Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 47 (2018) 157–174

www.elsevier.com/locate/csl

Optimal sensor placement in electromagnetic articulography recording for speech production study☆

Ashok Kumar Pattem[a], Aravind Illa*[,a], Amber Afshan[b], Prasanta Kumar Ghosh[a]

[a] Department of Electrical Engineering, Indian Institute of Science, Bangalore, Karnataka 560012, India
[b] Department of Electrical Engineering, University of California, 420 Westwood Plaza, Los Angeles, CA 90095, USA

Received 27 April 2017; received in revised form 13 July 2017; accepted 26 July 2017
Available online 27 July 2017

Abstract

Electromagnetic articulography (EMA) is one of the technological solutions, widely used to measure the articulatory movement useful for speech production research. EMA is typically used to track articulatory flesh points by placing sensors, often heuristically, on the key articulators including lips, jaw, tongue and velum in the mid-sagittal plane. In this work, we address the problem of optimal placement of EMA sensors by posing it as the optimal selection of points for minimizing the reconstruction error of the air-tissue boundaries in the real-time magnetic resonance imaging (rtMRI) video frames of vocal tract (VT) in the mid-sagittal plane. We propose an algorithm for optimal placement of EMA sensors using dynamic programming. Experiments are performed using rtMRI video frames for read speech from four subjects with upper and lower lips as two fixed points. One optimal sensor on the upper VT boundary is found to be at an average distance of 21.41(±25.54) mm from the velum tip. Similarly, for the lower VT boundary, one optimal sensor is found at the lower incisor at a distance of 26.37(±8.08) mm from lower lip and three optimal sensors on tongue – at tongue tip (19.93(±11.45) mm from tongue base) and 38.2(±11.52) mm and 80.51(±13.51) mm away from the tongue tip.

Keywords: Electromagnetic articulography; Sensor placement; Speech production

1. INTRODUCTION

Recording of the dynamics of the speech articulators (e.g., lips, tongue, jaw, velum) is critical for the study of speech production (Rubin and Vatikiotis-Bateson, 1998). Articulatory movement data for speech production research are acquired using different modalities such as mid-sagittal X-ray diagrams (Ladefoged et al., 1978), X-ray microbeam imaging (XRMB) (Westbury et al., 1990), Ultrasound (Watkin and Rubin, 1989), Electropalatography (Stone and Lundberg, 1996), tagged MRI (Parthasarathy et al., 2007), Electromagnetic Articulography (EMA) (Maurer et al., 1993) and real-time magnetic resonance imaging (rtMRI) (Demolin et al., 2000; Narayanan et al., 2004). rtMRI provides a complete 2D mid-sagittal view of articulatory dynamics during read speech (Narayanan et al., 2014). Among different modalities, only MRI technique provides a three-dimensional images of the vocal tract for sustained vowels (Demolin et al., 1996). The air-tissue boundaries from rtMRI images provide a time-varying

☆ This paper has been recommended for acceptance by Prof. R. K. Moore.
* Corresponding author.
E-mail address: aravindece77@gmail.com (A. Illa).

http://dx.doi.org/10.1016/j.csl.2017.07.008
0885-2308/

description of the vocal tract shape in the mid-sagittal plane. However, the rtMRI data has low temporal resolution (23.18 frames/s) (Narayanan et al., 2014). It also remains a challenge to record a good quality speech from the subject while he/she undergoes rtMRI scan, due to the loud MRI scanner noise.

