




Article

Data-Driven Critical Tract Variable Determination for European Portuguese [†]

Samuel Silva ^{1,*}, Nuno Almeida ¹ , Conceição Cunha ², Arun Joseph ³, Jens Frahm ³ 
and António Teixeira ¹ 

¹ Telecommunications and Informatics (DETI), Department of Electronics, Institute of Electronics and Informatics Engineering of Aveiro (IEETA), University of Aveiro, 3810-193 Aveiro, Portugal; nunoalmeida@ua.pt (N.A.); ajst@ua.pt (A.T.)

² Institute of Phonetics and Speech Processing, Ludwig-Maximilians-Universität München, 80333 München, Germany; cunha@phonetik.uni-muenchen.de

³ Max Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany; arun-antony.joseph@mpibpc.mpg.de (A.J.); jfracm@gwdg.de (J.F.)

* Correspondence: sss@ua.pt

[†] This article is an extended version of work presented by the authors at the International Conference on Computational Processing of Portuguese (PROPOR).

Received: 9 August 2020; Accepted: 16 October 2020; Published: 21 October 2020



Abstract: Technologies, such as real-time magnetic resonance (RT-MRI), can provide valuable information to evolve our understanding of the static and dynamic aspects of speech by contributing to the determination of which articulators are essential (critical) in producing specific sounds and how (gestures). While a visual analysis and comparison of imaging data or vocal tract profiles can already provide relevant findings, the sheer amount of available data demands and can strongly profit from unsupervised data-driven approaches. Recent work, in this regard, has asserted the possibility of determining critical articulators from RT-MRI data by considering a representation of vocal tract configurations based on landmarks placed on the tongue, lips, and velum, yielding meaningful results for European Portuguese (EP). Advancing this previous work to obtain a characterization of EP sounds grounded on Articulatory Phonology, important to explore critical gestures and advance, for example, articulatory speech synthesis, entails the consideration of a novel set of tract variables. To this end, this article explores critical variable determination considering a vocal tract representation aligned with Articulatory Phonology and the Task Dynamics framework. The overall results, obtained considering data for three EP speakers, show the applicability of this approach and are consistent with existing descriptions of EP sounds.

Keywords: critical articulator; critical tract variable; speech production model; data-driven approach; real-time magnetic resonance

1. Introduction

Major advances on phonetic sciences in the last decades contributed to better description of the variety of speech sounds in the world languages, to the expansion of new methodologies to less common languages and varieties contributing to a better understanding of spoken language in general. Speech sounds are not sequential nor isolated, but sequences of consonants and vowels are produced in a temporally overlapping way with coarticulatory differences in timing being language specific and varying according to syllable type (simplex, complex), syllable position (Marin and Pouplier [1] for timing in English, Cunha [2,3] for European Portuguese), and many other factors. Because of the coarticulation with adjacent sounds, articulators that are not relevant (i.e., noncritical) for the

production of the analysed sound can be activated. In this regard, noncritical articulators can entail gestures which are not actively involved in a specific articulation (e.g., a bilabial gestures for /p/), but show anticipatory movement influenced by the following vowel.

Despite the advances, our knowledge regarding how the sounds of the world languages are articulated is still fragmentary, being knowledge particularly limited for temporal organization, coarticulation, and dynamic aspects. These limitations in knowledge prevent both the advances in speech production theories, such as Articulatory Phonology, language teaching, speech therapy, and technological applications. More knowledge, particularly regarding the temporal organization and dynamic aspects, is essential for technologies such as articulatory and audiovisual synthesis. Recent developments show that articulatory synthesis is worth revisiting as a research tool [4,5], as part of text-to-speech (TTS) systems or to provide the basis for articulatory-based audiovisual speech synthesis [6].

Some articulatory based phonological descriptions of speech sounds appeared for different languages boosted by the increased access to direct measures of the vowel tract, such as electromagnetic articulography (EMA) or magnetic resonance imaging (MRI). For EP there are some segmental descriptions using EMA [7] and MRI (static and real time, for example, Reference [8]) and some work on onset coordination using EMA [2,3]. Also an initial description of EP adopting the Articulatory Phonology framework was proposed [7]. However, these descriptions analysed a reduced set of images or participants and need to be revised and improved.

