Segmentation and Analysis of Vocal Tract from MidSagittal Real-Time MRI

Samuel Silva[1], António Teixeira[1], Catarina Oliveira[2], and Paula Martins[2]

[1] DETI/IEETA, University of Aveiro, 3810–193 Aveiro, Portugal,
[2] ESSUA/IEETA, University of Aveiro, 3810–193 Aveiro, Portugal
{sss,ajst,coliveira,pmartins}@ua.pt

Abstract. The articulatory description of European Portuguese (EP) requires the analysis of different anatomical structures (e.g. tongue dorsum and velum), and the study of dynamic aspects of speech production. The use of real-time magnetic resonance imaging (RT-MRI), with frame rates above 10 frames/s, provides adequate support for these studies and results in a large amount of images that need to be processed to extract relevant data to be analysed by linguists. To tackle the required data processing and analysis this article presents methods to perform segmentation of the vocal tract from midsagittal real-time MR image sequences and provide researchers with visualizations of the relevant extracted data. Examples are provided illustrating the analysis of dynamic aspects of EP nasal vowels.

1 Introduction

The acoustic and articulatory properties of nasal vowels have been the subject matter of several studies over the past decades using a large variety of techniques. Most of these studies have focused on velum activity and on the effects of coupling between the naso-pharyngeal and oral tracts.

The main difference between nasal and oral vowels (e.g., the second sound in "canto" ([I] sing) and "cato" (cactus)) has traditionally been considered to reside essentially on the lowering of the velum, for nasal vowels, without any additional articulatory adjustment. Nevertheless, recent articulatory studies have shown evidence that modifications in both tongue and lips might also occur [1–3].

The articulation of European Portuguese (EP) nasal vowels has been addressed by the authors in several studies (e.g. [4, 5]), but with an emphasis on velum dynamics or with limited information regarding the tongue provided by electromagnetic midsagittal articulography (EMMA). To extend these studies, with a characterization of the oral configuration of EP nasal vowels, important, for example, for articulatory synthesis, real-time magnetic resonance imaging (RT-MRI) data of the vocal tract was acquired. This imaging modality provides adequate data regarding the position and coordination of the different articulators over time and might also be advantageous since it avoids the hyper-articulation effect observed in sustained productions (i.e., the speaker sustains vowel production while a single static image is acquired) [6].

M. Kamel and A. Campilho (Eds.): ICIAR 2013, LNCS 7950, pp. 459–466, 2013.

To perform a systematic study on this subject, one of the main challenges concerns how to deal with the large amount of data resulting from the RT-MRI in order to provide linguists with the data and visualizations that allow analysis. The comparison among different image frames is of limited use and, therefore, relevant data regarding the vocal tract and different articulators must be extracted for analysis.

Analysis of dynamic MRI vocal tract data has typically been performed looking into pattern variations at pixel level, without a segmentation of the anatomical structures of interest [7], or focusing on specific regions, e.g. tongue dorsum [8].

Following on work previously presented for the segmentation of the oral and nasal cavities from coronal-oblique RT-MRI [9] we present a segmentation method for the vocal tract from midsagittal real-time MR image sequences and provide methods to allow quick and systematic exploration of the existing data by linguists and researchers interested in model development or even articulatory synthesis [10].

This article is organized as follows: section 2 briefly describes the image database; in section 3 the proposed methods are presented followed, in section 4, by examples illustrating how the obtained data can be used for articulatory analysis. Finally, section 5 presents some conclusions and ideas for future work.

2 Image Database

Image sequences were acquired containing: a) the five European Portuguese (EP) nasal vowels uttered in three word positions: initial, internal and final (e.g. the nonsense words "**am**pa, p**am**pa, p**an**" or "**em**pa p**em**pa p**en**"); and b) the eight EP oral vowels (e.g., "p**a**pa" or "p**u**pa").

Images were acquired at the midsagittal plane of the vocal tract (see figure 1 for example image with notable anatomical landmarks highlighted) using an Ultra-Fast RF-spoiled Gradient Echo (GE) pulse sequence and yielding a frame rate of 14 frames/second. Each recorded sequence contained 75 images.

