IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Exploring Advances in Real-time MRI for studies of European Portuguese

Conceição Cunha[1], *Samuel Silva*[2], *António Teixeira*[2], *Catarina Oliveira*[3], *Paula Martins*[4], *Arun Joseph*[5], *Jens Frahm*[5]

[1]IPS, Ludwig-Maximilians-Universität, Germany
[2]DETI/IEETA, Universidade de Aveiro, Portugal
[3]ESSUA/IEETA, Universidade de Aveiro, Portugal
[4]ESSUA/IBIMED/IEETA, Universidade de Aveiro, Portugal
[5]Max Plank Institute for Biophysical Chemistry, Germany

cunha@phonetik.uni-muenchen.de, sss@ua.pt, ajst@ua.pt, coliveira@ua.pt, pmartins@ua.pt, arun-antony.joseph@mpibpc.mpg.de, jfrahm@gwdg.de

Abstract

Recent advances in real-time magnetic resonance imaging (RT-MRI) for speech studies, providing a considerable increase in time resolution, potentially improve our ability to study the dynamic aspects of speech production. To take advantage of the sheer amount of the resulting data, automated methods can be used to select, process, and analyze the data, and previous work could tackle these challenges for an European Portuguese (EP) corpus acquired at 14 frames per second (fps).

Aiming to further explore RT-MRI in the study of the dynamic characteristics of EP sounds, e.g., nasal vowels and diphthongs, we present a novel 50 fps RT-MRI corpus and assess the applicability, in this new context, of our previous proposals for processing and analyzing these data to extract relevant articulatory information. Importantly, at this stage, we were interested in assessing if and to what extent the new data and the proposed methods are able to support and corroborate the articulatory analysis obtained from the previous corpus. Overall, and although this new corpus poses novel challenges, it was possible to process and analyze the 50 fps data. A comparison of automated analysis performed for the same sounds, for both corpora (i.e., 14 fps and 50 fps), yields similar results, corroborating previous results and demonstrating the envisaged replicability. Moreover, we updated the processing in order to be able to analyze dynamic information and provide first insights on the temporal organization of complex sounds such as nasal vowels and diphthongs.

Index Terms: speech production, dynamic, real-time magnetic resonance, European Portuguese, processing and analysis.

1. Introduction

A major challenge in modern linguistics is to understand how continuous and dynamic speech movements are related to perceptual categories like consonants and vowels. Physiological studies have shown that phonological segments can be defined based on the synchronization of primary articulators in speech (the lips, tongue, soft-palate, jaw, and vocal folds) with each other in time [1] using methods such as Eletromagnetic Articulography (EMA) and Magnetic Resonance Imaging (MRI).

Many advances in real-time magnetic resonance imaging (RT-MRI) resolutions (spatial and temporal) have been driven by the need to investigate phonetic and phonological phenomena [2], such as vowel nasalization in Portuguese [3].

In the beginning of RT-MRI application to the study of speech production the temporal resolution was quite low, but this application was already successful in providing for the first time information about the entire tongue contour instead of some points, as in EMA and it updated significantly the information about the velum. The increase of temporal resolution, updates the quality of the analysis and allows the description of a wider set of sounds, including faster sounds as thrills and more complex sounds as diphthongs and nasal vowels. The significant increase of the number of participants is crucial to take apart individual from languages specific characteristics. The state-of-the-art as summarized in [2] includes currently (1) frame rates that already surpass 100 Hz; (2) databases for several languages were recorded at high frame rates (English [4], German [5], French [5]); (3) use of a wide variety of RT-MRI analysis techniques, that can be classified in four classes [2]: basis decomposition or matrix factorization techniques at the level of the raw or processed images, pixel or region-of-interest (ROI) based, grid-based, and contour-based; (4) initial studies regarding the dynamics of articulators and gestural timing relationships.