Recent advances in the collection and processing of real-time MRI (RT-MRI) [9,10] are promising to improve existing descriptions providing high temporal and spatial resolutions along with automated extraction of the relevant structures (for example, the vocal tract [11,12]). The huge amounts of collected data are only manageable by proposing automatic data-driven approaches. In these scenarios with huge amounts of data need to be tackled (e.g., RT-MRI, EMA), the community has made an effort to contribute with methods to extract and analyse features of interest [11,13–16]. In order to determine the more important articulator for each specific sound, articulator criticality, several authors have proposed data-driven methods, for example, References [17–22]. Of particular interest, the statistical method proposed by Jackson and Singampali [19] considers the position of the EMA pellets as representative of the articulators and uses a statistical approach to determine the critical articulators for each phone.

The authors' previous work [23–25] explored those computational methods initially proposed for EMA [19,26], to determine the critical articulators for EP phones from real-time MRI data at 14 and 50 fps. demonstrating the applicability of the methods to MRI data, and present (albeit preliminary) interesting results. To further pursue this topic, this work presents first results towards critical variable (articulator) determination considering a representation of the vocal tract aligned with the Articulatory Phonology and the Task Dynamics framework [27]. It extends previous work [25], further confirming our previous findings, by: (a) considering more articulatory data, both increasing it for previously considered speakers and adding a novel speaker; (b) improving on the computation of tract variables (particularly the velum); (c) adding additional analysis details for all speakers (e.g., regarding the correlation among tract variables); and (d) tackling a more indepth exploration of individual tract variable components (e.g., constriction degree or velopharyngeal passage).

The remainder of this document is organized as follows: Sections 2 and 3 provide a brief overview on relevant background and related work; Section 4 provides an overall description of the methods adopted to acquire, annotate, process, and revise articulatory data obtained from RT-MRI of the vocal tract to determine critical articulators; Section 5 presents the outcomes of the main stages of the critical articulator determination obtained considering tract variables aligned with the Task Dynamics framework and these are discussed in Section 6; finally, Section 7, highlights the main contributions of this work and proposes routes for future efforts.

2. Background: Articulatory Phonology, Gestures and Critical Tract Variables

Speech sounds are not static target configurations clearly defined, their production involves complex tempo-spatial trajectories in the vocal tract articulators responsible for their production from the start of the movement till the release and back (e.g., in bilabial /p/ both lips move until the closure that produces the bilabial and the lips open again). All this movement is a so called articulatory gesture. Instead of phonological features, the dynamic gestures are the unities of speech in Articulatory Phonology [27,28] and define each particular sound. Therefore, gestures are, on one hand, the physically tractable movements of articulators that are highly variable, depending, for example, on context and speaking rate, and, on the other hand, the representations of motor commands for individual phones in the minds of the speakers which are invariant. In other words, they are both instructions to achieve the formation (and release) of a constriction at some place in the vocal tract (for example, an opening of the lips) and abstract phonological units with a distinctive function [27].

Since the vowel tract is contiguous, more articulators are activated simultaneously than the intended ones. Consequently it is important to differentiate between the actively activated (critical) articulators and the less activated or passive ones: For example, in the production of alveolar sounds as /t/ or /l/, tongue tip needs to move up in the alveolar region (critical articulator) and simultaneously also tongue back and tongue body show some movement, since they all are connected. For laterals, for example, also tongue body may have a secondary importance in their production [29,30]. Some segments can be defined only based on one or two gestures, bilabials are defined based on the lips trajectories; laterals as mentioned before, are more complex and may include tongue tip and tongue body gestures.

Gestures are tempo-spatial entities, structured with a duration and a cycle. The cycle begins with the movement's onset, continues with the movement toward the target – that can be reached or not –, then to the release, where the movement away from the constriction begins, ending with the offset, where the articulators cease to be under active control of the gesture. Individual gestures are combined to form segments, consonant clusters, syllables, words.

