Medical Image Analysis 44 (2018) 98–114

Contents lists available at ScienceDirect

Medical Image Analysis

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

Multi-hypothesis tracking of the tongue surface in ultrasound video recordings of normal and impaired speech

Catherine Laporte[a,*], Lucie Ménard[b,c]

[a] Department of Electrical Engineering, École de technologie supérieure, Montreal H3C 1K3, Canada
[b] Department of Linguistics, University of Quebec in Montreal, H2X 1L7, Canada
[c] Center for Research on Brain, Language, and Music, Montreal H3G 2A8, Canada

ARTICLE INFO

Article history:
Received 17 January 2017
Revised 28 November 2017
Accepted 2 December 2017
Available online 5 December 2017

Keywords:
Image segmentation
Tracking
Ultrasound
Particle filter
Speech

ABSTRACT

Characterizing tongue shape and motion, as they appear in real-time ultrasound (US) images, is of interest to the study of healthy and impaired speech production. Quantitative anlaysis of tongue shape and motion requires that the tongue surface be extracted in each frame of US speech recordings. While the literature proposes several automated methods for this purpose, these either require large or very well matched training sets, or lack robustness in the presence of rapid tongue motion. This paper presents a new robust method for tongue tracking in US images that combines simple tongue shape and motion models derived from a small training data set with a highly flexible active contour (snake) representation and maintains multiple possible hypotheses as to the correct tongue contour via a particle filtering algorithm. The method was tested on a database of large free speech recordings from healthy and impaired speakers and its accuracy was measured against the manual segmentations obtained for every image in the database. The proposed method achieved mean sum of distances errors of 1.69 ± 1.10 mm, and its accuracy was not highly sensitive to training set composition. Furthermore, the proposed method showed improved accuracy, both in terms of mean sum of distances error and in terms of linguistically meaningful shape indices, compared to the three publicly available tongue tracking software packages Edgetrak, TongueTrack and Autotrace.

1. Introduction

Observing and characterizing tongue motion is of interest to the study of normal and impaired speech production. Ultrasound (US) imaging is widely used for speech research (Stone, 2005) as it provides a spatially dense representation of the moving tongue at relatively good sampling rates, interferes minimally with natural articulatory processes, and is safe, affordable and portable, and thus suitable for field work.

Ultrasound imaging has been used successfully in studying the shapes adopted by the tongue in the sounds of normal adult speech (Slud et al., 2002; Stone and Lundberg, 1996; Stone et al., 1997), the mechanisms of speech development in children (Zharkova et al., 2011) as well as the effect of deafness (Turgeon et al., 2015), blindness (Ménard et al., 2016) and partial glossectomy (Bressmann et al., 2005) on articulation, to name only a few applications. Most such studies focus on static tongue shape at key instants in the recording and do not fully exploit the dynamic nature of speech data. This is somewhat surprising given the known benefits of dynamic speech analysis methods (Lancia and Tiede, 2012) and the relatively high frame rates achievable in US imaging. Part of the explanation for this apparent paradox is that extracting the tongue contour in US video sequences is difficult. While several automated methods exist for this purpose, they remain error prone (Csapó and Lulich, 2015), which means time-consuming manual interventions are almost always required.

This paper presents a new, highly robust automated method for tongue contour tracking in US video sequences based on particle filtering. Preliminary work (Laporte and Ménard, 2015) showed that this approach recovers gracefully and rapidly from occasional tracking failures. This paper describes the method in more detail, also introducing a new contour length consistency constraint that improves the stability of the method. More importantly, it presents a thorough validation study on a much larger data set of long free speech recordings, designed to evaluate the usefulness of the method for studies of impaired speech production The study includes recordings both from healthy speakers and from speak-

* Corresponding author.
E-mail addresses: catherine.laporte@etsmtl.ca (C. Laporte), menard.lucie@uqam.ca (L. Ménard).

https://doi.org/10.1016/j.media.2017.12.003
1361-8415/© 2017 Elsevier B.V. All rights reserved.

ers with Steinert's disease, a form of myotonic dystrophy leading to slow speech, distorted vowels and consonants, and low overall speech intelligibility (Guimaraes et al., 2007; Sjögreen et al., 2011). The automated segmentation results on these test recordings are compared to manually segmented tongue contours in all images (over 20,000), and the comparison is based on a broad set of accuracy measures, including some pertaining to lingusitically relevant tongue shape indices (Ménard et al., 2012), which to the best of the authors' knowledge, has never been done before. The study documents the influence of the method's key parameters (number of particles and training set composition) on its accuracy and performance, and compares the method to three publicly available tongue tracking software packages: Edgetrak (Li et al., 2005), TongueTrack (Tang et al., 2012) and Autotrace (Fasel and Berry, 2010).

The remainder of this paper is structured as follows. Section 2 clarifies this paper's contribution in relationship with the existing literature. Section 3 describes the proposed particle filtering method in detail. Section 4 provides the methodological details of the validation study. The results are presented and analyzed in Section 5. Section 6 concludes and outlines directions for future work.

