Speech Communication 99 (2018) 27–46

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Automatic segmentation of speech articulators from real-time midsagittal MRI based on supervised learning☆

Mathieu Labrunie[a], Pierre Badin[a,*], Dirk Voit[b], Arun A Joseph[b], Jens Frahm[b], Laurent Lamalle[c], Coriandre Vilain[a], Louis-Jean Boë[a]

[a] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
[b] Biomedizinische NMR Forschungs GmbH am Max-Planck-Institut für biophysikalische Chemie, Göttingen, Germany
[c] Univ. Grenoble Alpes, INSERM, CHU Grenoble Alpes, CNRS, UMS IRMaGe - Inserm US 17 - CNRS UMS 3552, 38043, Grenoble, France

ARTICLE INFO

Keywords:
Real-time MRI
Speech articulation
Articulator segmentation
Multiple Linear Regression
Active Shape Models
Shape Particle Filtering

ABSTRACT

Speech production mechanisms can be characterized at a peripheral level by both their acoustic and articulatory traces along time. Researchers have thus developed very large efforts to measure articulation. Thanks to the spectacular progress accomplished in the last decade, real-time Magnetic Resonance Imaging (RT-MRI) offers nowadays the advantages of frame rates closer than before to those achieved by electromagnetic articulography or ultrasound echography while providing very detailed geometrical information about the whole vocal tract. RT-MRI has thus become inescapable for the study of speech articulators' movements. However, making efficient use of large sets of images to characterize and model speech tasks implies the development of automatic methods to segment the articulators from these images with sufficient accuracy. The present article describes our approach to develop, based on supervised machine learning techniques, an automatic segmentation method that offers various useful features such as (1) capability of dealing with individual articulators independently, (2) ensuring hard palate, jaw and hyoid bone to be adequately tracked as rigid structures, (3) delivering contours for a full set of articulators, including the epiglottis and the back of the larynx, which partly reflects the vocal fold abduction / adduction state, (4) dealing more explicitly and thus more accurately with contact between articulators, and (5) reaching an accuracy better than one millimeter.

The main contributions of this work are the following. We have recorded the first large database of high quality RT-MRI midsagittal images for a French speaker. We have manually segmented the main speech articulators (jaw, lips, tongue, velum, hyoid, larynx, etc.) for a small training set of about 60 images selected by hierarchical clustering to represent the whole corpus as faithfully as possible. We have used these data to train various image and contour models for developing automatic articulatory segmentation methods. The first method, based on Multiple Linear Regression, allows to predict the contour coordinates from the image pixel intensities with a Mean Sum of Distances (MSD) segmentation error over all articulators of 0.91 mm, computed with a Leave-One-Out Cross Validation procedure on the training set. Another method, based on Shape Particle Filtering, reaches an MSD error of 0.66 mm. Finally the modified version of Active Shape Models (mASM) explored in this study gives an MSD error of a mere 0.55 mm (0.68 mm for the tongue). These results demonstrate that this mASM approach performs better than state-of-the-art methods, though at the cost of the manual segmentation of the training set. The same method used on other MRI data leads to similar errors, which testifies to its robustness. The large quantity of contour data that can be obtained with this automatic segmentation method opens the way to various fruitful perspectives in speech: establishing more elaborate articulatory models, analyzing more finely coarticulation and articulatory variability or invariance, implementing machine learning methods for articulatory speaker normalization or adaptation, or illustrating adequate or prototypical articulatory gestures for application in the domains of speech therapy and of second language pronunciation training.

