Towards a Systematic and Quantitative Analysis of Vocal Tract Data

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

[1]DETI/ IEETA, University of Aveiro (Aveiro, Portugal)
[2]ESSUA/ IEETA, University of Aveiro (Aveiro, Portugal)
sss@ua.pt, ajst@ua.pt, coliveira@ua.pt, pmartins@ua.pt

Abstract

Articulatory data can nowadays be obtained using a wide range of techniques, such as real-time magnetic resonance (RT-MRI), enabling acquisitions of large amounts of data. A major challenge arises: analysing these new large data sets to extract meaningful information regarding speech production in an expedite and replicable way. Traditional approaches such as superimposing vocal tract profiles and qualitatively characterizing relevant properties and differences, although providing valuable information, are rather inefficient and subjective. Therefore, analysis must evolve towards a more automated, quantitative approach. To tackle this issue 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 data regarding differences pertaining meaningful regions under the influence of various articulators. Visual representation of such data is a key part of the proposal and some concrete forms of visualization are proposed to depict the differences found and corresponding direction of change. Application examples concerning the articulatory characterization of EP vowels are presented with promising results, paving the way towards automated and objective analyses of articulatory data.

Index Terms: vocal tract analysis, quantitative comparison, real-time MRI

1. Introduction

The articulation of European Portuguese (EP) nasal vowels has been studied by the authors mainly focusing on velar dynamics as provided by electromagnetic midsagittal articulography (EMMA) [1]. To extend these studies, with a characterization of the oral configuration of EP nasal vowels, static [2] and, more recently, real-time magnetic resonance imaging (RT-MRI) data of the vocal tract was acquired [3, 4]. This imaging modality provides adequate data regarding the position and coordination of the different articulators over time [5] and might provide a good choice to tackle the hyperarticulation effect observed in sustained productions [6]. Additionally, it might also help reduce the gravity effect on articulators for acquisitions in supine position [7].

After image acquisition the different regions of interest must be segmented (e.g., [8]), or points of interest identified, often resulting in contours delimiting the vocal tract and/or specific structures such as the tongue or velum.

Analysis of different vocal tract contours is typically performed visually by characterizing the position of the different articulators or by describing articulator differences between different sounds (e.g. [9, 10]). This is often done by superimposing contours and performing qualitative analysis of the main differences. Adding to the subjective nature of such analysis, when the database is large, as happens when RT-MRI is used [11], it becomes an almost infeasible task to explore all available data.

To attain a systematic analysis it is important to define quantitative methods that allow it to be performed automatically, in an expedite and replicable way, resulting in data/visualizations depicting a summary of the most important features which researchers can analyse. Quantitative analysis is also important to reduce variability among characterizations performed by different researchers, provides grounds to perform comparisons intra- and inter-speaker and, on the long run, inter-language comparisons.

Quantitative analysis of the vocal tract has already been performed, to some extent, by several authors. Notable works include the detection of constriction regions along the tract [12, 5] and estimation of articulator trajectories [13, 14] contributing to improved analysis of articulatory gestures and their coordination. Other authors have performed quantitative comparison of specific regions such as the tongue (e.g. [15]). When considering the whole vocal tract, Proctor et al. [16] have extracted distance functions, which might not be a clear way to detect which articulator has the most influence, and Ramanarayanan et al. [17] have proposed a set of vocal tract descriptors covering lip aperture, tongue tip, dorsum and root constrictions and velum aperture.

Nevertheless, to the best of our knowledge, no method has been proposed to support a more complete vocal tract profile comparison, providing meaningful data regarding the regions under the influence of the different articulators.

In a first attempt to deal with this issue, particularly when considering large amounts of data, we propose that: 1) differences between vocal tract profiles should be obtained using objective, comparable measures; 2) differences must be computed for the different anatomical regions of interest to provide a meaningful regional measure of difference; 3) difference should not be limited to a simple number but provide information on how the difference occurred (e.g. in what direction did the tongue back move); 4) meaningful visual representations should be provided to help users understand the resulting data.

In this paper we present a set of methods that are a first approach to articulatory analysis based on these goals.

This paper is organized as follows: Section 2 presents a description of the proposed methods for vocal tract comparison and representation of the resulting data; Section 3 applies those methods to the articulatory characterization of EP oral and nasal vowels providing a brief illustration of their application in a real scenario; finally, Section 4 presents conclusions and ideas for further work.

2. Methods

Analysis of vocal tract data, using an image modality such as RT-MRI, involves a set of steps including image acquisition, segmentation, identification of the relevant frames for analysis and extraction of information to support articulatory characterization (see figure 1).