Gestures are specified by a set of tract variables and their constriction location and degree: Tract variables are related to the articulators and include: Lips (LIPS), Tongue Tip (TT), Tongue Body (TB), Velum (VEL) and Glottis (GLO); Constriction location specifies the place of the constriction in the vocal tract and can assume the values: labial, dental, [alveolar, postalveolar, palatal, velar, uvular and pharyngeal]; Constriction degree includes: closed (for stops), critical (for fricatives), narrow, mid and wide (approximants and vowels). For example, a possible specification for the alveolar stop /t/ in terms of gestures is Tongue Tip [constriction degree: closed, constriction location: alveolar] [31].

The tract variables involved in the critical gestures are considered critical tract variables and the involved articulators the critical articulators.

The articulatory phonology approach has been incorporated into a computational model by Haskins Laboratories researchers [32,33]. It is composed by three main processes, in sequence: (1) Linguistic Gestural Model, responsible for transforming the input into a gestural score (set of discrete, concurrently active gestures); (2) Task Dynamic Model [32,33], that calculates the articulatory trajectories given the gestural score; and (3) Articulatory Synthesizer, capable of, based on the articulators' trajectory, obtaining the global vocal tract shape, and, ultimately, the speech waveform.

3. Related Work

The following sections provide a summary of existing descriptions for European Portuguese sounds and overview previous work on using data-driven methods to determine critical articulators deemed relevant to contextualize the work presented in this article.

3.1. Gestural Descriptions of (European) Portuguese

Essentially manual analyses of a limited set of MRI images and contours made possible the first description adopting the principles of Articulatory Phonology for European Portuguese [31,34]. In a very summarized form, those aspects deemed relevant as background for the present work are References [31,34]:

- Vowels can be adequately characterized using one or two tract variables: Tongue Body alone or combined with Lips;
- Stops and fricatives consist essentially in the definition of constriction degree and location;
- For nasal consonants velum must also be defined;
- Laterals demand the additional inclusion of a shape variable;
- Taps and trills imply control of parameters such as stiffness;
- The bilabials /b, m/ are specified here for the gestures Lips [closed]. Additionally, the nasal is specified for Velum [wide];
- The correct definition of laterals would demand the addition of a constriction shape variable to the initial set proposed by Articulatory Phonology;
- In the production of the alveolar lateral, /l/, two coordinated oral gestures, TT [narrow, alveolar] and a TB [narrow, velar], are involved .
- The tap ([R]) is produced with a short duration TT gesture [closed, alveolar].

3.2. Computational Approaches to Critical Gesture Determination

As previously mentioned, several authors have proposed data-driven approaches to harness large amounts of articulatory data to advance our knowledge regarding speech production [21,22] and critical articulators, in particular, for a wide range of contexts, such as emotional speech [17], and exploring different approaches, for example, time–frequency features [18]. In a notable example, Jackson and Singampalli [19], consider a large set of articulatory data from EMA to build statistical models for the movement of each articulator. This is performed by selecting data samples, at the midpoint of each phone, and computing statistics describing: (1) the whole articulator data (the grand statistics), used to build the models for each articulator; and (2) the data for each phone (phone statistics). The critical articulators, for each phone, are determined by analysing the distances between the grand and phone probability distributions. By considering a static tract configuration for each of the phones, the dynamic nature of the data is not explored. Nevertheless, by doing so, Jackson and Singampalli present a method that is quite useful in providing clear and interpretable insights regarding articulation and enabling a comparison with existing phonological descriptions.

For European Portuguese (EP), the authors have been exploring approaches based on articulatory data extracted from midsagittal RT-MRI images of the vocal tract to automatically determine the critical articulators for each EP sound. While RTMRI offers a large amount of data of the whole vocal tract, over time, the main goal, in a first stage, was to leave the dynamic aspects out and pursue a data-driven method that could provide easily interpretable results for an automated phonological description. This entailed privileging approaches with a specific consideration of the anatomical regions of interest, rather than methods dealing with the overall shape of the vocal tract.