Despite all these evolutions, particularly in extraction of information from the images (e.g., [5, 6]), the number of studies going beyond the extraction of contour or articulators is very scarce. A representative example of working in the facilitation of high level analysis is [7]. As frame rates rise, more work is needed in these frameworks for quantitative systematic analysis, that will be essential to make possible exploring the highly increased amount of images.

Several studies used MRI for studying EP, as summarized in Section 2. Compared to the state-of-the-art [2], these studies used a low frame rate, possibly providing not enough information to the adequate characterization of the investigated sounds. Further limitations are: (1) the scarce amount of publications and missing analysis of some sounds, namely rhotics and diphthongs; (2) preliminary character of the approaches to other sounds such as nasal vowels, due to the limited frame rate used and the amount of participants.

The main objectives of this paper are the following: (1) adding Portuguese to the small set of languages studied using the last evolutions in RT-MRI; (2) contribute to a better understanding of the dynamical aspects in the production of speech sounds; (3) assess previous results obtained with lower temporal resolution; (4) profit from the improved temporal resolution to start covering sounds characterized by their dynamic nature which could not be contemplated in previous studies, such as the diphthongs.

Even if essential for these objectives, the acquisition of

10.21437/IberSPEECH.2018-29

the novel data for new classes of sounds and more complex phonetic contexts with higher temporal resolution poses several challenges on how to process and analyze the resulting database. Our team has previously addressed this challenge, for different data [8, 9] by proposing methods to process [6] and analyze [7, 10] the image sequences to extract articulatory data. However, the nature of the novel database raises several new questions, opportunities and challenges, including the applicability of previous analyses.

The remainder of this article is organized as follows: Section 2 provides a short summary of related work in speech production studies using RT-MRI and other techniques, focusing in studies addressing EP; Section 3 presents information regarding a novel RT-MRI database acquired for EP, from corpus description to the methods considered for acquisition, processing and analysis of the image sequences; in Section 4, illustrative results are presented, including novel information regarding diphthongs production; finally, the article concludes with a brief summary of the contributions and ideas for future work.

2. Related Work

Nowadays, speech production studies can be supported by a wide range of technologies including imaging modalities (e.g., Ultrasound, MRI) and other instrumental techniques (e.g., EMA [11, 12]). In this context, real-time magnetic resonance imaging has been received particular relevance due to its non-invasiveness, non-use of ionizing radiation, and the remarkable improvements in spatio-temporal resolution achieved in the last few years [13, 2].

From the first attempts to acquire real-time imaging with MRI to date, much has been evolved in the area, achieving sampling rates close to those obtained with EMA and Ultrasonography. This has been possible by exploiting several technological advancements that involve the use of high field strengths, more powerful gradients, dedicated coils, non-Cartesian K-space trajectories, high degree of undersampled data and more efficient image reconstruction algorithms allowing for high sampling rates and improved image quality [5, 14].

For European Portuguese (EP), several of these techniques (e.g., EMA and MRI) have been used. These efforts include not only data acquisition but also data analysis. Early studies addressed dynamic aspects of nasal vowel production using EMA [15, 16, 17] and MRI (e.g. [9]

Internationally, several research groups have been using MRI to gather information for different languages using different approaches. Comprehensive reviews of these studies are summarized in [13, 2, 7].

The first MRI study for EP included 2D and 3D data regarding static configuration of all EP vowels and consonants for one speaker [9]. A deeper study on EP laterals (3D) was conducted later with data from 7 participants [8, 18]. A study of co-articulation resistance in EP was presented in [19]. First results of RT-MRI for EP were presented in 2012 [3]. At this stage, a RT-MRI dataset which included nasal vowels, laterals, taps and trills was acquired with a frame rate of 14 fps. This represented an important first step towards a better characterization of the dynamic aspects involved in the production of these sounds of Portuguese [3]. The configuration of the vocal tract during the production of nasal vs. oral vowels was investigated using RT-MRI in [20]. More recently, RT-MRI was used for studying the temporal coordination of oral articulators and the velum during the production of nasal vowels [21], providing additional support for the delayed coordination of oral and nasal gestures in Portuguese [22, 21].