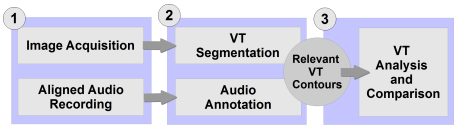

Figure 1: Pipeline depicting main steps involved in articulatory data analysis.

To provide context to the methods being proposed we briefly present the main aspects of the targeted data and corpus and summarize the pre-processing steps already carried out.

2.1. Data Acquisition and Processing

The RT-MRI image acquisition protocol (further details in [18, 3]) provides a frame rate of 14 frames/s. The corpus, which has been acquired for three speakers, includes: 1) the five EP nasal vowels ([ɐ̃], [ɛ̃], [ĩ], [õ], [ũ]) in word initial, medial and final positions (e.g. [ɐ̃pɐ], [pɐ̃pɐ], [pɐ̃]); and 2) the eight oral vowels ([a], [ɐ], [ɛ], [e], [i], [ɔ], [o] and [u]) inserted in CV_1CV_2 sequences (e.g. papa [papɐ], pupa [pupɐ]), where C is a voiceless bilabial plosive. A synchronized audio signal has been recorded during image acquisition, using a fiberopic microphone, and annotated manually using Praat [19], allowing the identification of which image frames correspond to each sound.

Vocal tract segmentation has been performed semi-automatically for all image sequences as described in Silva et al. [20] and resulted in sequences of vocal tract contours as those presented in figure 2. A first approach to the analysis of the articulatory characteristics of EP vowels, including the dynamic behaviour of articulators during nasals, has been presented in [3]. Even though it provided interesting results, the methods used to explore the data were still rather inefficient to handle such a large amount of data with contour selection, superposition and subjective analysis performed by the researchers.

The methods presented in what follows are a first step towards a more objective and systematic approach to the vocal tract data available. These consider that vocal tract contours have already been extracted from the image sequences and the relevant image frames automatically identified based on the audio annotation [3].

2.2. Vocal Tract Comparison

The following sections describe how the different anatomical regions of interest are identified, in the segmented vocal tract profiles, and how we propose to analyse the variations of corresponding regions between profiles.

2.2.1. *Anatomical Regions of Interest*

The identification of the different anatomical regions [8] in the segmented contours is performed with the help of landmark points defined manually (see figure 2). Since speakers kept their position throughout the whole acquisition session, the landmarks defined for each speaker can be used for the whole image set. Additional processing is necessary in order to separate the tongue back from the tongue dorsum, to locate the tongue tip and to consider just the bottom side of the velum (for nasal vowels). This is roughly based on the analysis of the first derivative of the contour segments identified using the landmarks and is performed automatically.

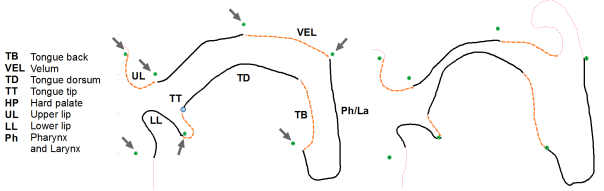

Figure 2: Different vocal tract regions identified using manually defined landmark points and analysis of the first derivative (for tongue back/tongue dorsum separation, tongue tip identification and bottom side of velum).

2.2.2. *Comparison of VT Regions*

The **pharynx/larynx** (Ph), **tongue back** (TB), **tongue dorsum** (TD) and **velum** (VEL) are compared by computing the Pratt index [21] for each pair of corresponding contours and given by:

$$P = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+\alpha d_i^2},$$

where N is the number of corresponding points between the compared contours (e.g., tongue back), d_i is the euclidean distance between two corresponding points, and α is a constant set to 1/9, based on Pratt's work and similar works in the literature. At this stage the same constant value has been used for all regions but it might be tuned for each region if different sensibilities to differences were desired. To obtain corresponding points between contours, the contour with the smallest number of points is selected and for each point in this contour the closest point in the other contour is considered the corresponding point. The Pratt's figure of merit provides values in the range $]0, 1]$ where 1 is attained when there are no differences between the contours.