In a first exploratory work [23], the authors asserted that critical articulator identification could be performed by extending the applicability of the method proposed for EMA data by Jackson and Singampalli [19]. Those first results were obtained for RT-MRI at 14Hz and, since the original method worked with 100 Hz EMA, the question remained regarding if higher RT-MRI frame rates, along with a larger dataset, could have a positive impact on the outcomes. Since one frame is used as representative of the vocal tract configuration for each sound, a higher frame rate, might enable capturing key moments of articulation, for example, alveolar contact for the /n/. To address these aspects, in Silva et al. [24], the authors explored 50Hz RT-MRI data showing both the applicability of the methods to the novel data along with noticeable improvement of the results. At this point, the representation of the vocal tract configurations was still being performed based on landmarks placed on the lips,

tongue surface, and velum, to establish similar conditions as those of the method proposed for EMA. However, the tract data available from RT-MRI can potentiate the consideration of different tract variables moving beyond simple landmarks. In this regard, the authors have explored tract variables aligned with the Task Dynamics framework [25], adopting the concept of constrictions (except for the velum), and showed that these provided an alternative that was not only more compact (less variables involved), but also provided interesting critical articulator results with the benefit of having a more direct relation with existing Articulatory Phonology descriptions, supporting easier comparison with the literature. Nevertheless, the authors identified a few points deserving immediate attention to further assess and improve the tested approach: (1) enlarge the amount of data considered per speaker; (2) consider additional speakers; and (3) explore a novel representation for the velum, to completely avoid landmarks. These aspects are addressed in the present work and described in what follows.

4. Methods

The determination of critical articulators is performed from articulatory data extracted from real-time magnetic resonance imaging (RT-MRI). The overall pipeline is presented in Figure 1 and its main aspects are detailed in what follows.

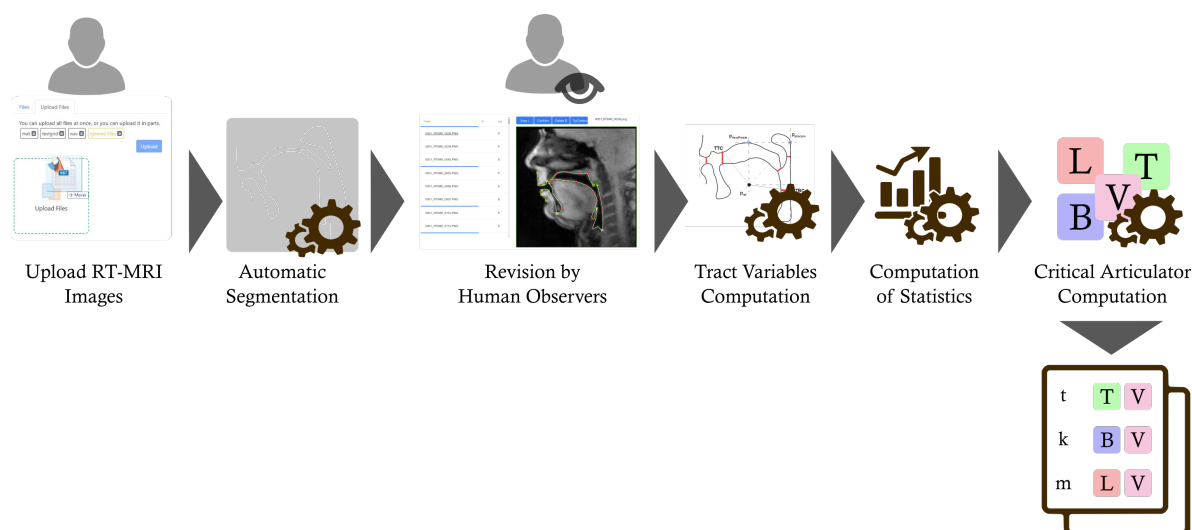


Figure 1. Overall steps of the method to determine the critical articulators from real-time MRI (RT-MRI) images of the vocal tract. After MRI acquisition and audio annotation, the data is uploaded to our speech studies platform, under development [35], and its processing and analysis are carried out resulting in a list of critical tract variables per phone. Refer to the text for additional details.

