IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Data-driven analysis of nasal vowels dynamics and coordination in bilabial contexts

Conceição Cunha[*2], *Nuno Almeida*[*1], *Jens Frahm*[3], *Samuel Silva*[1], *António Teixeira*[1]

[*]These authors contributed equally to this work
[1]IEETA, DETI, University of Aveiro, Aveiro, Portugal
[2]Institute of Phonetics and Speech Processing, LMU Munich, Germany
[3]Max Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany
cunha@phonetik.uni-muenchen.de, nunoalmeida@ua.pt

Abstract

One of Portuguese distinctive marks is the large nasals inventory, including five phonemic nasal vowels and three diphthongs. Previous studies argued for an initial oral part and a short nasal consonant, probably related to synchronization between oral and nasal gestures. These studies have considered discrete descriptions with EMA-flesh points, limiting our grasp of the whole vocal tract, and preliminary work using real-time MRI (RT-MRI) considered a small frame rate (14fps) and a reduced number of speakers, yielding a rather small time-resolution to study an intrinsically dynamic process. The recent advances of RT-MRI, with frame rates of 50fps, have made possible a finer detail of the dynamics of nasals. However, new challenges need to be tackled to deal with the resulting large amount of data and to foster analyses to tackle a larger number of speakers. Grounded on a new RT-MRI corpus for European Portuguese, this paper explores the capabilities of recent data-driven methods, to analyze dynamic aspects of nasal vowels and coordination. To this end, we consider data for 11 EP speakers and investigate vocal tract configurations, over time, and the coordination of velum and lip aperture in bilabial (oral and nasal) contexts. Overall, the results show changes of the vocal track and the model explores the dynamic behavior of the vowel tract along the production of oral and nasal vowels.

Index Terms: Speech production studies, European Portuguese, RT-MRI, nasals, data-driven analysis

1. Introduction

European Portuguese (EP) distinguishes five nasal vowels and three diphthongs. Nasal vowels are complex sounds visible also in the acoustic antiformants caused by the inclusion of the nasal cavity in the production and the difficult parsing of articulation and acoustics. EP nasal vowels are usually divided in an initial oral portion, a nasal part and a consonantal tail, probably related to a late alignment of the velum relatively to the vowel tract configuration for the vowel production. However, there is no robust evidence for this partition and this will be one of the advances of this paper. These dynamic aspects cannot be caught with static analysis of vowel midpoints and do have to consider the whole or a greater part of its duration.

The study of such complex sounds, exhibiting characteristic dynamic patterns contributes to improve our knowledge on speech production supporting a range of applications from speech therapy to articulatory speech synthesis. Initial studies of EP nasal vowels production dynamics resorted to Electromagnetic Articulography (EMA) data, and contemplated: quantitative analysis of velum movement in context of stop consonants [1]; comparison with French nasal vowels [2]; study of gestures timing, characterization of the gesture in terms of average duration, investigation of factors influencing such durations, and characterization of inter-gestural coordination [3]; and speech rate effects [4]. With advances in magnetic resonance imaging (MRI) and real time magnetic resonance imaging (RT-MRI), this studies were complemented by exploring the coverage of the complete vocal tract in, for example [5, 6, 7, 8].

It is important to expand the body of knowledge for EP nasals by moving into data acquisition techniques that provide a more complete view over the vocal tract, when compared to EMA, along with a finer grasp of tract dynamics. Advances in RT-MRI of the vocal tract [9] allow improving on previous research by providing better image quality at a higher frame rate of 50fps, enabling to move beyond past limitations on time resolution. However, with this new data, several challenges arise mostly resulting from the sheer amount of data and from how to systematically process and analyze it to obtain useful results to address the research questions in a quantitative manner [10, 11]. Additionally, the amount of available data also opens new possibilities to move into more data-driven analysis, e.g., adopting methods to model the behavior of relevant sounds by gathering the data from multiple repetitions and/or speakers.

In view of recent methods proposed to process and analyze speech production data extracted from real-time MRI of the vocal tract [12] it is important to understand how these, proposed for a general scope, might be of use to investigate the dynamics of nasals. In this context, the work presented here considers a novel RT-MRI database for European Portuguese and aims to:

- Explore recent methods proposed to support dynamic analysis and modeling from articulatory data extracted from RT-MRI of the vocal tract;
- Assess the applicability of these new methods to the analysis of nasal vowels;
- Provide first reports regarding vowel dynamics and coordination for EP nasal cardinal vowels in bilabial contexts.