The **tongue tip** (TT) position is compared by computing the distance between the tongue tips in both contours normalized by the longest distance from each tongue tip to the alveolar ridge (AR).

$$TT = 1.0 - \frac{d_{TT}}{max(dA_{(TT-AR)}, dB_{(TT-AR)})}.$$

Lip protrusion (LP) is obtained by computing the horizontal displacement of the mid-point between the upper and lower lips (LMP). The mid-point is computed considering the line that connects the lowest point of the upper lip with the highest point of the lower lip. To perform normalization the horizontal distance between the mid-points and the alveolar ridge (AR) is obtained ($dA_{(LMP-AR)}$ and $dB_{(LMP-AR)}$) and used as follows:

$$LP = 1.0 - \frac{\left|dB_{(LMP-AR)} - dA_{(LMP-AR)}\right|}{max(dA_{(LMP-AR)}, dB_{(LMP-AR)})}.$$

Lip aperture (LA) is computed based on the lip aperture values for both vocal tract profiles (LA_A, LA_B) and is normalized by considering the longest of the two:

$$LA = 1.0 - \frac{|LA_B - LA_A|}{max(LA_A, LA_B)}.$$

2.3. Visualization

This set of seven comparison values, corresponding to the seven anatomical regions/features provide quantitative evidence on the differences between two vocal tract configurations. Nevertheless, the analysis of a set of numbers on a table is still difficult to interpret/compare.

Therefore, to attain a visual representation of the difference data we propose to represent it in a diagram as the one depicted in figure 3. Each anatomical region corresponds to an angular orientation, in close relation to its position in the vocal tract. Starting at the zero degrees position and moving counterclockwise: tongue back (TB), velum (VEL), tongue dorsum (TD), tongue tip (TT), lip protusion (LP), lip aperture (LA) and pharynx/larynx (PH).

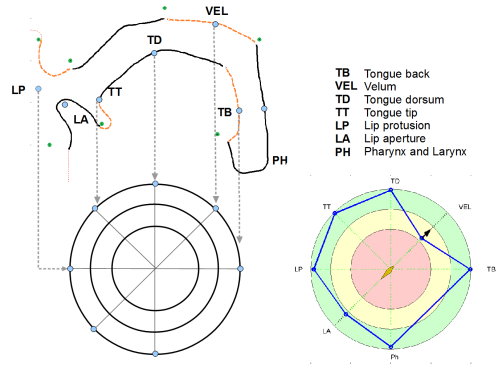

Figure 3: Each vocal tract region is associated with an angular orientation and the computed difference value is depicted over each axis with zero represented at the origin and one over the outer circumference. An example difference diagram is presented on the bottom right.

The comparison value corresponding to each region is represented by a point over its axis and all the points connected to form a polygon. Circumferences are added to the representation to provide reference for the values 0.75 and 0.5. The resulting circular coronas are coloured according to a possible interpretation of the associated values (similar to what is typically performed for the Pratt index): green, non-significant difference ($[0.75, 1]$); yellow, mild difference ($[0.50, 0.75]$); and red ($[0, 0.50]$), strong difference.

When representing the average difference polygon obtained, for example, from multiple cross comparisons, it might also be important to represent variability data such as the standard deviation, for each of the parameters. As an initial proposal we depict this data by showing a polygon, centred at the origin, with each vertex position defined by the standard deviation represented over each parameter's axis. Figure 5 presents some examples of difference diagrams with standard deviation presented at their centre.

2.3.1. *Visualization Calibration*

The vocal tract profiles obtained for the different occurrences of the same vowel, over all the image sequences, present some variability due to the segmentation method or due to naturally occurring differences in the articulator positions. When performing a quantitative analysis these slight differences work as an offset that might add to the relevant differences. In order to tackle this issue a possible approach is to cross compare all the occurrences for each vowel to assess how much variability is found. Then, the average difference considering all in-vowel differences can be subtracted from all difference computations, i.e., the difference diagram for comparisons inside each vowel will tend to present values close to 1.0.

This calibration is not performed blindly. It takes in consideration that the differences found among all occurrences of the same vowel were small (i.e., falling on the green circular corona in the difference diagrams). This is important because the intent is not to disguise important differences (if they exist). For our data, relevant differences have not been observed for the cross comparisons inside each vowel.

All difference diagrams presented in what remains of this paper have been subject to calibration.

2.3.2. *Articulator Movement Direction*

The diagram can also integrate additional data on how the depicted difference occurred. For example, if there is a significant difference in the tongue back between two vocal tract configurations, which direction did the tongue back move from the first to the second configuration? To provide such data our current approach is to compute the centre of gravity for the contour and then depict how it moved over the diagram. We chose only to represent this movement over each region axis, i.e., it will only point to the centre or out. For the sake of simplicity, movement direction is only represented for comparison parameters presenting values below 0.75. Figure 4 illustrates the comparison between vowels [ẽ] and [e] presenting the vocal tract profiles for all occurrences of both vowels, all the comparison polygons resulting from the cross comparisons and the average polygon depicting each region's movement direction.