Audio was recorded simultaneously with the RT images inside the MR scanner, at a sampling rate of 16000 Hz, using a fiberoptic microphone and manually annotated, using the software tool Praat (`http://www.fon.hum.uva.nl/praat/`), in order to identify the time intervals corresponding to different sounds. The time intervals allow the determination (because both data are aligned) of the corresponding image frames.

Data was acquired for three female speakers, aged between 21 and 33, phonetically trained, with no history of hearing or speech disorders.

Further details concerning the image acquisition protocol and corpus can be found in [11].

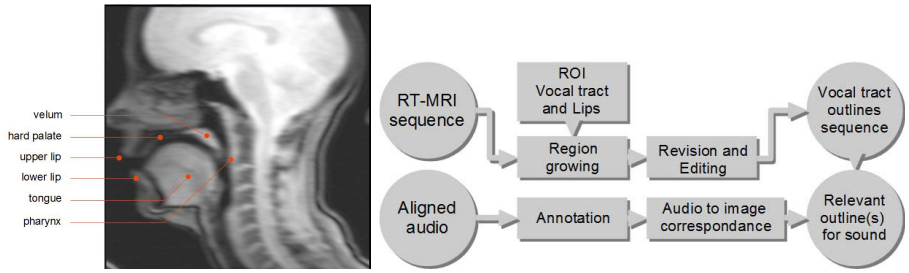

Fig. 1. Left, sample midsagittal MRI image of the vocal tract with notable anatomical regions highlighted. Right, pipeline depicting the main steps involved in processing the acquired image and audio data.

3 Segmentation and Analysis Methods

The proposed framework provides methods to allow a quick segmentation of the image sequences and to allow easier exploration of the resulting data. A description of the devised methods (pipeline depicted on figure 1) is provided in what follows.

3.1 Image Segmentation

The first approach followed considered the vocal tract as a whole, from lips to pharynx and including the nasal cavity. Initial experiments revealed that, due to slightly different characteristics regarding intensities, size of structures and artefacts resulting from proximity between structures (e.g. lips or velum positioned near the tongue or pharynx) the segmentation results could be improved if several regions were considered: the lips, the velar region and the remaining vocal tract.

Segmentation starts with the definition of a region of interest (ROI) roughly encompassing the vocal tract. This ROI just needs to be defined over one of the image frames and, since there is spatial coherence among images acquired for a speaker, it can be replicated (as happens with the other ROIs) over the entire image series. Its main purpose is to avoid the segmentation to go beyond the nasal cavity and above the hard palate. A second ROI is defined to encompass the lips and a third to encompass the velar region and defined over the first since its purpose is just to refine the region between velum and pharynx. Although not particularly important, an additional ROI can be defined to allow chin profile segmentation (for example, enlarging the lips region and defining a sub-region over the chin). The chin is not used for analysis but provides context to the vocal tract profile. These ROIs are defined using Live Wire in order to ease their definition following anatomical structures' boundaries (e.g., pharynx and hard palate). Figure 2a shows the different ROIs defined for one of the speakers.

Region growing is then applied to each of the ROIs (over all the image series) by manually defining a seed inside of each, in a hypointense region belonging