The remainder of this document is organized as follows. Section 2 presents an overall overview and background regarding the study of nasal vowels and the consideration of data-driven methods to process and analyze speech production data; section 3 describes the considered data and provides an overview of the methods to tackle it; section 4 presents and discusses the obtained results; and, finally, section 5 presents some conclusions and routes for further work.

10.21437/IberSPEECH.2021-46

2. Background and Related Work

2.1. On nasal vowels

The production of nasal vowels involve much more than lowering the velum. The way this aperture, and other articulators, vary in time and their coordination is important [13]. Nasal vowels can be regarded as diphthongs [14], starting with dominant lips radiation and ending in a nasal radiation dominant configuration. This dynamic nature of nasal vowels is important for their perception [15].

It is often unclear when a vowel is a phonemic nasal or simply contextually nasalized. For example, for Brazilian Portuguese, Meireles [16] found a synchronous coordination of the nasal vowel with the preceding consonant, while Desmeules-Trudel [17] reported a very late alignment of nasal and oral gestures, arguing against the phonemic status of the nasal vowels.

Despite efforts such as [3, 8, 10], it is not yet completely clear how oral and nasal gestures are synchronized in EP nasal vowels production, mainly due to the limitations of the analyzed data (restricted number of speakers and partial information regarding the tract, for EMA, or reduced temporal resolution, for RT-MRI).

2.2. Analysis and Modeling of Articulatory Data

With the increase of the data available from different technologies supporting speech production studies, as is the case for RT-MRI [18, 9], it is paramount to pursue methods that enable its systematic quantitative assessment through unsupervised approaches, to take the most out of the available data. In this regard, several authors have proposed data-driven methods for their processing and analysis (e.g. [19, 20, 21].

In this regard, the authors have explored data-driven approaches to determine critical articulators from vocal tract data extracted from RT-MRI. [22, 11] expanding an approach proposed for EMA [23]. One notable aspect of the presented approach is that, even though the method considers statistical modeling for the different sounds, the consideration of tract variables aligned with the Articulatory Phonology framework [24], as grounds for the method, yields results that keep a connection to the tract anatomy (e.g., constrictions on the tongue tip and body) and are, hence, more interpretable towards a critical discussion of the outcomes and an improved knowledge of speech production. Nevertheless, while this provides valuable information for a wide range of sounds, in an unsupervised manner, in its current state it still only considers a static representation for each phone (i.e., one time point along the production). Even though, for nasal vowels, three time points were selected, to grasp some of the dynamics, the granularity of the resulting data does not enable taking conclusions about the subtlety of the underlying dynamics and coordination.

In a recent article, Carignan et al. [12] explore how vocal tract data extracted from RT-MRI can be explored by modeling the dynamics of speech production based on multiple repetitions of each sound. In their method they adopt generalized additive mixed models (GAMMs) applied to vocal tract aperture functions, along with validations of the resulting models using functional linear mixed models (FLMMs) at 20% and 80% of the vowel interval. Overall, these methods model vocal tract aperture, over time, for given sounds, and can be useful to gain insight over how sounds are produced considering data from multiple speakers, at once. One notable point addressed is the interpretability of the results obtained with the proposed methods. To this end, the authors establish a correspondence between data points on the tract aperture functions and anatomical regions and apply this principle to all speakers and repetitions, which then allows an understanding of the resulting GAMMs.

3. Methods

In what follows, an overall description of the considered data and methods adopted for analysis is provided.

3.1. Data Aquisition

The RT-MRI dataset recordings were performed at the Max-Planck-Institute in Göttingen, Germany, using a 3T Siemens Prisma Fit MRI System equipped with a 64-channel head coil. The MRI acquisitions involved a low-flip angle gradient-echo sequence with radial encodings and a high degree of data under sampling [9]. The procedure allowed for real-time image sequences of the vocal tract in a midsagittal plane of the speaker at 50 fps. Speech sound was synchronously recorded using an optical microphone (Dual Channel-FOMRI, Optoacoustics) and annotated using Praat [25].