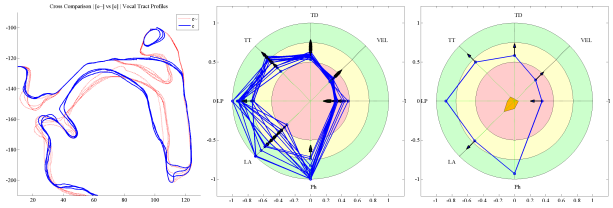

Figure 4: EP vowels [ẽ] and [e] compared by superimposing the different vocal tract profiles, the individual difference polygons and the corresponding comparison diagram depicting difference per region and movement direction.

3. Application Examples

To illustrate the use of the proposed methods we perform a characterization of the articulatory differences among EP vowels revisiting some of our previous work [3]. The main difference is that the characterization is performed relying solely on the computed quantitative difference data, rather than on the observation of the superimposed contours, in order to assess their performance.

3.1. EP Vowels Comparison

A first analysis was performed by comparing the static configuration of EP vowels by considering the annotated interval and selecting the centre frame, for oral vowels, and the frame presenting a lowered velum and the lips still open (typically the last of the interval), for nasal vowels [3].

Figure 5 shows the difference diagrams for all the comparisons. Each diagram presents a difference polygon for each compared pair. Observation of these diagrams allows the fol-

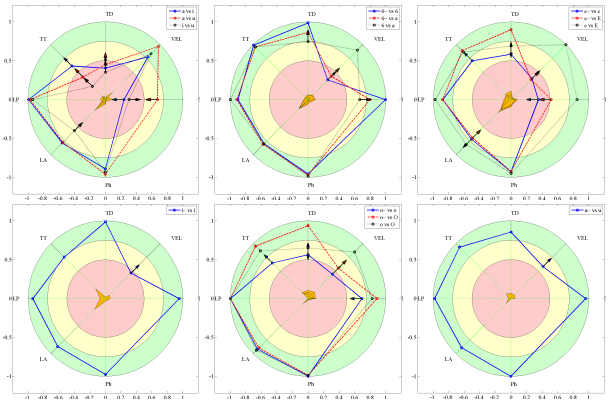

Figure 5: Comparison among nasal vowels and their oral congeners: top, [a][i][u], [ɐ̃][ɐ][a], [ẽ][e][ɛ]; bottom, [ĩ][i], [õ][o][ɔ] and [ũ][u].

lowing conclusions:

[a][i][u] – as expected, since these are cardinal vowels, differences among them are strong. In addition to differences in tongue height (TD) and backness (TB), it is also noteworthy from the diagram a variation in lip aperture between [i] and [u].

[ɐ̃][ɐ][a] – apart from the lowering of the velum, these vowels exhibit very few differences between them. Tongue and labial configuration of the nasal vowel is closer to [ɐ] than [a].

[ẽ][e][ɛ] – oral vowels were produced by the speaker with a significantly more fronted tongue body than the nasal [ẽ]. There are also differences in tongue height between [ẽ] and [e].

[ĩ][i] – the configuration of this oral-nasal pair is very similar, except for the lowering of the velum.

[õ][o][ɔ] – the tongue configuration of [ɔ] and [õ] is quite similar and the main difference between this vowel pair lies in velum height. In comparison with the nasal vowel, the tongue for [o] is somewhat more fronted and lowered.

[ũ][u] – in comparison with oral congener, TD is slightly more raised for [ũ] and the velum is in a lower position, as expected.

3.2. Dynamic Analysis

To perform dynamic analysis we considered all the image frames corresponding to each sound, based on the audio annotation, and compared them with a reference configuration. The dynamic analysis is supported by a difference diagram between the reference and the first frame (beginning of production), a line graph depicting the evolution of the different parameters along the vowel and a difference diagram between the reference and last frames (end of production).

The first row of figure 6 shows the dynamic analysis for the nasal vowel [ɐ̃], using the first frame as a reference. This provides data regarding the changes occurring, for the different articulators, along the duration of the vowel. The major adjustments occur at the velum, which opens slowly along the vowel. This suggests that nasality in Portuguese is typically incremental over the vowel, with a movement from oral to nasal, as shown previously by other studies [3, 22, 23].