2. Related work

A natural framework for segmenting deformable structures such as the tongue surface in images is that of active contours or *snakes* (Kass et al., 1988), wherein the locations of a discrete set of vertices defining the deformable contour are estimated such as to minimize an energy functional expressing the expected appearance of the image in the vicinity of the contour (via an *external* energy functional) as well as constraints on the allowed deformations (via an *internal* energy functional). Akgul et al. (1999) used a snake formulation for tongue segmentation and tracking in US video sequences and proposed the use of dynamic programming (Amini et al., 1990) for contour optimization in each frame. Li et al. (2005) modified Akgul et al. (1999)'s external energy functional to include a *band* energy factor that constrains the tongue surface to lie immediately below a bright white band, thereby reducing the likelihood of the snake latching onto speckle noise in the US image. This method is publicly available as the Edgetrak software. Edgetrak requires the user to identify a few points near the tongue contour in the first frame of the sequence and then iteratively proceeds to optimize the snake for that frame and copy it to be used as an initial guess for the next frame, thereby enforcing some degree of temporal consistency.

Edgetrak fails in the presence of rapid tongue curvature increases and occasionally produces tongue shapes that are uncharacteristic of those encountered during speech. One way to overcome this problem is to enforce global shape constraints on the segmented tongue contour by fitting an active shape or active appearance model (Cootes et al., 1995; 2001) of the tongue to the US data, as proposed by Roussos et al. (2009). They built an active appearance model using a training database of segmented X-ray and US images to respectively establish shape constraints and characterize US image texture in the vicinity of the tongue. Active shape models (ASMs) can also be used to constrain and iteratively drive snake optimization (Hamarneh and Gustavsson, 2000; Ghrenassia et al., 2013). In this case, sufficiently accurate ASMs can be acquired from segmented US data only, thus making X-ray acquisitions unnecessary (Ghrenassia et al., 2013).

If a database of segmented US images is available, as is assumed by the shape-constrained snake methods described above, then this database can also be used to learn the relationship between image features and the sought tongue contours using machine learning. Fasel and Berry (2010) investigated this approach using deep neural networks, leading to the publicly available Autotrace software. In Autotrace, one deep neural network is trained on individual segmented US images to build an abstract and compact generative model of the relationship between image features and the location of contour vertices. Another deep neural network establishes the relationship between unlabeled image data and this abstract model so that labels (i.e. tongue segmentations) can then be inferred based on image data only. The method was extended to optimize the selection of training data, thereby improving training efficiency and segmentation accuracy (Berry et al., 2012). Fabre et al. (2015) proposed a similar approach, based on a simpler neural network architecture, that establishes a relationship between the principal component representation of images and that of the sought tongue contour. These methods based entirely on machine learning require no manual initialization. They also perform segmentation independently on each image in a recording, thereby avoiding the propagation of segmentation errors from one frame to the next, but also neglecting the temporal consistency of tongue shape as a useful source of regularizing information. Jaumard-Hakoun et al. (2015) modified the Autotrace method to employ automatically labeled rather than manually labeled training data. The automatic labeling method is coarse and based on simplistic image processing operations, but enforces weak temporal consistency constraints between consecutive frames, which later become embedded in the neural network.

Strong temporal consistency constraints for tongue tracking were introduced by Tang et al. (2012). They modeled the tongue segmentation problem as a global optimization problem over a higher order Markov random field whose vertices represent the tongue contour evolving in 2D+t and whose edges represent adjacency constraints in space (along the tongue contour) and in time. Their method, implemented in their publicly available software TongueTrack, is mathematically elegant and accounts for spatial (i.e. shape) and temporal (i.e. motion) constraints in a flexible manner. It also allows future frames to condition past ones, which may be useful to resolve some ambiguities. This method does not require any training data. As a result, it incorporates very little contextual knowledge. It is also computationally intensive.

Instead of relying on heuristics and/or segmented datasets to constrain tongue shape and kinematics, Loosvelt et al. (2014) used a biomechanical model as prior information for tongue tracking. Rather than tracking a contour in the image, their method involves fitting the biomechanical model (using the finite element method) to a set of tracked point features belonging to the entire tongue (not necessarily its surface). Their preliminary results suggest that realistic tongue motion constraints improve tongue tracking when parts of the tongue are invisible in the image. However, the technique relies on tracking point-based landmarks in US images, a notoriously error-prone task. Further validation would be necessary to reliably assess the performance of the method.

Csapó and Lulich (2015) recently published an analysis of the segmentation errors produced by Edgetrak, Autotrace and TongueTrack in their default configurations using 8 utterances of a short English sentence by 4 speakers. They compared the automatic segmentation results to manual tracings as well as the results from a baseline "algorithm" which copied the manually traced contour from the first frame in the sequence across the entire video. Though the study was limited by the small size and limited diversity of the test data set, it led to interesting findings. One was that for small training data sets, Autotrace's accuracy was highly dependent on the training and test data being similar. Also, for all automated techniques, the evolution of segmentation error over time followed a similar pattern to that produced by the baseline, suggesting that none of the automated approaches entirely adapted to rapid tongue movement. This effect was particularly marked for