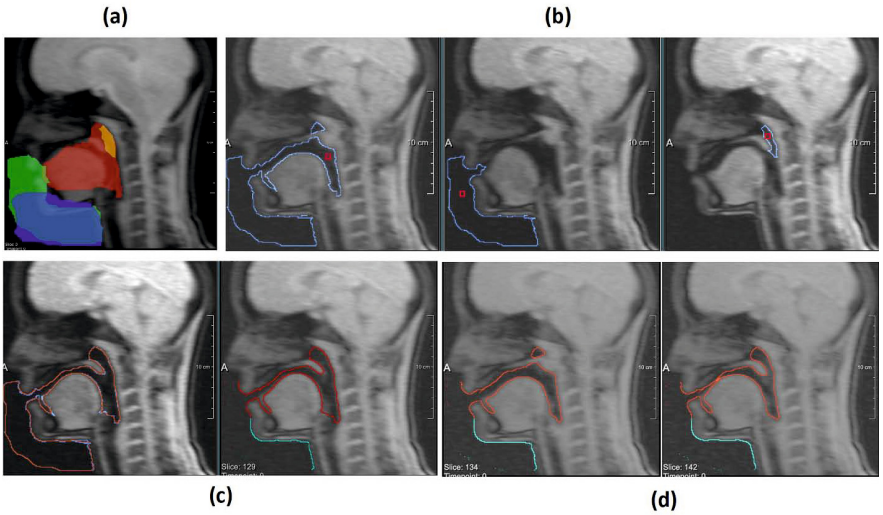

Fig. 2. Different stages of segmentation: (a) Definition of ROIs; (b) Definition of a seed inside each ROI to apply region growing (only region contours shown); (c) Add contributions from different regions, convert to contours and eliminate those which are not relevant; (d) examples of final vocal tract segmentations

to the vocal tract (figure 2b). Since different regions are being processed, the region growing parameters can be tuned for the different characteristics of each. For example, for the lips the intensity interval is often narrower than for the remaining vocal tract.

Given the spatial coherence alluded above and similar image acquisition conditions between series, the ROIs, seed positions and intensity intervals for the region growing are roughly the same for all image series and, therefore, need only the be defined once for each speaker. The different segmented regions are added and the output is presented to the user by depicting just the contour of the resulting segmentation (figure 2c). Any contour not belonging to the vocal tract is removed by detecting those situated beyond the lips that do not vary much over the sequence.

Implementation was performed using MeVisLab (MeVis Medical Solutions, `http://mevislab.de`).

3.2 Editing

Even considering the different ROIs there are situations when the proximity between articulators might render it impossible to separate between the two without over segmenting the remaining vocal tract. For these cases an editing tool can be used to perform the separation. Typically the regions that need correction are small sized and the proposed tool is quite simple and allows the user to add regions to the vocal tract by defining contours around them using Live Wire. The use of Live Wire just helps the user follow tongue or velum

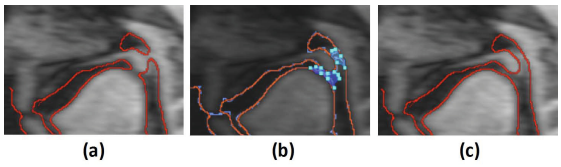

Fig. 3. Editing example: a) initial conditions, with velum connected to pharynx and tongue; b) added regions; and c) final result with the velum separated from the tongue and pharynx

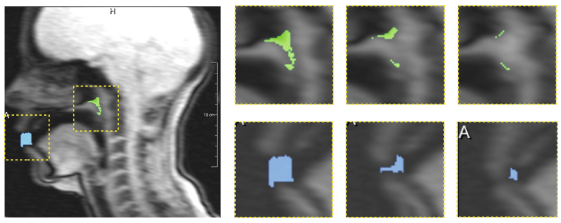

Fig. 4. Velum heigh and lip aperture data extracted based on the number of active pixels inside a ROI

contours more easily (e.g., when separating the tongue from the velum). Figure 3 shows an editing example. On the first image notice that the velum is connected to the pharynx and to the tongue. The user then added regions over the image, separating the velum from the other structures.

The segmentation of a time series (75 image frames), including a revision of the full sequence, takes one to five minutes depending on if editing is required.

3.3 Velum and Lips Movement Extraction

Adding to the segmentation of the vocal tract, which already allows the comparison of different vocal tract configurations over time, it might also be relevant to extract specific data regarding particular articulators such as the velum or lips. For example, assessment of velum movement is important to analyse differences in velum height between the vowels or differences in velum movement throughout the production of the vowel.

To gather data concerning velum movement, an image presenting the velum completely lowered is chosen and a region of interest defined between the velum and pharynx (which is a hypointense region). Then, region growing is applied to that region over the whole image series and the number of active pixels inside it computed. When the velum is lowered, the number of active pixels inside the defined ROI is higher than when the velum is closed (figure 4).

To extract data concerning lip movement a similar approach is used, but considering a ROI between the lips, defined using an image where they are fully open.