Abbreviations: AAM, Active Appearance Models; ASM, Active Shape Models; EMA, ElectroMagnetic Articulography; fps, frames per second (also images per second); HOGs, Histogram of Oriented Gradients; mASM, modified Active Shape Models; MLR, Multiple Linear Regression; MRI, Magnetic Resonance Imaging; MSD, Mean Sum of Distances; PCA, Principal Components Analysis; RMS, Root Mean Square; ROI, Region of Interest; RT-MRI, real time MRI; SPF, Shape Particle Filtering; SURF, Speed-Up Robust Features

☆ This article presents an extension of work published in Labrunie et al. (2016b)

* Corresponding author.
E-mail addresses: mathieu.labrunie@gipsa-lab.grenoble-inp.fr (M. Labrunie), pierre.badin@gipsa-lab.grenoble-inp.fr (P. Badin).

https://doi.org/10.1016/j.specom.2018.02.004
Received 24 May 2017; Received in revised form 23 January 2018; Accepted 26 February 2018
Available online 01 March 2018
0167-6393/

1. Introduction

Speech production mechanisms can be characterized at a peripheral level by both their acoustic and vocal tract articulatory movements along time. Researchers have thus developed very large efforts to obtain acoustic and articulatory data in order to analyze the sounds of the world's languages (Ladefoged and Maddieson (1996)), to develop articulatory models of speech production (*e.g.* Lindblom and Sundberg (1971), Mermelstein (1973), Maeda (1990), Badin et al. (2002), or Kröger and Birkholz (2007)), to infer articulatory speech motor control schemes (*e.g.* Perkell et al. (2000), Perrier (2012)), or still to model relations between gestures and sounds by statistical methods (*e.g.* Toda et al. (2008), Hueber et al. (2015)). Some works deal with overall vocal tract geometry in relation with vocal tract acoustics (*e.g.* Fant (1960), Story and Titze (1998)), while others are focused on individual articulators for developing biomechanical articulatory models (*e.g.* Stavness et al. (2013), Perrier et al. (2016)) or for studying speech temporal coordination (*e.g.* Fowler and Saltzman (1993), Wang et al. (2013)).

The techniques implemented to obtain articulatory data for speech have evolved from cineradiography (Moll (1960), Simon (1967), Bothorel et al. (1986), Badin (1991), Stark et al. (1998)), to X-ray microbeam (Kiritani (1986)), ElectroMagnetic Articulography (EMA) (Perkell, Cohen et al. (1992), Hoole and Nguyen (1997), Geng et al. (2013), Savariaux et al. (2017)), or ultrasound echography (Stone et al. (1983); Morrish et al. (1985); Hueber and Denby (2009)). As mentioned in Masaki et al. (2008)'s brief history of MRI in speech production studies, MRI has been used to study speech articulation since the 1980 s (Rokkaku et al. (1986)), but was limited to sustained sounds until the beginning of the 1990 s (for one of the first dynamic MRI studies of speech, see *e.g.* Foldvik et al. (1990)). The fast techniques are limited to the characterization of the position of a few fiducial markers (EMA) or provide images with poor spatial resolution and Signal to Noise Ratio (SNR) and only partial coverage of the Field of View (FoV) of interest (ultrasound echography). Non-invasiveness makes MRI a method of choice to perform systematic studies of speech production on healthy volunteers. The technique however suffers from limited SNR, imposing compromises between spatial resolution, volume coverage and temporal resolution. However, continuous progress on hardware, acquisition approaches, and reconstruction algorithms has markedly improved image quality and spatiotemporal resolution. Relevant hardware changes include higher magnetic fields (3.0T *vs* 1.5T or lower), improved receiving coils as well as gradients with higher strengths, slew rates, and duty cycles. Faster acquisition schemes are associated with spiral and radial k-space trajectories in combination with parallel encoding, data undersampling, and appropriate reconstruction algorithms (for overviews on these subjects, *cf.* Uecker et al. (2012), Lingala et al. (2016a)) or Lingala et al. (2016b))). This spectacular progress has thus boosted RT-MRI acquisition rate up to 20 – 100 images / second (henceforth *fps*) for single slice images, while decreasing pixel size down to 1.5 mm with reasonable image quality (see Table 1). Nowadays, this medical imaging technique offers frame rates closer than before to those provided by EMA or ultrasound echography while providing very detailed information about the whole vocal tract. RT-MRI has thus become inescapable for the study of speech articulators' movements in various types of tasks, such as speaking (*e.g.* Silva and Teixeira (2015)), swallowing (Olthoff et al. (2014); Olthoff et al. (2016)), singing (Echternach et al. (2010)), or even blowing musical wind instruments (Iltis et al. (2015)). Nowadays, it is therefore feasible to acquire voluminous corpora of real time midsagittal images, as illustrated by the work of Narayanan et al. (2014) and of Silva and Teixeira (2015).