Despite all these relevant contributions, the sounds of Portuguese are not yet well described: diphthongs and trills still missing, nasal vowels need an improved characterization due to the lower frame rate and the small amount of participants.

From the perspective of processing and quantitative information extraction a lot has been achieved recently for EP exploring, essentially, automatic techniques to deal with the quantity and diversity of the information obtained, both from 3D static [18] and real-time MRI [20, 6, 7, 10], However, the problem of information extraction for speech production studies and development of relevant articulation models is far from solved. The consideration of these imaging technologies poses challenges to extract articulatory-relevant information profiting from the full range of available data.

3. Methods

Gathering novel insights on the dynamic nature of complex sounds, for EP, taking advantage of recent advances in real-time MRI, involves several steps from corpus definition to articulatory analysis, as described in what follows.

3.1. Corpus

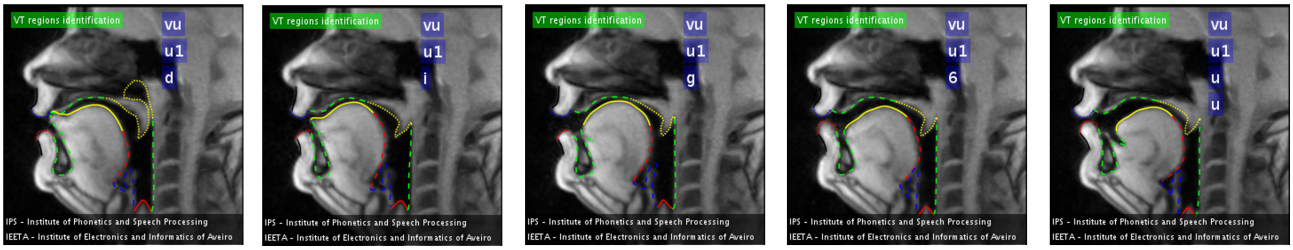
The corpus consists of minimal pairs containing all stressed oral [i, e, ɛ, a, ɔ, o, u] and nasal vowels [ɐ̃, ẽ, ĩ, õ, ũ] in one and two syllable words. Nasal diphthongs /ɐ̃w, ẽj, ɐ̃j/ and the oral counterparts /aw, aj, oj/ as well as /ej, ew, iw, ow, uj/ in monosyllabic words were also included. Additional materials were recorded for further modeling of variability in the production of nasality.

All words were randomized and repeated in two prosodic conditions embedded in one of three carrier sentences alternating the verb as follows (Diga 'Say'—ouvi 'I heard'— leio 'i read') as in 'Diga **pote**, diga **pote** baixinho' ('Say **pot**, Say **pot** gently'). So far, this corpus has been recorded from twelve native speakers (8m, 4f) of EP. The tokens were presented from a timed slide presentation with blocks of 13 stimuli each. The single stimulus could be seen for 3 seconds and there was a pause of about 60 seconds after each block of 13. The first three participants read 7 blocks in a total of 91 stimuli and the remaining nine participants had 9 blocks of 13 stimuli (total of 117 tokens).

3.2. RT-MRI Acquisition

RT-MRI recordings, as exemplified in Fig. 1, were conducted at the Max Planck Institute for biophysical Chemistry, Göttingen, Germany, using a 3 Tesla Siemens Magnetom Prisma Fit MRI System equipped with high performance gradients (Max ampl=80 mT/m; slew rate = 200 T/m/s).

A standard 64–channel head coil was used with a mirror mounted on top of the coil. The speaker was lying down, in a comfortable position, and was instructed to read the required sentences. Real-time MRI measurements were based on a recently developed method, where highly under-sampled radial FLASH acquisitions are combined with nonlinear inverse reconstruction (NLINV) providing images at high spatial and temporal resolutions [23]. Acquisitions were made at 50 fps, resulting in images as the ones presented in Fig. 1. Speech was synchronously recorded using an optical microphone (Dual Channel-FOMRI, Optoacoustics, Or Yehuda, Israel), fixed on the head coil, with the protective pop-screen placed directly against the speaker's mouth.



d i g ɐ u

Figure 1: *Selected frames for multiple sounds occurring as in "Diga vu" (Say 'vu').*

Before their enrollment on the study, all volunteers provided informed written consent and filled an MRI screening form in agreement with institutional rules. None of the participants had any known language, speech or hearing problems and were compensated for their participation.