3.2. Corpus and Speakers

The analysed corpus consists of minimal pairs containing the three stressed oral and nasal point vowels [i, u, a] and [ĩ, ũ ɐ̃] preceded by bilabial oral or nasal consonants, as in the following words: 'pato' [patu], 'panto' [pɐ̃tu], 'mato' [matu], 'manto' [mɐ̃tu]. 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 pato, diga pato baixinho' ('Say duck, Say duck gently'). The data considered for the analysis presented in this article include 11 native speakers of EP and is part of a larger corpus being acquired to study EP nasals.

3.3. MRI data Processing and Analysis

Overall, our purpose was to explore the applicability of the method proposed by Carignan et al. [12]. In short, the processing pipeline consists of five steps: (1) **Image registration**, in which, images are aligned to compensate some movement of the speaker; a (2) **Semi-polar grid** is placed throughout the vocal tract, it consists of 28 lines distributed from the glottis to the anterior edge of the alveolar ridge; (3) **Air-tissue boundary detection** processes semi-automatically each frame to find the outer boundary of the vocal tract for each grid line; the (4) **Aperture estimation** is obtained by counting the pixels from the boundary that are below a determined threshold; (5) Principal Component Analysis (PCA) of the **velum and lip aperture** are estimated by an approach based on region-of-interests (ROI) and considering pixel intensities (e.g., the amount of darker pixels is higher for the inter-lip region when the lips are open).

Following the same approach as Carignan et al. [12], but since we are interested in lip and velar aperture, we have complemented the semi-polar grid data with the data obtained from the velum and lips to test if Generalized Additive Mixed Models (GAMMs) visualizations could provide any useful insight on their changes, over time. To this end, 1) we have added two more grid lines in the lips area with the computed values; and 2) in the velum area, we have subtracted a factor of the value obtained in the velum PCA from the tract aperture function, in that region, in an attempt that velar opening would induce a discernible change, in the GAMMs, between oral and nasal sounds, our object of study. To get a grasp of what is happening

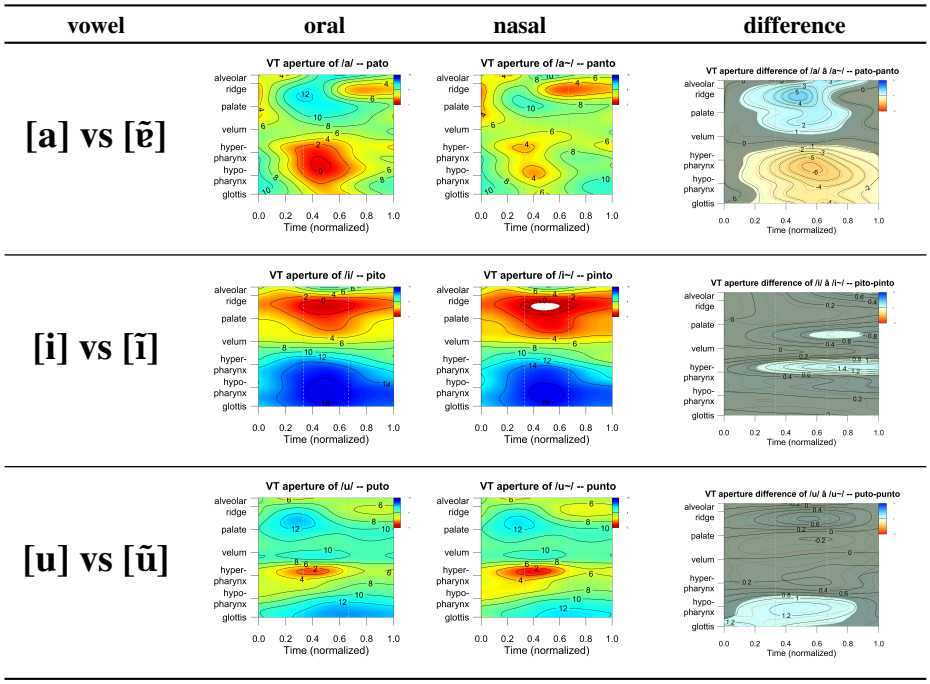

Figure 1: *Results from GAMMs for the three vowel pairs in /p/ context. From left to right, in each row, the plots refer to the oral, the nasal, Each GAMMs plot encompasses the previous consonant /p/, the vowel (at the center, from 0.33 to 0.66), and the next consonant. The results for the /m/ context presented similar patterns and are not shown for the sake of brevity.*

to the vocal tract, we also added the preceding and following consonant, as context. The obtained GAMMs show the vowel (at their center, between 0.33 and 0.66), but also show the preceding consonant (from 0 to 0.33) – a /p/ or a /m/, and the following consonant (from 0.66 to 1) – a /t/ or a /d/. See Figure 1 for examples.