4.1. Materials and Participants

The corpus consisted of lexical words containing all oral [i, e, ε, a, o, ɔ, u, ʊ] and nasal vowels [ẽ, ê, î, ô, û] in one and two syllables words. Oral and nasal diphthongs as well as additional materials including alternations of nasal monophthongs and diphthongs as in ‘som’ (‘sound’) and ‘são’ (‘they are’) or ‘antaram’ (‘they sang’) and ‘cantarão’ (‘they will sing’) were recorded for further investigation of variability in the production of nasality. Due to the strong research question on nasality behind these recordings, the occurrence of single segments is strongly unbalanced. Unstressed oral and nasal vowels were added to the corpus after the third recording. All words were randomized and repeated in two prosodic conditions embedded in one of three carrier sentences, alternating the verb (‘Diga’—‘Say’; ‘ouvi’—‘I heard’; or ‘leio’—‘I read’) as in ‘Diga pote, diga pote baixinho’ (‘Say pot, Say pot gently’). The sentences were presented on a screen in randomized order with three repetitions. So far, this corpus has been recorded for sixteen native speakers (8 m, 8 f) 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 s and there

was a pause of about 60 s after each block of 13 stimuli. 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).

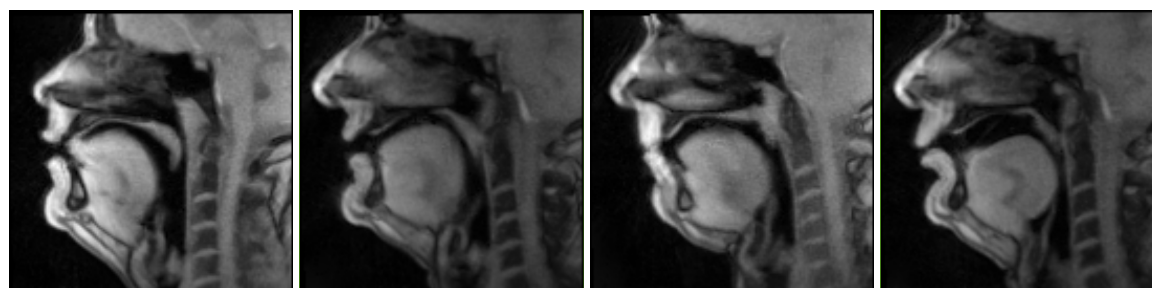
The participants were familiarized with the corpus and the task previously, in which they read the corpus with 2 or 3 repetitions in a noise reduced environment. During the RT-MRI experiment, they were lying down, in a comfortable position, and were instructed to read the required sentences projected in front of them. All volunteers provided informed written consent and filled an MRI screening form in agreement with institutional rules prior to the enrollment on the study. They were compensated for their participation and none of them reported any known language, speech or hearing impairment.

4.2. Image Acquisition and Corpus

Real time MRI acquisition was performed at the Max Planck Institute for Biophysical Chemistry, Göttingen, Germany, using a 3T Siemens Magnetom Prisma Fit MRI System 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. Real-time MRI measurements were based on a recently developed method, where highly undersampled radial FLASH acquisitions are combined with nonlinear inverse reconstruction (NLINV) providing images at high spatial and temporal resolutions [36]. Acquisitions were made at 50 fps, resulting in images as the ones presented in Figure 2. 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.

Figure 2 presents some illustrative examples of MRI images for different speakers and sounds and Figure 3 shows the image sequence corresponding to the articulation of /p/ as in [pənɛtɐ].

After the audio has been annotated all the data concerning a speaker (images, audio, and annotations) is uploaded to a speech studies platform in development by the authors [35] where the following steps take place.



(a) speaker 8458, /n/ (b) speaker 8460, /ɛ/ (c) speaker 8545, /p/ (d) speaker 8460, /ū/

Figure 2. Illustrative examples of midsagittal real-time MRI images of the vocal tract, for different speakers and sounds.

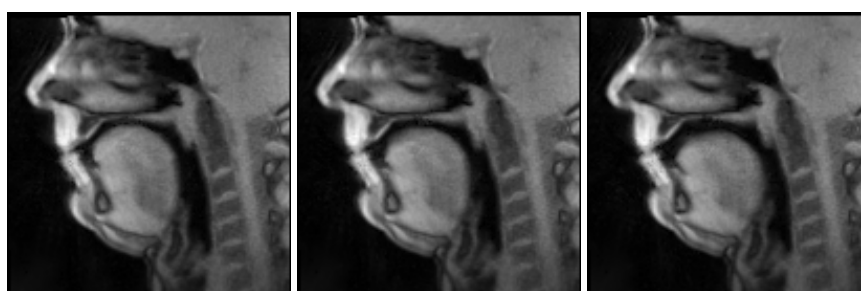


Figure 3. Cont.