3.3. Speech annotation

The speech recordings were preprocessed to filter MRI acquisition noise and the target segments manually delimited by the first author using Praat [24, 25]. The produced annotation information was used in Matlab for extraction of relevant images and synchronized speech signal.

3.4. Processing and Analysis

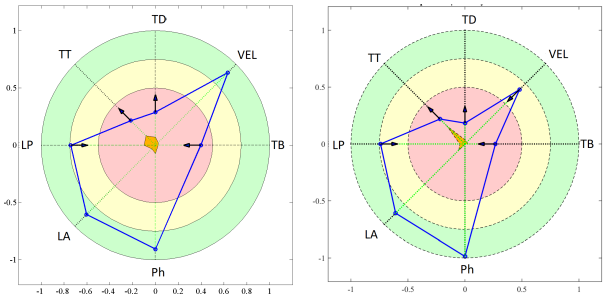
In line with the semi-automated methods described in Silva et al. [7], twenty eight images were manually annotated to train two active appearance models [26] for oral and nasal configurations of the vocal tract. For the work presented in this article, the RT-MRI sequences for one of the speakers (male, 39 yo) were processed to extract sequences of vocal tract contours. Figure 1 illustrates the overall outcome of the processing stage by presenting a few selected image frames from the sentence 'Diga vu' (say vu) depicting the identified vocal tract contours, with the different articulators shown in different colors and line types, and the corresponding audio annotation.

The static and dynamic analysis and comparison among vocal tract configurations was performed by adopting the framework proposed in Silva et al. [10] providing objective normalized analysis and visualization. Accordingly, vocal tract configurations are compared for seven different regions: velum (VEL), tongue dorsum (TD), tongue back (TB), tongue tip (TT), lip protrusion (LP), lip aperture (LA) and pharynx (Ph). For each, the comparison yields a score, from 1 (no difference) to 0 (strong difference), which is represented over the unitary circle (Fig. 2) or graph (Fig. 3) for static and dynamic analysis, respectively. For the sake of brevity, the reader is forwarded to [10] for additional details regarding the adopted analysis framework.

4. Results

Considering our initial goals, to start exploring the new corpus and assess the applicability of the previously proposed processing and analysis methods to this new context, we processed the data for one of the male speakers. Our aim was to explore two important aspects: confirm overall previous findings, as a proof of replicability, particularly for oral and nasal vowels, and obtain a first insight over EP diphthongs.

As an example, Fig. 2 shows the comparison between [a] and [i] performed for the previous 14 fps corpus and using data



14 fps corpus 50 fps corpus

Figure 2: *Results obtained for the comparison among vowels [a] and [i] for two different RT-MRI corpora: a previous 14 fps acquisition and the current 50 fps corpus.*

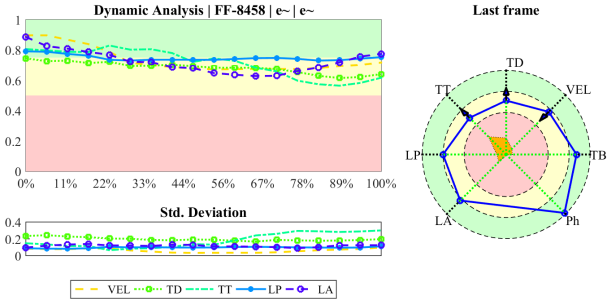


Figure 3: *Changes to the different articulators along the production of [ẽ].*

from the new 50 fps corpus. The visual representation, in Fig. 2, should be interpreted as the answer to the question: "What happens when I move from [a] to [i]?". As can be observed, the difference polygon, inscribed in the unitary circle, has grossly the same shape, for both corpora, and depicts the same notable differences, namely (dots in yellow and red circular coronas): the tongue dorsum moving up, and an advance of the tongue tip and tongue back. Overall, the similarity between the diagrams (results) obtained for both corpora corroborates the reproducibility of the analysis methods among datasets.