To complement the GAMMs visualizations with more detailed data for analysis, Functional Linear Mixed Models (FLMMs) were computed for tract apertures at 20%, 50% and 80% of the vowel interval. Also, a first attempt is presented of using FLMMs to model the behaviour of the lips and velum, over the time, to support the analysis of their coordination.

4. Results

This section presents examples of analyses aiming at both assessing the capabilities of the methods and contributing to augment knowledge regarding the temporal aspects of nasal vowel production. The results gathered for 11 EP speakers covering an overall analysis of tract dynamics, a more detailed analysis of tract configuration for key timepoints along the vowels, and an overall analysis of lip and velar coordination.

4.1. Overall Vocal Tract Dynamics

The analysis started by an overall assessment of how the tract evolves for oral and nasal vowels by applying GAMMs to the considered bilabial contexts for the 3 vowels. The resulting 'areograms' showing data for the vowel, at the center of each plot (from 0.33 to 0.66), along with the flanking consonants ([p] and [t]) are presented in Figure 1. Each row shows the "areogram" for the oral, its nasal congener, and the difference between the two. For the sake of brevity and given no major differences between both bilabial contexts, only the oral [p] contexts are shown.

Overall, the differences for the pairs [u]/[ũ] and [i]/[ĩ] are very small, for both contexts, noticeable by the dominant grayish color of the difference diagram. Differences between [a] and [ɐ̃] are more noticeable. For both contexts, the oral is more backed than the nasal (the redish area on the difference GAMMs, for the lower tract, and blueish around the alveolar ridge). Finally, the dynamic pattern for each vowel pair is similar across the considered contexts. The GAMMs representations, obtained based on the tract area functions, which we modified by including velar and labial aperture, provide good insight on oral configurations, but nasality differences do not arise, in the representation, as we attempted.

4.2. Detailed Analysis at Specific Timepoints

While the GAMMs in Figure 1 show the overall evolution of tract aperture, to have a more detailed grasp of the tract's configuration, Figure 2 shows superimposed plots of tract apertures at specific times: beginning (20%), middle (50%), and end (80%) of the vowel interval.

For [a] and [ɐ̃] these plots further confirm the previously described results. For both contexts, the oral vowel exhibits a smaller aperture at the back of the tract pointing to a more backed configuration than the nasal. At the alveolar ridge, it is the opposite effect, consistent with the observed backness of the oral and hinting on a wider lip aperture (although the plots do not explicitly include lip data), and also, possibly, of the jaw.

For vowels [i] and [u], the differences towards their nasal congeners are very small (slightly higher for [u]) and no notable difference appears across contexts. For [u], the nasal shows a slight difference for the alveolar ridge, from 50% onward, possibly due to the tongue movement to produce the [n], in the considered contexts.

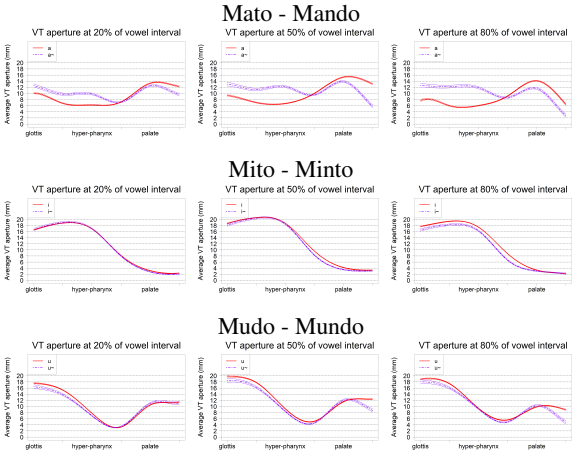

Figure 2: *FLMMs showing the aperture function of the vocal tract for three time points. Each row presents the aperture function for 20%, 50% and 80%, along the vowel. Each of the presented plots shows, in red, the curve for the oral and, in blue, the curve for the nasal. The curves for the oral bilabial context [p] showed similar patterns.*

4.3. Coordination

Since the vowels were produced in bilabial context, we will analyse the coordination of the lower lip with the velum. Figure 3 shows FLMMs for the lip and velar aperture, over time, for the different oral/nasal pairs. By modeling lip and velar behavior considering the data for the 11 speakers, we obtain a first grasp over coordination, an important (and challenging) aspect to study for nasal vowels. For vowel [ɐ̃], in both contexts, lip aperture's peak occurs earlier than for the oral vowel and to a smaller aperture. The smaller aperture is consistent with the more closed vowel quality of the nasal congener. Vowel [i] seems to show a similar (albeit less pronounced) pattern. For the nasals in [p] context, the onset of velar opening seems to happen slightly after lip occlusion release.