Comparison might also be performed using a different frame (e.g., vocal tract profile for another sound) as reference. The bottom row of figure 6 shows the dynamic analysis for nasal vowel [ẽ] using [e] as reference. Notice that [ẽ] starts with a configuration similar to an [e] and differences arise along the production. In addition to the adjustments in tongue height and

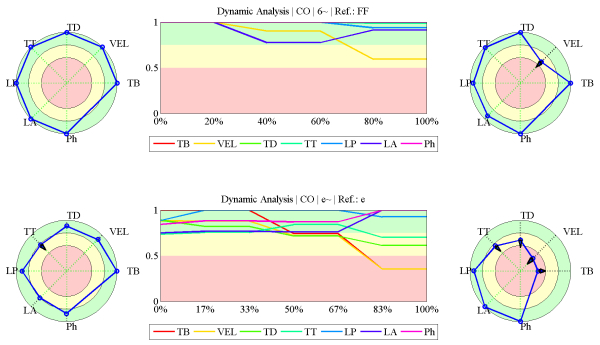

Figure 6: Dynamic analysis of: vowel [ɐ̃] using first frame as reference (top); and vowel [ẽ] using [e] as reference (bottom).

backness, the velum is almost closed at the beginning of both oral and nasal vowels. This will open along the nasal vowel production, which points to the existence of different phases in the production of nasal vowels: an oral onset followed by a nasal portion at the end of the vowel [3, 24, 25].

4. Conclusions

A set of methods is proposed that uses objective measures to compare the configurations assumed by the vocal tract during the production of different sounds considering different meaningful regions under the influence of various articulators. Visual representation of such data is also proposed depicting the differences found and corresponding direction of change.

Using the proposed method to perform an analysis of the articulatory differences between EP oral and nasal vowels lead to the same general conclusions as our previous analysis based on visual inspection of contour superpositions [3], a good indication of its adequacy: nasal vowels [ẽ] and [õ] exhibited more articulatory adjustments with respect to oral congeners than [ĩ] and [ũ]. Although, data from this speaker did not show evidence of anteriority of [ɐ̃] with respect to [a]. Analysis including difference diagrams for the remaining speakers will help to clarify this aspect.

The proposed methods can be further developed in different aspects. Comparison of regions using the Pratt index is performed by determining the corresponding point by a proximity criteria. This might be improved by using a criteria that searches for the corresponding points along the expected principal direction of movement, e.g. radial for tongue dorsum.

Diagram calibration can be improved to better adapt to the conditions relevant to each comparison performed (e.g., to the vowels involved).

Comparison considering additional vocal tract variables [26, 8] should also be considered, in particular for the dynamic analysis.

The dynamic analysis is still only presented for a particular vowel occurrence. As performed with the static analysis, it would be interesting to use all utterances available in the data set, for a particular nasal, to provide an overall summary of what happens over time.

Acknowledgments.: Research partially funded by FEDER through the Program COMPETE and by National Funds (FCT) in the context of HERON II (PTDC/EEA-PLP/098298/2008) and IEETA Research Unit funding FCOMP-01-0124-FEDER-022682 (FCT-PEst-C/EEI/UI0127/2011). The fourth author acknowledges PhD grant (FCT-SFRH/BD/65183/2009).

5. References

[1] C. Oliveira, P. Martins, and A. Teixeira, "Speech rate effects on European Portuguese nasal vowels," in *Proc. Interspeech 2009*, Brighton, 2009.

[2] P. Martins, I. Carbone, A. Pinto, A. Silva, and A. Teixeira, "European Portuguese MRI based speech production studies," *Speech Comm.*, vol. 50, no. 11-12, pp. 925–952, 2008.

[3] C. Oliveira, P. Martins, S. Silva, and A. Teixeira, "An MRI study of the oral articulation of European Portuguese nasal vowels," in *Proc. Interspeech 2012*, 2012.

[4] P. Martins, C. Oliveira, S. Silva, and A. Teixeira, "Velar movement in European Portuguese nasal vowels," in *Proc. IberSpeech 2012 — VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop*, 2012, pp. 231–240.

[5] C. Hagedorn, M. I. Proctor, and L. Goldstein, "Automatic analysis of singleton and geminate consonant articulation using Real-Time Magnetic Resonance Imaging," in *Proc. Interspeech 2011*, 2011, pp. 409–412.

[6] O. Engwall, "A revisit to the application of MRI to the analysis of speech production - testing our assumptions," in *Proc. 6th Int. Seminar on Speech Production (ISSP)*, 2003.

