Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 36 (2016) 307–329

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Quantitative systematic analysis of vocal tract data☆

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

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

Received 26 June 2014; received in revised form 21 May 2015; accepted 25 May 2015
Available online 3 June 2015

Abstract

Articulatory data can nowadays be obtained using a wide range of techniques, with a notable emphasis on imaging modalities such as ultrasound and real-time magnetic resonance, resulting in large amounts of image data.

One of the major challenges posed by these large datasets concerns how they can be efficiently analysed to extract relevant information to support speech production studies. Traditional approaches, including the superposition of vocal tract profiles, provide only a qualitative characterisation of notable properties and differences. While providing valuable information, these methods are rather inefficient and inherently subjective. Therefore, analysis must evolve towards a more automated, replicable and quantitative approach.

To address these issues we propose the use of objective measures to compare the configurations assumed by the vocal tract during the production of different sounds. The proposed framework provides quantitative normalised data regarding differences covering meaningful regions under the influence of various articulators. An important part of the framework is the visual representation of the data, proposed to support analysis, and depicting the differences found and corresponding direction of change.

The normalised nature of the computed data allows comparison among different sounds and speakers in a common representation.

Representative application examples, concerning the articulatory characterisation of European Portuguese vowels, are presented to illustrate the capabilities of the proposed framework, both for static configurations and the assessment of dynamic aspects during speech production.

Keywords: Vocal tract analysis; Quantitative comparison; RT-MRI

1. Introduction

Speech production studies are currently served by a wide range of technologies that allow research on the dynamic aspects of speech. Methods such as ultrasound (US) and real-time magnetic resonance imaging (RT-MRI) (Scott et al., 2014) provide data regarding the position and coordination of the different articulators over time (Hagedorn et al., 2011). Furthermore, they offer the possibility to improve on the studies based on information regarding a static sustained production by reducing the hyperarticulation effect (Engwall, 2003).

☆ This paper has been recommended for acceptance by Shrikanth Narayanan.

* 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.2015.05.004
0885-2308/

After image acquisition, the different regions of interest must be segmented (e.g., Bresch and Narayanan, 2009; Silva and Teixeira, 2015), or points of interest identified, often resulting in contours delimiting the vocal tract or specific structures such as the tongue or velum.

Analysis of different vocal tract contours is typically performed visually by characterising the position of the different articulators or by describing articulator differences between different sounds (e.g., Delvaux et al., 2002; Shadle et al., 2008). This is often done by superimposing contours and performing qualitative analysis of the main differences (Martins et al., 2008; Cleland et al., 2011; Badin et al., 2014). Adding to the subjective nature of such analysis, when the database is large, e.g., as happens when RT-MRI is used (Niebergall et al., 2013), it becomes an almost infeasible task to explore all available data.

1.1. Challenges

Beyond the sheer amount of data made available by current technologies, the field of speech production faces several challenges that should be addressed to allow further advances, harnessing the full potential of the data available. A framework should be proposed that tackles the large amounts of data addressing, among others, the following aspects:

- **Objectivity** — The subjective nature of the methods used to describe articulatory differences, for example, results in variability among researchers that precludes true comparison among works in the literature describing the same phenomena.
- **Intra-speaker assessment** — The analysis of articulatory features for different sounds produced by one speaker lacks methods to profit from multiple repetitions and common grounds for comparison among sounds.
- **Inter-speaker assessment** — The comparison among speakers lacks common grounds for comparison, e.g., a common normalised measure of difference, without losing sight of the contributions provided by each articulator.
- **Variability** — Not only average behaviour is relevant for the researchers, there is also a strong need to have information on variability (across repetitions, across speakers, etc.).
- **Inter-language comparison** — Data from multiple speakers of one language could be jointly used to provide overall quantitative characterisation of its main features. This would allow new ways of comparing sounds in multiple languages, advancing on the current status of inter-language comparisons mostly based on qualitative assessment of data from different speakers.
- **Multimodality** — Data provided by different modalities and concerning similar phenomena or providing complementary data might benefit from joint analysis. For example, several technologies that support speech production studies (e.g., EMA, Kim et al., 2014 and ultrasound, Laprie et al., 2014) are used in combination with MRI (Scott et al., 2014). Regardless of how the different data is analysed, if their individual contributions to the understanding of specific phenomena could be gathered in joint representations it might motivate a generalisation of multimodal studies and an easier interpretation of the data.

One important route to attain a systematic analysis addressing these issues is to move towards quantitative methods that allow it to be performed automatically, in an expedite and replicable way, resulting in data providing a summary of the most important features which researchers can analyse. In the work presented here we consider these challenges in the scope of real-time MRI data.

1.2. Related work

Considering that we are mostly addressing vocal tract data analysis from vocal tract profiles extracted from image data (e.g., RT-MRI), in this overview of related work we focus on methods applied to image data of the vocal tract, or to full (or partial) vocal tract profiles extracted from them. The rationale is to overview notable recent literature that describes some level of quantification regarding the analysis and comparison of the data extracted from these images. Therefore, works supported on subjective analysis based, for example, on visual assessment of articulatory configuration differences, are not covered.

Regarding pixel-based methods, i.e., without an explicit extraction of vocal tract profiles, notable works include the detection of constriction regions along the tract (Lammert et al., 2010; Hagedorn et al., 2011) and estimation of