However, making efficient use of these data to characterize and model speech tasks implies the development of automatic methods to determine the contours of articulators from these images with a quality as precise and reliable as in the traditional manual methods (*e.g.* Soquet et al. (1998), Serrurier and Badin (2008)). Indeed, an expert typically takes around 10 – 15 minutes to trace all articulators' contours on one midsagittal image, which makes it unrealistic to process large quantities of images. The present article, which is a largely extended version of Labrunie et al. (2016b), describes our approach to develop an automatic method able to determine the contours of most speech articulators taken individually from RT-MRI midsagittal images.[1] The ambition of our work was in particular to offer various features such as (1) capability of dealing with individual articulators independently, (2) tracking hard palate, jaw and hyoid bone as rigid structures, (3) delivering contours for a full set of articulators, (4) dealing more explicitly with contact between articulators, and (5) reaching an accuracy better than one millimeter. The specific option of individually treating the articulators aims to avoid implicit or explicit hypotheses about inter-articulator correlations. This is important to maximize the accuracy and reliability of the segmentation, as explained further in Section 3. It is also crucial to characterize temporal inter-articulatory coordination without initial bias on the data. Moreover, this choice does not preclude developing global models taking into account inter-articulatory correlations, whether related to biomechanical constraints or to speech articulatory control strategies (*cf.* Beautemps et al. (2001)).

In this context, we consider two types of structures: the rigid bony structures and the deformable articulators. The skull is a rigid structure that must be tracked to monitor the head movements of the subjects. In addition to the (hard) palate which belongs to the skull structure, other interesting rigid structures are the jaw and the hyoid bone which have specific movements in speech or swallowing. The deformable articulators comprise the lower and upper lips, the tongue, the epiglottis, the velum (soft palate), the pre-epiglottic fat pad (*hyoEpiFat* in the following), the naso-oropharyngeal posterior wall (*pharynx*), and the *back larynx*, as illustrated in Fig. 1.

Note that what we call *back larynx* is not actually an articulator, but corresponds to the variable region of contact of the rear structures of the larynx in the midsagittal plane. During phonation, the action of the transverse and oblique arytenoidian muscles brings tissues surrounding arytenoids in midsagittal contact, which increases the visible contact region, whereas this region may be considerably shrunk during breathing. The horizontal movements of the anterior edge of this region are thus related to vocal fold adduction. The vertical movements of this region are related to vertical movements of the larynx. Tracking this pseudo-articulator can thus bring information about laryngeal position and vocal fold state.

2. Related work

As presented in the very detailed work of Silva and Teixeira (2015), some studies aim to recover individualized organ contours (Bresch and Narayanan (2009); Eryildirim and Berger (2011); Silva and Teixeira (2015)). Others rather aim to determine vocal tract contours so as to derive a midsagittal distance function, *i.e.* the variation of the distance between the upper and lower midsagittal vocal tract contours along a midline running from the glottis to the lips, irrespective of the contribution of individual organs (Bresch et al. (2006); Proctor et al. (2010); Proctor et al. (2011); Lammert et al. (2013); Kim et al. (2014)). Some methods are based on image processing techniques (Proctor et al. (2010); Proctor et al. (2011); Lammert et al. (2013); Kim et al. (2014)), while others make use of more or less complex models of contours or appearance that must be trained from expert labeled data (Silva and Teixeira (2015)), or can be more implicit such as the snakes used by Bresch et al. (2006). In order to allow the comparison of performances of these different methods described in Section 2.2, we first introduce

[1] The terminology found in the literature to refer to the "determination of the contours of an object in an image" is diverse: registration, delineation, segmentation, contour determination or tracking. In the article, *segmentation* will be used.