Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 33 (2015) 25–46

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Unsupervised segmentation of the vocal tract from real-time MRI sequences☆

Samuel Silva[a,*], António Teixeira[b,a]

[a] *Inst. Electronics Eng. and Telematics of Aveiro (IEETA), University of Aveiro, Aveiro, Portugal*
[b] *Dept. Electronics, Telecommunications and Informatics (DETI), University of Aveiro, Aveiro, Portugal*

Received 21 November 2013; received in revised form 4 November 2014; accepted 15 December 2014
Available online 25 December 2014

Abstract

Advances on real-time magnetic resonance imaging (RT-MRI) make it suitable to study the dynamic aspects of the upper airway. One of the main challenges concerns how to deal with the large amount of data resulting from these studies, particularly to extract relevant features for analysis such as the vocal tract profiles. A method is proposed, based on a modified active appearance model (AAM) approach, for unsupervised segmentation of the vocal tract from midsagittal RT-MRI sequences. The described approach was designed considering the low inter-frame difference. As a result, when compared to a traditional AAM approach, segmentation is performed faster and model convergence is improved, attaining good results using small training sets. The main goal is to extract the vocal tract profiles automatically, over time, providing identification of different regions of interest, to allow the study of the dynamic features of the vocal tract, for example, during speech production. The proposed method has been evaluated against vocal tract delineations manually performed by four observers, yielding good agreement.

Keywords: Vocal tract; Segmentation; Real-time MRI

1. Introduction

During speech production, the vocal tract configuration continuously changes over time. This dynamic nature has long been recognised as of paramount importance to study speech production (Saltzman and Munhall, 1989) and should cover the characterisation of the position and movement of the different articulators involved, such as the tongue, lips and velum (Fig. 1).

Several techniques have been proposed and used to gather data to enable the study of these dynamic aspects such as electromagnetic midsagittal articulography (EMMA) (e.g., Oliveira and Teixeira, 2007) or ultrasound (e.g., Wrench et al., 2011). Even though they provide high frame rates, they present some limitations by focusing in a restricted set of regions of the vocal tract. In recent years, real-time magnetic resonance imaging (RT-MRI) has been used for speech studies (e.g., Narayanan et al., 2011; Teixeira et al., 2012), allowing enough frame rate to provide useful data regarding

☆ This paper has been recommended for acceptance by P. Jackson.

* Corresponding author at: IEETA, Campus Univ. de Santiago, 3810-193 Aveiro, Portugal. Tel.: +351 234370500.
E-mail addresses: sss@ua.pt (S. Silva), ajst@ua.pt (A. Teixeira).

http://dx.doi.org/10.1016/j.csl.2014.12.003
0885-2308/

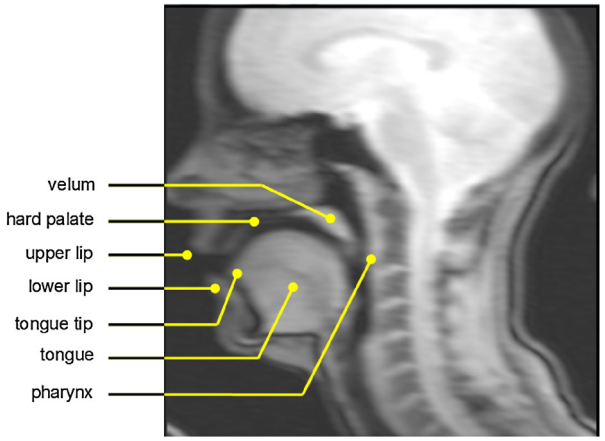

Fig. 1. Example of an MRI image depicting relevant anatomical structures.

the position and coordination of the different articulators, over time, while potentially avoiding the hyperarticulation effect observed in sustained productions (i.e., the speaker sustains vowel production while a single static image is acquired) (Engwall, 2003). Although typically acquired at the midsagittal plane, any other plane of interest may be considered (Silva et al., 2012).

A wide range of applications can profit from these dynamic studies (Höwing et al., 1999), such as the assessment of swallowing disorders (Kumar et al., 2013) or the characterisation of the articulatory properties of speech (Oliveira et al., 2012; Shosted et al., 2012; Silva et al., 2013). A notable example of the latter is the study of nasal and oral vowels (for example, considering European Portuguese, the second sound in "canto" ([I] sing) and "cato" (cactus)). These have traditionally been considered to differ essentially on the lowering of the velum, for nasal vowels, without any additional articulatory adjustment, but a few studies have recently shown evidence of modifications occurring also at the tongue and lips (Engwall et al., 2006; Carignan, 2011; Shosted et al., 2012a).

The articulation of European Portuguese (EP) nasal vowels has been addressed by the authors and colleagues (Martins et al., 2008; Oliveira et al., 2009), focusing on velum dynamics or using limited tongue information obtained from EMMA studies. Real-time MRI has been acquired (Teixeira et al., 2012) to extend these studies with a characterisation of the oral configuration of EP nasal vowels, important, for example, for articulatory synthesis (Teixeira et al., 2005; Birkholz et al., 2011).

The often used visual comparison among different image frames is of limited use due to the subjective nature of the comparison and the difficulty to cross compare among multiple images. Therefore, relevant data regarding the vocal tract and different articulators must be extracted for analysis. Real-time MRI studies rapidly result in several thousands of images and one of the main challenges concerns how to deal with such a large amount of data in order to extract relevant features and provide researchers with the materials and visualisations that allow systematic analysis (e.g., Silva et al., 2013).

In this scenario, manual segmentation of each image is not only unthinkable, given the large number of images, but also inadvisable in a scenario where a large number of observers might be available to perform the annotation, since the noisy nature of the images makes it difficult to maintain consistency intra- and inter-observer. Therefore, a systematic (semi-)automatic method to extract the relevant data should be used.

1.1. Related work

Analysis of dynamic MRI vocal tract data has typically been performed looking into pattern variations at pixel level, with no segmentation of the anatomical structures of interest (Demolin et al., 2002), or focusing on specific regions, e.g., tongue dorsum (Stone et al., 2001). In recent years, a few authors have presented methods to segment the vocal tract (or specific articulators) from MRI.