[7] M. K. Tiede and E. Vatikiotis-Bateson, "Contrasts in speech articulation observed in sitting and supine conditions," in *Proc. 5th Seminar on Speech Production*, 2000, pp. 25–28.

[8] E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway Real-Time Magnetic Resonance images," *IEEE Transactions on Medical Imaging*, vol. 28, no. 3, pp. 323–338, 2009.

[9] V. Delvaux, T. Metens, and A. Soquet, "Propriétés acoustiques e articulatoires des voyelles nasales du Français," in *Proc. Journées d'Étude sur la Parole*, Nancy, juin 2002.

[10] C. Shadle, Proctor, M.I., and K. Iskarous, "An MRi study of the effect of vowel context on English fricatives," in *Proc. Acoustics '08 Paris: Joint meeting of the ASA, EAA & Société Française d'Acoustique*, 2008.

[11] A. Niebergall, S. Zhang, E. Kunay, G. Keydana, M. Job, M. Uecker, and J. Frahm, "Real-time MRI of speaking at a resolution of 33 ms: Undersampled radial FLASH with nonlinear inverse reconstruction," *Magnetic Resonance in Medicine*, vol. 69, no. 2, pp. 477–485, 2013.

[12] A. C. Lammert, M. I. Proctor, and S. S. Narayanan, "Data-driven analysis of realtime vocal tract MRI using correlated image regions," in *Proc. Interspeech 2010*, 2010, pp. 1572–1575.

[13] M. I. Proctor, A. C. Lammert, A. Katsamanis, L. M. Goldstein, C. Hagedorn, and S. S. Narayanan, "Direct estimation of articulatory kinematics from Real-Time Magnetic Resonance image sequences," in *Proc. Interspeech 2011*, 2011, pp. 281–284.

[14] R. Shosted, B. P. Sutton, and A. Benmamoun, "Using magnetic resonance to image the pharynx during Arabic speech: Static and dynamic aspects," in *Proc. Interspeech 2012*, 2012.

[15] A. Teixeira, P. Martins, A. Silva, and C. Oliveira, "An MRI study of consonantal coarticulation resistance in Portuguese," in *Proc. 9th Int. Seminar on Speech Production (ISSP)*, 2011.

[16] M. Proctor, D. Bone, A. Katsamanis, and S. Narayanan, "Rapid semi-automatic segmentation of real-time Magnetic Resonance images for parametric vocal tract analysis," in *Proc. Interspeech 2010*, Japan, 2010.

[17] V. Ramanarayanan, D. Byrd, L. Goldstein, and S. S. Narayanan, "Investigating articulatory setting - pauses, ready position, and rest - using real-time MRI," in *Proc. Interspeech 2010*, 2010, pp. 1994–1997.

[18] A. Teixeira, P. Martins, C. Oliveira, C. Ferreira, A. Silva, and R. Shosted, "Real-time MRI for Portuguese: database, methods and applications," in *Proc PROPOR 2012, LNCS vol. 7243*, 2012, pp. 306–317.

[19] P. Boersma and D. Weenink. (2013, March) Praat: doing phonetics by computer [computer program]. version 5.3.42. [Online]. Available: http://www.praat.org/

[20] S. Silva, A. Teixeira, C. Oliveira, and Martins, "Segmentation and analysis of vocal tract from midsagittal real-time MRI," in *Proc. ICIAR 2013, LNCS vol. 7950*, 2013, pp. 459–466.

[21] W. K. Pratt, *Digital Image Processing*. Wiley-Interscience, 2007.

[22] A. Lacerda and B. Head, "Análise de sons nasais e sons nasalizados do português," *Revista do Laboratório de Fonética Experimental de Coimbra*, vol. 6, pp. 5–70, 1966.

[23] R. Sampson, *Nasal Vowel Evolution in Romance*. Oxford University Press, 1999.

[24] F. N. Gregio, "Configuração do trato vocal supraglótico na produção das vogais do português brasileiro: dados de imagens de ressonância magnética," Tese de Mestrado, PUC/SP, 2006.

[25] L. Lovatto, A. Amelot, L. Crevier-Buchman, P. Basset, and J. Vaissière, "A fiberscopic analysis of nasal vowels in Brazilian Portuguese," in *Proc. 16th International Congress of Phonetic Sciences (ICPhS 2007)*, Saarbrücken, 2007, pp. 549–552.

[26] E. Saltzman and K. Munhall, "A dynamical approach to gestural paterning in speech production," *Ecological Psychology*, vol. 1, no. 4, pp. 333–382, 1989.