The definition of these ROIs does not need to be very accurate. As long as they include the region between the lips and the region between the lowered velum

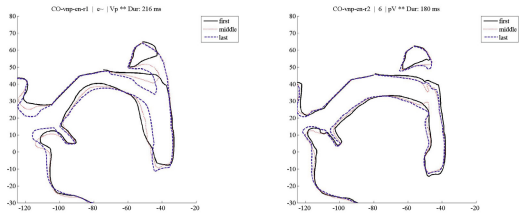

Fig. 5. On the left, first, middle and last vocal tract profiles (on the annotated interval) during the production of [ẽ] ("**em**pa"). On the right, vocal tract profiles along the production of [ɐ] ("empa**a**").

and the pharynx, the part of the ROI that is not affected by the movement (and therefore remains constant over most images) is detected and removed.

3.4 Analysis of Extracted Data

A first approach to analyse the extracted data is to allow the user to load specific vocal tract profiles and superimpose them for comparison. This can be easily performed using MeVisLab by providing a simple interface where the user can choose the speaker, image series and image frame (based on the annotated audio). This allows a first exploration of the data and a preliminary discussion of the articulatory differences between oral and nasal vowels, based on comparisons using this first approach, can be found in [12]. Nevertheless, this is still a rather non-systematic time consuming task. Therefore, based on the segmented data and on the image frame intervals, identified using the annotated audio, an automatic method was implemented to generate figures with the different vocal tract profiles along the production of each sound or from different sounds at a specific time frame (typically at the centre of the relevant time interval) for comparison (e.g., oral vowel and its nasal counterpart). This is also an important step towards automated analysis of the data (e.g., automatic profile comparison).

Line graphs are also generated depicting velum and lip movement data over the entire image sequences and for the image frame intervals corresponding to each nasal vowel. Velum movement can then be further characterized by computing the different phases of its movement (opening, plateau and closing times).

4 Application Examples

To illustrate how the extracted data and created figures can help to obtain important data, a few examples are presented. These do not intend to illustrate articulatory tendencies, but how the provided figures and graphs allow to detect differences and dynamic aspects of the articulators.

Figure 5, on the left, shows a superposition of the vocal tract profiles obtained from the image frames acquired along the production of [ẽ]("**em**pa"). It can be

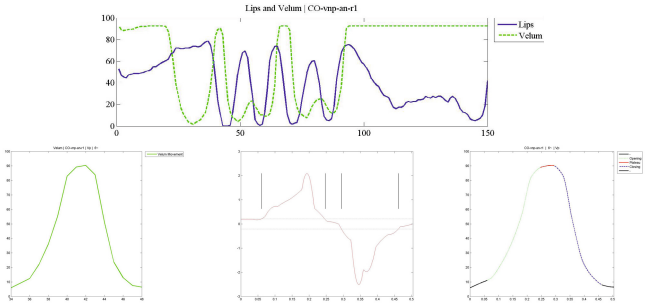

Fig. 6. Top, lips and velum movement curves for the sequence "ampa pampa pan"; bottom, from left to right: velum movement curve for [ɐ̃] as in "**am**pa"; first order derivative of velum movement curve with 10% of maximum used as criterion to detect transitions; and velum movement curve with different segments of interest identified

noted that differences occur at the velum, gradually opening, at the tongue, which moves slightly back and at the lips. On the right, the vocal tract profiles for the oral vowel [ɐ] ("emp**a**") are shown. Notice how the velum region has no movement.

Figure 6 shows velum and lip movement curves (the number of active pixels in each of the ROIs, per frame) obtained while the speaker is uttering "ampa pampa pan". The minima of the lips curve correspond to moments where the lips are closed (uttering [p]) and the maxima of the velum curves correspond to when a nasal vowel is produced ([ɐ̃], "**am**pa p**am**pa p**an**"). Notice that at the beginning and end of the sequence the velum is open because the speaker is breathing. These curves are important to provide a first idea on how the coordination between lips and velum movement occurs and illustrates the dynamic nature of both articulators along the nasal vowels. For example, one important aspect to note is that when the lips are closed, the velum is still slightly open.

The bottom row of figure 6 shows the velum movement curve for the [ɐ̃] in "p**am**pa", the first order derivative used to detect transitions between the movement phases and their depiction over the curve: opening, plateau and closing.