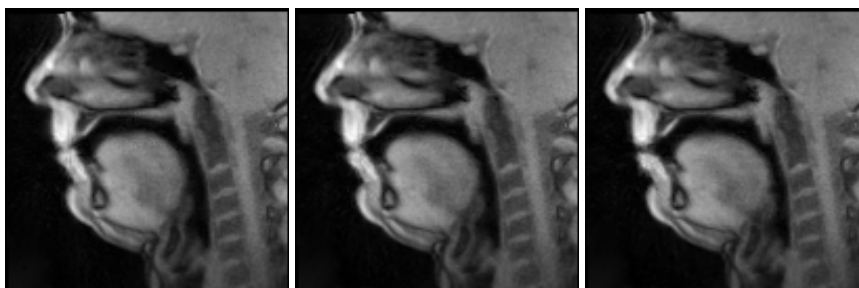


Figure 3. Midsagittal real-time MRI image sequence of speaker 8545 articulating /p/ as in the nonsense word [pɛnetɛ]. The images have been automatically identified considering the corresponding time interval annotated based on the audio recorded during the acquisition. Note the closed lips, throughout and their opening, in the last frame, to produce the following /ɛ/.

4.3. Vocal Tract Segmentation and Revision

The RT-MRI sequences were processed considering the method presented in Reference [11] to extract the vocal tract profiles. Based on a small set of manually annotated images, typically one for each of the phones present in the corpus, to initialize the method, the RT-MRI image sequences are automatically segmented to identify the contour of the vocal tract. Figure 4 presents a few examples of the resulting segmentations for a representative frame of /p/ and /n/ for the three speakers.

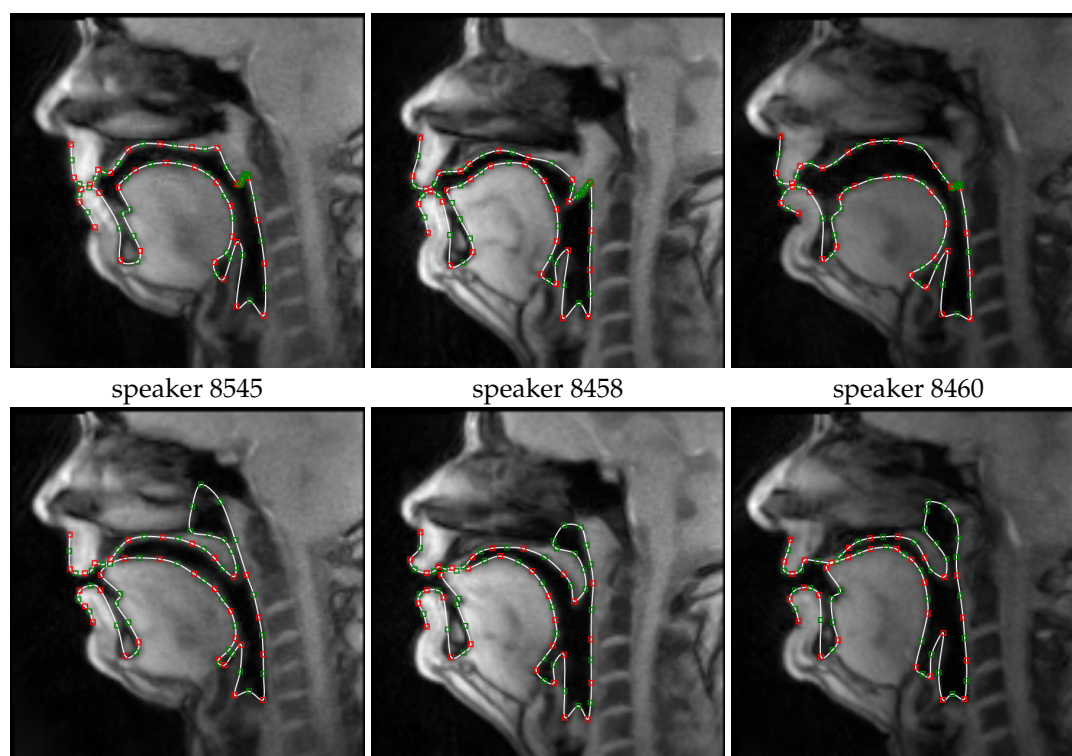


Figure 4. Illustrative examples of the automatically segmented vocal tract contours represented over the corresponding midsagittal real-time MRI images for three speakers uttering /p/, on the top row, and /n/ on the bottom row.