For the dynamic analysis of nasal vowels, Fig. 3 shows, as an example, the evolution of the vocal tract configuration along the production of [ẽ]: it is essentially characterized by velum movement (noticeable by a trajectory entering in the yellow zone of the graph at around 30%). Of note is, also, the variation for the tongue dorsum and tongue tip, observed for the

end of the vowel and a possible consequence of coarticulation effects. For instance, the influence of the carrier sentence's second 'diga' (say), following the token, as in 'Diga **vem, diga** . . . ' (Say come, say . . .).

4.1. EP Diphthongs

Given the exploratory nature of this first effort regarding diphthongs, to have a first grasp of what is happening, we started by oral diphthongs. As an example, Fig. 4 shows results for [aw]

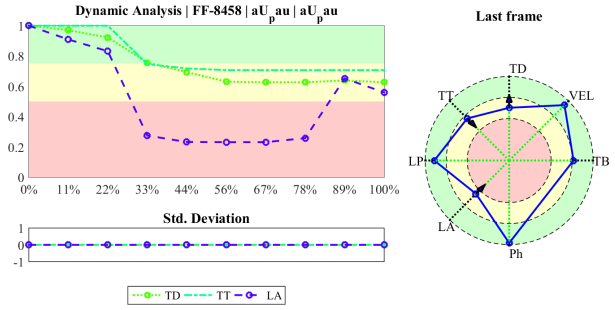


Figure 4: *Variation, over time, of the articulators during the production of the EP oral diphthong [aw], as in 'pau' (stick).*

as in 'pau' (stick). The diphthong is covered by 10 images (10 points in the graph), including the initial [p] and the diphthong (around 200 ms). Note that these representations only present lines, in the graph, for those regions where, along production, changes fall into the yellow and red stripes. It is clear an abrupt change in lip aperture (LA) at 20-30 % of diphthong duration, from [p] to [a], and a reduction, close to the end, for the [w].

The nasal counterpart of this diphthong, [ɐ̃w̃], as in 'pão' (bread), is analyzed in Fig. 5, showing a gradual variation of lip aperture (LA), similar to the one observed in paw. Additionally, there is movement of the tongue back, as opposed to the oral counterpart, probably due to the need of adjustment in the nasal passage, which is hinted by the changes also noted at the velum.

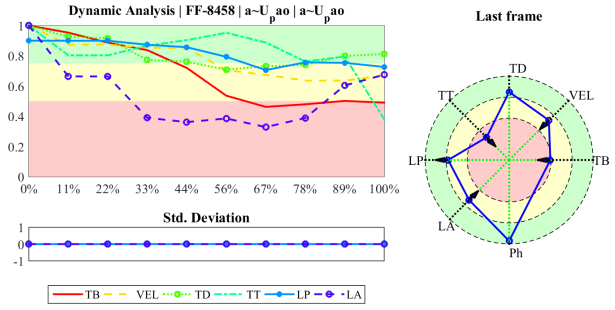


Figure 5: *Variation, over time, of the articulators during the production of the EP oral diphthong [ɐ̃w̃], as in 'pão' (bread).*

Our corpus and methods also enable investigating the complex context of a diphthong after a nasal consonant, and, for instance, in comparison with other diphthongs. In Fig. 6, we compare the production of 'mão' ([mɐ̃w̃], 'hand') with 'pão' ([pɐ̃w̃], 'bread'), showing that, as expected, the velum behaves