5 Conclusions and Future Work

This article presents a framework for analysis of midsagittal RT-MRI images of the vocal tract including a segmentation method and automated generation of images of relevant phenomena for analysis. The devised methods, although simple in nature, allow a quick and robust way of dealing with the main challenge of processing the large amount of image data gathered using real-time MRI. The data extracted, and the methods used to process it, allow comparison of vocal tract configurations over time and over the different nasal and oral congeners of each vowel.

The existing RT-MRI database (a rare resource, even considering other languages), and the proposed framework constitute a valuable source of information to support linguistic description, anatomical modelling or articulatory synthesis (to determine coordination and temporal aspects).

Further work is still envisaged to improve the proposed framework. Even though the final results are supervised by a radiographer, evaluation of the segmentation method to assess its performance is an important step. Furthermore, additional methods should be provided to support the comparison among the different vocal tract contours by using some local measure of difference and adequate representation of its outputs. This would allow a more quantitative comparison, particularly among data from different speakers (after normalization), and reducing the need for user dependant judgements.

Acknowledgements. Research was partially funded by FEDER through the Operational Program Competitiveness factors (COMPETE) and by National Funds through Foundation for Science and Technology (FCT) in the context of the Project HERON II (PTDC/EEA-PLP/098298/2008). IEETA is funded by FEDER through the Operational Program Competitiveness Factors - COMPETE and by FCT – FCOMP-01-0124-FEDER-022682 (FCT reference PEst-C/EEI/UI0127/2011). The fourth author acknowledges the PhD grant from FCT (SFRH/BD/65183/2009).

References

1. Engwall, O., Delvaux, V., Metens, T.: Interspeaker variation in the articulation of French nasal vowels. In: Proc. 7th Int. Sem. Speech Prod. (2006)
2. Carignan, C.: Oral articulation of nasal vowels in French. In: 17th ICPhS, pp. 408–411 (2011)
3. Shosted, R., Carignan, C., Rong, P.: Managing the distinctiveness of phonemic nasal vowels: Articulatory evidence from Hindi. JASA 131(1), 455–465 (2012)
4. Martins, P., Carbone, I., Pinto, A., Silva, A., Teixeira, A.: European Portuguese MRI based speech production studies. Speech Comm. 50(11-12), 925–952 (2008)
5. Oliveira, C., Martins, P., Teixeira, A.: Speech rate effects on European Portuguese nasal vowels. In: Proc. InterSpeech 2009, Brighton (2009)
6. Engwall, O.: A revisit to the application of MRI to the analysis of speech production - testing our assumptions. In: Proc. 6th Int. Sem. Speech Prod. (2003)
7. Demolin, D., Hassid, S., Metens, T., Soquet, A.: Real-time MRI and articulatory coordination in speech. Comptes Rendus Biologies 325(4), 547–556 (2002)
8. Stone, M., Davis, E.P., Douglas, A.S., Aiver, M.N., Gullapalli, R., Levine, W.S., Lundberg, A.J.: Modeling tongue surface contours from cine-MRI images. J. Speech Lang. Hear. Res. 44(5), 1026–1040 (2001)
9. Silva, S., Martins, P., Oliveira, C., Silva, A., Teixeira, A.: Segmentation and analysis of the oral and nasal cavities from MR time sequences. In: Campilho, A., Kamel, M. (eds.) ICIAR 2012, Part II. LNCS, vol. 7325, pp. 214–221. Springer, Heidelberg (2012)
10. Teixeira, A.: Síntese Articulatória das Vogais Nasais do Português Europeu. Phd thesis, Universidade de Aveiro (2000)
11. Teixeira, A., Martins, P., Oliveira, C., Ferreira, C., Silva, A., Shosted, R.: Real-time MRI for Portuguese: database, methods and applications. In: Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F. (eds.) PROPOR 2012. LNCS, vol. 7243, pp. 306–317. Springer, Heidelberg (2012)
12. Oliveira, C., Martins, P., Silva, S., Teixeira, A.: An MRI study of the oral articulation of European Portuguese nasal vowels. In: Proc. InterSpeech 2012 (2012)