One aspect that was considered paramount, at the current stage of the research, was to ensure that the vocal tract data considered for the analysis had been segmented properly, that is, that all relevant structures had no segmentation errors. A few of these errors, for example, a wrongly segmented tongue tip, could disturb the tract variable data distributions and have unpredictable influences over the statistics affecting critical tract variable determination. Such effects could, then, hinder a correct

assessment of the capabilities of the method. Therefore, all the segmentations considered were checked by human observers and, when required, revised to perform, for example, fine adjustment of the segmentation at the tongue tip, velum or lips. The revisions were performed by five observers, who revised the segmentations for different images, using the same revision tool, and care was taken so that each observer would only revise a subset of any sound/context to avoid observer bias effect as a factor potentially influencing the outcomes of the critical articulator analysis.

Finally, one aspect that was observed for the revised segmentations was that the hard palate could be prone to slight variations of its shape due to a difficulty of the automatic method in establishing its precise location. This was mostly due to the fact that it is a region that is not imaged very clearly with MRI, which also made its manual revision difficult. While these differences were not large, overall, they could impact, for example, the determination of the location of the highest constriction of the tongue body, potentially causing differences of this variable among occurrences of the same sound/context (e.g., with location transitions between tongue back and tongue blade). Therefore, for all the computations presented, the hard palate for each sample was replaced by the mean hard palate across all considered contours, for each speaker.

4.4. Tract Variables Computation

Differently from the original application of the method to EMA [19] and subsequent extension to RT-MRI data by our team [24], for this work the considered variables are not landmarks placed over the vocal tract, but tract variables aligned with the Task Dynamics framework [27,37], that is, mostly based on the concepts of constrictions defined by their location and degree (distance). After preliminary work by the authors [25], which provided interesting results and has shown the applicability of the critical articulator method to this new set of variables, this work further extends this approach by also abandoning the landmark representation for the velum, identified as a limitation in our previous studies.

4.4.1. Choice and Computation of Tract Variables

Aligned with Articulatory Phonology, we considered the variables depicted in Figure 5. With this set of variables we move away from choosing fixed points over the tongue, as in previous work. Instead, maximal constrictions are determined between different tract segments (e.g., tongue tip and hard palate, identified based on the segmentation data [16]) as follows:

- The **tongue tip constriction (TTC)** is determined as the minimal distance between the tongue tip region and the hard palate. A small segment of the tongue contour, in the neighborhood of the tongue tip, is selected and the distances from each point to the hard palate contour segment points are computed. Of those, the minimal distance is determined and the constriction distance (TTCd) and location (TTCl) obtained;
- The **tongue body constriction (TBC)** is determined as the minimal distance between the tongue body and the pharyngeal wall or hard palate (not including the velar region). The distance between all points of the tongue contour segment (minus those considered for the the tongue tip neighborhood) and all the points in the pharyngeal and palatal segments is computed. The smallest distance obtained corresponds to the point of maximal tongue body constriction. The constriction distance (TBCd) and location (TBCl) is thus obtained. In the example presented in Figure 5, the smallest distance was found between a point located in the tongue back and a point in the pharyngeal wall;
- The **velar configuration (V)** is determined by obtaining the constriction distance between the velum and the pharyngeal wall contour segments (velopharyngeal port, Vp) and between the velum and the tongue body (oral port, Vt);
- The **lips (LIPS)** configuration is characterized by their aperture (LIPa), computed as the minimum distance between the contour segments of the upper and lower lips, and protrusion (LIPp) as the horizontal distance from the left most point of the lips to the reference point p_{ref} . While this

does not provide just the lip protrusion distance (having rest as a reference), it is suitable for the intended analysis without having to determine minimum lip protrusion beforehand.