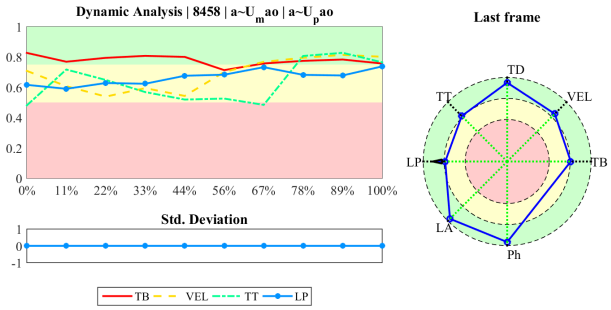


Figure 6: *Comparison, over time, of 'mão' ([mɐ̃w̃], 'hand') with 'pão' ([pɐ̃w̃], 'bread').*

differently, in the initial part, since it is open, from the beginning, in 'mão'.

5. Conclusion

The major contribution of this paper is the presentation of a novel RT-MRI database for EP recorded with a frame rate of 50 Hz, contributing to augment the very reduced amount of languages with such a valuable resource for speech production studies. This database, after its completion and pre-processing will be partially made available for other researchers. Additionally, we demonstrate the applicability of previously proposed methods for segmentation and analysis, by illustrating previous findings for oral and nasal vowels and performing a first exploration of EP diphthongs. The application of this methodology to more data and speakers will enable a detailed description of nasal sounds in European Portuguese and a better understanding of their implementation in production.

The work presented here can still profit from several improvements and provides the grounds for exploring new routes of speech production studies in EP. Even though the image quality is better than our previous 14 fps corpus, the different nature of the corpus, with a large number of dental, alveolar and palatal contacts (in the support words and sentences, e.g., [t] before [ɐ̃w̃] as in 'sotão' – attic) poses new challenges to vocal tract segmentation, with a few segmentations still requiring a final manual revision, an aspect to improve as new speakers are included, by training better models [6].

Finally, now that the grounds for work have been established, future developments should be propelled by addressing concrete hypotheses regarding EP nasals, such as the one of delayed coordination of oral and nasal gestures in Portuguese [27, 21, 22].

6. Acknowledgements

This work is partially funded by the project 'Synchrone Variabilität und Lautwandel im Europäischen Portugiesisch', with funds from the German Federal Ministry of Education and Research, by IEETA Research Unit funding (UID/CEC-/00127/2013), by Portugal 2020 under the Competitiveness and Internationalization Operational Program, and the European Regional Development Fund through project SOCA – Smart Open Campus (CENTRO-01-0145-FEDER-000010) and project MEMNON (POCI-01-0145-FEDER-028976). We thank Philip Hoole for the scripts for noise supression and all the participants of the experiment for their time and voice.

7. References

[1] L. Goldstein and M. Pouplier, "The temporal organization of speech," *The Oxford handbook of language production*, p. 210, 2014.

[2] V. Ramanarayanan, S. Tilsen, M. Proctor, J. Töger, L. Goldstein, K. S. Nayak, and S. Narayanan, "Analysis of speech production real-time MRI," *Computer Speech & Language*, 2018.

[3] A. Teixeira, P. Martins, C. Oliveira, C. Ferreira, A. Silva, and R. Shosted, "Real-time MRI for portuguese," in *International Conference on Computational Processing of the Portuguese Language*. Springer, 2012, pp. 306–317.

[4] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein *et al.*, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc)," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1307–1311, 2014.

[5] M. Labrunie, P. Badin, D. Voit, A. A. Joseph, J. Frahm, L. Lamalle, C. Vilain, and L.-J. Boë, "Automatic segmentation of speech articulators from real-time midsagittal MRI based on supervised learning," *Speech Communication*, vol. 99, pp. 27 – 46, 2018.

[6] S. Silva and A. Teixeira, "Unsupervised segmentation of the vocal tract from real-time MRI sequences," *Computer Speech and Language*, vol. 33, no. 1, pp. 25–46, Sep. 2015.

[7] ——, "Quantitative systematic analysis of vocal tract data," *Computer Speech & Language*, vol. 36, pp. 307 – 329, 2016.