Unlike the rtMRI, the XRMB provides the articulatory movement data at a rate more than 100 Hz (Westbury, 1994). In spite of a high temporal resolution, the XRMB technique is limited, in the sense that it does not provide a complete mid-sagittal view of articulatory dynamics, since only a few pellets placed sparsely on various articulators are tracked (Westbury et al., 1990). Ultrasound also provides a high temporal resolution (50 frames/s or more Slłrdahl et al. (2001)) and a good quality audio can be recorded simultaneously too. But the ultrasound images are noisy and detect only the first air-tissue boundary (Bresch et al., 2008). Hence, it is not possible to record the dynamics of anterior tongue tip and lips in ultrasound imaging. On the other hand, EMA has a high temporal resolution (sampling rate of ∼500 Hz). But it cannot capture the structure of pharyngeal wall unlike the rtMRI recording. The EMA data provides the co-ordinates of sensors sparsely placed on different articulators. Another advantage of EMA recording is that a good quality audio can be recorded in parallel to the EMA recording. However, proper care has to be taken while doing the EMA recording to minimize the measurement errors. The accuracy in the measurement of the articulatory movements by EMA is affected due to sensor failures, electromagnetic interference, sensors going out out of the measurement region and also by numerical instabilities (Yunusova et al., 2009; Stella et al., 2012). Attempts are made to handle out of range issues (Kroos, 2008) and to improve the measurement accuracy of EMA (Kroos, 2012; Uchida et al., 2016). For acquiring articulatory movements during speech production, it has been claimed that AG501 provides greater accuracy and is more user-friendly than AG500 (Stella et al., 2013). Therefore, it is apparent that different modalities capture different amount of spatial and temporal information about articulatory movements (Bresch et al., 2008) depending on the imaging technique used or the placement of the sensors and pellets. In this work, we focus on optimal placement of sensors in the mid-sagittal plane for EMA recording such that it provides maximal information about the air-tissue boundaries as observed in rtMRI recording.

EMA data has been crucial for several speech production studies, analysis and modeling including the study of experimental phonetics, the articulatory movement modeling (Perkell et al., 1992; King and Wrench, 1999), examining the variability of coarticulation (Cho, 2004; Bombien et al., 2007; Hardcastle et al., 1996; Recasens, 2002; Hoole et al., 1993; Hoole and Gfoerer, 1990; Hoole and Nguyen, 1997; Mooshammer and Hoole, 1993; Mooshammer and Schiller, 1996; Katz et al., 1990; West, 2000) understanding coupling dynamics (Van Lieshout, 2001; Van Lieshout et al., 2002) of motor primitives in speech movements in case of the normal and disordered speech (Schulz et al., 2000; Maassen et al., 2007; Van Lieshout, 2007; Van Lieshout et al., 2007) as well as during stuttering (Peters et al., 2000; McClean and Runyan, 2000; Namasivayam and Van Lieshout, 2001; 2008) and swallowing (Steele and Van Lieshout, 2004; 2005; Bennett et al., 2007; Steele and Van Lieshout, 2009). EMA data of the articulatory kinematics available through MOCHA-TIMIT (Wrench, 2000) and USC-TIMIT (Narayanan et al., 2014) are widely used for acoustic-articulatory modeling for speech recognition (Frankel et al., 2000; Wrench and Richmond, 2000; Richardson et al., 2003), text-to-articulatory-movement prediction and analysis of critical articulators Zhang and Renals (2008); Ling et al. (2010), mapping from articulatory movements to vocal tract spectrum (Payan and Perrier, 1997; Toda et al., 2004b; Steiner et al., 2013), acoustic-to-articulatory inversion (Toutios and Margaritis, 2003; Toda et al., 2004a; Ghosh and Narayanan, 2010; Uria et al., 2011; Ghosh and Narayanan, 2011), multimodal speech animation (Kim et al., 2014; Engwall, 2003).

Given these wide-spread uses of the EMA data, it is important to develop a principled approach in the placement of sensors during EMA recording. Since EMA data provides movement of few sparsely placed sensors, it is required to place them optimally in order to capture maximal information of the articulatory dynamics. For most of the existing EMA recordings, the sensors are typically placed following some heuristic rules. For example, for recording using Carstens system (AG100), the suggested three EMA sensors positions on the tongue are 1 cm from the tongue tip, midpoint of the tongue body and 4 cm from the tongue tip as tongue dorsum (UCLA, 2017). The TORGO Database of Dysarthric Articulation (Rudzicz et al., 2012) was recorded from dysarthria patients. It consists of both acoustics and articulatory data of EMA and 3D-reconstruction from binocular video sequences. The sensors were placed on the tongue at three different locations, namely, tongue tip being 1 cm, tongue middle 4 cm and tongue back 6 cm behind from the anatomical tongue tip. In another study of pharyngealization using an articulograph (Ouni and Laprie, 26–27 March, 2009), EMA data was collected by placing four sensors at 1.6, 3.6, 5.2 and 7 cm away from the tongue tip. Mücke et al. (2012) collected articulatory data from German speakers and they used only two sensors on the tongue, at 1 and 4 cm away from the tongue tip and called them tongue blade and tongue body,