Overall, velar behavior for both contexts is as expected. For [p] contexts, the velum starts closed, opens during the nasal vowel, and closes, again. For [m], the velum starts open and gradually closes during the oral vowel, or continues closing after the nasal vowel. Interesting to note is the higher error interval for the velar curve in [ũ]. For ɐ̃, and ĩ, a close observation of the curves seems to reveal a slight tendency of lip closure minima, after the vowel, occurring slightly before velar closure.

5. Conclusions

This paper presents a novel analysis of EP nasal vowels regarding dynamics of the tract configuration and coordination of velum movement with lip aperture based in the application of GAMMs to high frame rate RT-MRI data.

The methods explored in this article enabled an elegant approach to tackling data from multiple speakers to reach an overall model of the dynamic behavior of the tract, along the production of oral and nasal vowels. The results presented, although a first exploration of these methods, already highlight interesting aspects regarding nasal vowels for a considerable number of speakers. These results corroborate the similarity of the vocal tract configuration, over time, for [ĩ] and [ũ], when compared with their oral congeners and a stronger difference in configuration between [a] and [ɐ̃] that seems to be slightly more pronounced at the end of the vowel.

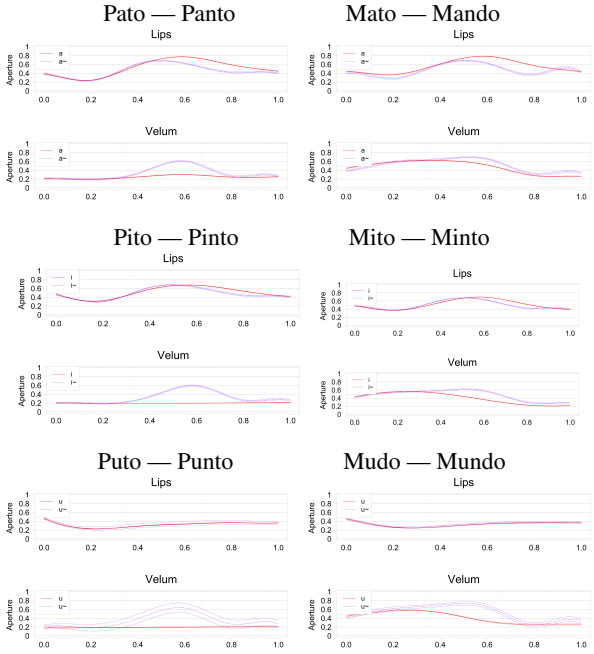

Figure 3: *FLMMs of lip and velar aperture, over time, for the different vowels and contexts, obtained from production data from 11 speakers. Curves for the oral vowels in red, and for the nasals in blue.*

A very interesting result is the stronger variation for the velar aperture for [ũ]. This is probably due to the backness of the vowel and the proximity of the tongue with the velum, potentially originating adjustments or less well defined timings (along with possible image artefacts, introducing a bit more uncertainty). This demands further investigation to disentangle the different possible causes. In this regard, also having data concerning the tongue body movement towards the pharynx might shed some light on what adjustments are involved. Regarding coordination, the FLMMs seem to provide a hint of an anticipatory closure/minimal area of the lips in respect to velar closure. However, to ascertain if this is confirmed – which would provide evidence supporting the occurrence of a small consonantic nasal tail – a more detailed analysis is required.

The changes introduced to the data input for the GAMMs, including data for the lips and velar aperture did not yield a clear depiction of the nasality differences among oral and nasal vowels. This might be a result of the smoothing effect of the GAMMs representation, given the limited spatial scope of the change. Further work to improve these aspects is required. In this regard, it would now be interesting to further test these methods with data extracted from tract segmentations to consider direct measures of the velopharyngeal passage and interlip distance as we considered in [11], also for RT-MRI.

