland et al. used ultrasound biofeedback for the rehabilitation of seven children with persistent speech sound disorders. In [3], Preston et al. performed a case study for a case of acquired apraxia, and validated the use of ultrasound to treat this pathology.

However, even if these studies tend to show the benefit of ultrasound biofeedback for speech rehabilitation, this imaging technique still presents various drawbacks. No information is provided about the other structures of the vocal cavity (palate, teeth...). The apex can be hidden by the shadow of the jawbone. Moreover, when the tongue is parallel to the ultrasound waves, parts of the contour can be missing.

The first goal of this PhD is to improve visual biofeedback systems based on ultrasound imaging. Two approaches are investigated:

1) Augmenting the raw ultrasound image by enlightening the tongue contour (automatic tongue contour segmentation) as well as the articulatory targets,

2) Replacing the raw ultrasound image by a so-called articulatory talking head, including the other structures of the vocal cavity (palate, teeth...).

The second goal of this PhD is to evaluate these systems in a practical, and ideally large-scale rehabilitation scenario. Several troubles involving bad tongue placement are envisioned. So far, we focused mainly on the first goal. Two systems have been proposed and described respectively in two Interspeech papers [4], [5]. Our main results are briefly summarized in the next section.

2. Augmenting the raw ultrasound image

2.1. State-of-the-art

One way to enhance an ultrasound image of the vocal tract is to enlighten the contour corresponding to the upper surface of the tongue. This is a difficult problem since the contour can be very poorly imaged for the following reasons: the ultrasound beam is not correctly back reflected to the probe in several tongue position (as in [i]), the tongue tip is hidden by the...
acoustic shadow of the mandible, the probe is not strictly in the midsagittal plane, etc.

Several approaches for extracting the tongue contour in ultrasound images have been proposed in the literature: active contours (snakes) [6], an Active Appearance Model (AAM), [7], Deep Belief Network (tDBN) [8]. So far, these methods have been developed in the context of phonetic researches but were not validated in the context of visual biofeedback and speech therapy. In fact, such context requires minimum manual intervention as well as training data. To that purpose, we proposed another technique, briefly described in the next section.

2.2. Proposed approach

The tongue segmentation method proposed here is in line with Fasel & Berry’s work. This method models the statistical relationships between pixels’ intensity over the entire ultrasound image and the tongue contour, using an artificial neural network (ANN). The proposed technique is based on: (1) a PCA-based decomposition technique named EigenTongue used to parameterize the pixels intensity and described in [9], (2) a PCA-based model of the tongue contour referred here to as EigenContour, and (3) an artificial neural network for modeling the relationships between these two representations. A block diagram of the method is given in Figure 3. The method was evaluated on 9 speakers. Its performance was evaluated using both a speaker-dependent and speaker-independent approaches.

3. Tongue model animation from ultrasound images

3.1. State-of-the-art

Another approach to display a target articulatory gesture is to use an articulatory talking head (ATH), i.e. a virtual head able to display the internal articulators (tongue, velum) using augmented reality (e.g. a transparent skin). Contrary to ultrasound images, an ATH makes the display very intuitive, since it displays all the internal structures of the vocal tract. Several approaches to use an ATH as a biofeedback tool have been proposed in the context of second language learning. In [10] for instance, pre-recorded animations of an ATH were used to teach Swedish phonemes to French learners.

In previous work [11, 12], Hueber et al. proposed a system in which the visual biofeedback was calculated automatically from the user’s voice. This approach gave encouraging results on non-pathological speakers. However, such an approach would be difficult for pathological speakers, since it relies on the speech acoustics only.

3.2. Proposed approach

Based on these considerations, we developed a system which combines 1) ultrasound imaging in order to capture directly the user’s tongue movements rather than estimating it from the acoustics, and 2) an ATH aimed to provide the most intuitive display. A Gaussian Mixture Regression (GMR) is used to convert the visual features extracted from the recorded ultrasound data into a sequence of control parameters of the ATH. A general overview of the system is presented in Figure 2.

The proposed technique was evaluated on a French female with no articulation disorders. The performance was assessed using an HMM-based articulatory recognizer. With 75% of the articulations correctly estimated, this pilot study validates the proposed method. Future work will be performed with multiple non-pathological and pathological speakers. Moreover, information on the lips and the jaw will be added to animate the entire ATH.

4. Conclusions and future work

The different methods presented in this paper aim at improving articulatory rehabilitation during speech therapy when using ultrasound biofeedback. These improvements can be performed in different ways: 1) by augmenting the raw ultrasound image with the tongue contour as well as the articulatory targets, and 2) by replacing the raw ultrasound image by an articulatory talking head revealing all the internal structure of the vocal tract.
However, these methods still have to be developed in real-time, in order to be evaluated in a clinical context. In parallel to these technological developments, we also establish links between the research and speech therapy to conduct several studies and validate, in France, the feasibility of the technique and the efficiency of ultrasound biofeedback for speech therapy.

5. Acknowledgements

The Région Rhône-Alpes ARC6 funding agency is thanked for supporting this work through doctoral funding.

6. References