[8] A. Teixeira, P. Martins, C. Oliveira, and A. Silva, "Production and modeling of the european portuguese palatal lateral." in *Computational Processing of the Portuguese Language, PROPOR 2012, Lecture Notes in Computer Science/LNAI, Vol. 7243*, 2012.

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

[10] S. Silva and A. J. S. Teixeira, "Critical articulators identification from RT-MRI of the vocal tract," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 626–630.

[11] J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta, and M. T. Jackson, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," *The Journal of the Acoustical Society of America*, vol. 92, no. 6, pp. 3078–3096, 1992.

[12] P. Hoole and N. Nguyen, "Electromagnetic articulography," *Coarticulation–Theory, Data and Techniques, Cambridge Studies in Speech Science and Communication*, pp. 260–269, 1999.

[13] A. D. Scott, M. Wylezinska, M. J. Birch, and M. E. Miquel, "Speech MRI: Morphology and function," *Physica Medica*, vol. 30, no. 6, pp. 604 – 618, 2014.

[14] J. Frahm, S. Schätz, M. Untenberger, S. Zhang, D. Voit, K. D. Merboldt, J. M. Sohns, J. Lotz, and M. Uecker, "On the temporal fidelity of nonlinear inverse reconstructions for real-time MRI–the motion challenge." *The Open Medical Imaging Journal*, vol. 8, pp. 1–7, 2014.

[15] A. Teixeira and F. Vaz, "European Portuguese Nasal Vowels: An EMMA study," in *7th European Conference on Speech Communication and Technology, EuroSpeech - Scandinavia*, vol. 2. Aalborg, Dinamarca: CPK/ISCA, Sep. 2001, pp. 1843–1846.

[16] C. Oliveira and A. Teixeira, "On gestures timing in european portuguese nasals," in *ICPhS*, 2007, pp. p. 405 – 408.

[17] S. Rossato, A. Teixeira, and L. Ferreira, "Les nasales du Portugais et du Français : une étude comparative sur les données EMMA," in *JEP'2006*, Rennes, França, 2006.

[18] P. Martins, C. Oliveira, C. Ferreira, A. Silva, and A. Teixeira, "3D MRI and semi-automatic segmentation techniques applied to the study of european portuguese lateral sound," in *International Seminar on Speech Production (ISSP'11)*, Montreal, Jun. 2011.

[19] A. Teixeira, P. Martins, A. Silva, and C. Oliveira., "An MRI study of consonantal coarticulation resistance in portuguese," in *International Seminar on Speech Production (ISSP'11)*, Montrreal, Canada, Jun. 2011.

[20] S. Silva, A. Teixeira, C. Oliveira, and P. Martins, "Segmentation and analysis of vocal tract from midsagittal real-time mri," in *International Conference Image Analysis and Recognition*. Springer, 2013, pp. 459–466.

[21] A. R. Meireles, L. Goldstein, R. Blaylock, and S. Narayanan, "Gestural coordination of brazilian portugese nasal vowels in CV syllables: A real-time MRI study." in *ICPhS*, 2015.

[22] C. Oliveira, "Do grafema ao gesto. contributos linguísticos para um sistema de síntese de base articulatória," Ph.D. dissertation, Universidade de Aveiro, 2009.

[23] M. Uecker, S. Zhang, D. Voit, A. Karaus, K.-D. Merboldt, and J. Frahm, "Real-time mri at a resolution of 20 ms," *NMR in Biomedicine*, vol. 23, no. 8, pp. 986–994, 2010.

[24] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–347, 2001.

[25] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program], version 6.0.40," 2018, retrieved 11 May 2018 from http://www.praat.org/.

[26] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[27] C. Cunha, S. Silva, A. Teixeira, C. Oliveira, P. Martins, , A. Joseph, and J. Frahm, "Analysis of nasal vowels and diphthongs in european portuguese," in *Workshop New Developments in Speech Sensing and Imaging (Labphon satellite event)*, Lisbon, Jun. 2018.