6. Acknowledgements

A word of thanks is due to Dr Christopher Carignan for sharing the scripts serving as basis for the methods explored here. This work is partially funded by the German Federal Ministry of Education and Research (BMBF, KZ:01UL1712X), by IEETA Research Unit funding (UIDB/00127/2020), by Portugal 2020 under COMPETE 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).

7. References

[1] 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.

[2] 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.

[3] C. Oliveira and A. Teixeira, "On gestures timing in European Portuguese," in *ICPhS*, 2007, pp. p. 405 – 408.

[4] C. Oliveira, P. Martins, and A. Teixeira, "Speech rate effects on European Portuguese nasal vowels," in *InterSpeech*, 2009.

[5] P. Martins, I. Carbone, A. Silva, and A. Teixeira, "An MRI study of European Portuguese nasals," in *Interspeech*, 2007.

[6] ——, "European Portuguese MRI based speech production studies," *Speech Communication*, vol. 50, pp. 925–952, 2008.

[7] A. Teixeira, P. Martins, C. Oliveira, C. Ferreira, A. Silva, and R. Shosted, "Real-time MRI for Portuguese," in *Computational Processing of the Portuguese Language, PROPOR 2012, Lecture Notes in Computer Science/LNAI, Vol. 7243*, 2012.

[8] C. Oliveira, P. Martins, S. Silva, and A. Teixeira, "An MRI study of the oral articulation of European Portuguese nasal vowels," in *13th Annual Conference of the International Speech Communication Association (InterSpeech)*, Portland, USA, September 2012.

[9] 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.

[10] C. Cunha, S. Silva, A. Teixeira, C. Oliveira, P. Martins, A. A. Joseph, and J. Frahm, "On the Role of Oral Configurations in European Portuguese Nasal Vowels," in *Proc. Interspeech 2019*, 2019, pp. 3332–3336. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2232

[11] S. Silva, N. Almeida, C. Cunha, A. Joseph, J. Frahm, and A. Teixeira, "Data-driven critical tract variable determination for european portuguese," *Information*, vol. 11, no. 10, p. 491, 2020.

[12] C. Carignan, P. Hoole, E. Kunay, M. Pouplier, A. Joseph, D. Voit, J. Frahm, and J. Harrington, "Analyzing speech in both time and space: Generalized additive mixed models can uncover systematic patterns of variation in vocal tract shape in real-time MRI," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 11, no. 1, 2020.

[13] A. Teixeira, F. Vaz, and J. C. Príncipe, "Nasal vowels after nasal consonants," in *5th Seminar on Speech Production: Models and Data*, Kloster Seon, Alemanha, May 2000.

[14] S. Parkinson, "Portuguese nasal vowels as phonological diphthongs," *Lingua*, vol. 61, no. 2-3, pp. 157–177, 1983.

[15] A. Teixeira, F. Vaz, and J. C. Príncipe, "Influence of Dynamics in the Perceived Naturalness of Portuguese Nasal Vowels," in *14th International Congress of Phonetic Sciences (ICPhS 99)*, San Francisco, CA, E. U. A., Agosto 1999.

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

[17] F. Desmeules-Trudel, "The aerodynamics of vowel nasality and nasalization in brazilian portuguese." in *ICPhS*, 2015.

[18] 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.

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

[20] Q. Chao, "Data-driven approaches to articulatory speech processing," Ph.D. dissertation, University of California, Merced, 2011.

[21] M. P. Black, D. Bone, Z. I. Skordilis, R. Gupta, W. Xia, P. Papadopoulos, S. N. Chakravarthula, B. Xiao, M. Van Segbroeck, J. Kim *et al.*, "Automated evaluation of non-native english pronunciation quality: combining knowledge-and data-driven features at multiple time scales." in *Proc. Interspeech*, 2015, pp. 493–497.

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

[23] P. J. Jackson and V. D. Singampalli, "Statistical identification of articulation constraints in the production of speech," *Speech Communication*, vol. 51, no. 8, pp. 695 – 710, 2009.

[24] C. P. Browman and L. Goldstein, "Gestural specification using dynamically-defined articulatory structures," *Journal of Phonetics*, vol. 18, pp. 299–320, 1990.

[25] P. Boersma, "Praat: doing phonetics by computer [computer program]," *http://www. praat. org/*, 2020.