In previous work [25], the authors have shown that a change from the vocal tract representation based on landmarks (mimicking the work done for EMA [19]) to an approach more aligned with the Task Dynamics framework reduced the correlation between variable components (e.g., x and y coordinates of the landmark) and among tract variables. This potentially entails that the data provided by each component and variable is overall, more independent, and might offer added insight regarding what is critical. One of the variables that remained as a landmark concerned the velum, represented by the x and y coordinates of a point placed at its back (see Reference [23] for details). To move into an approach more consistent with Articulatory Phonology a novel representation was adopted describing the velum configuration based on the velo-pharyngeal and orovelar constriction distances.

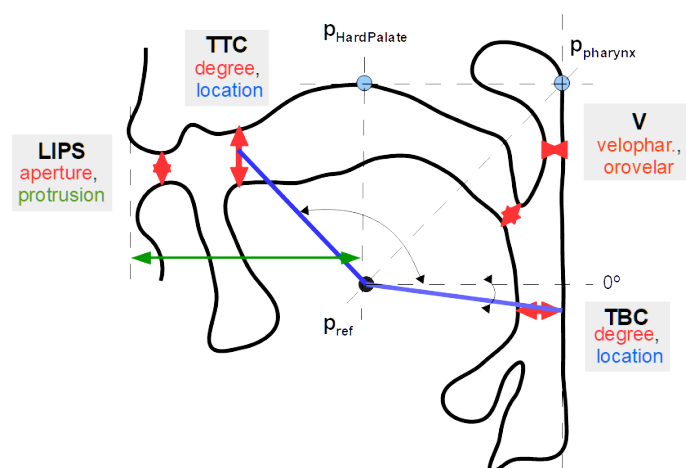


Figure 5. Illustrative vocal tract representation depicting the main aspects of the considered tract variables: tongue tip constriction (TTC, defined by degree and location); tongue body constriction (TBC, defined by degree and location), computed considering both the pharyngeal wall and hard palate; velum (V, defined by the extent of the velopharyngeal and orovelar passages); and lips (LIPS, defined by aperture and protrusion). The point p_{ref} is used as a reference for computing constriction angular locations. Please refer to the text for further details.

4.4.2. Determining Reference Point for Constriction Location

The first step for the determination of the constriction location is the definition of a referential to measure the location angle (see Figure 5). For each speaker, the highest point of the hard palate is determined ($p_{HardPalate}$) and the point of intersection between a tangent that passes this point and the orientation of the pharyngeal wall is computed ($p_{pharynx}$). The reference point (p_{ref}) is, then, obtained as the intersection of a vertical line passing $p_{HardPalate}$ and a line passing $p_{pharynx}$ at a 45° angle.

4.5. Data Selection

The critical articulator determination requires that a representative frame is selected for each occurrence of each phone. Table 1 shows the phones considered for analysis taking into account the contents of the corpus. Since that different sounds have a different progression, over time, the frame considered to represent the vocal tract configuration is selected from the annotated interval using slightly different criteria, as explained in Table 1. For instance, for /p/, the frame with the minimum inter-lip distance is selected while for oral vowels the middle frame is considered.

Similarly to Silva et al. [24,25], and considering the dynamic nature of nasal vowels [8,34,38,39], we wanted to have some additional information to assert if, for different timepoints, the determination of the critical articulators would highlight any relevant differences in behavior. Therefore, each nasal vowel was represented by three “pseudo-phones”, focusing the starting, middle and final frame of the

annotated interval and named, respectively, [vowel]^{start}, [vowel]^{mid} and [vowel], as in: \tilde{a}^{start} , \tilde{a}^{mid} , and \tilde{a} .

Table 1. Summary of the criteria used for selecting the representative frame for particular phones.

Phone (SAMPA)	Criterion
Oral Vowels 6, a, e, E, i, o, O, u	midpoint
Nasal Vowels 6̃, ẽ, ĩ, õ, ũ	three classes were created, taking the first, middle, and final frames
Nasal Consonants m, n	[m], frame with minimum inter-lip distance; [n], midpoint
Stops p, b, k, d, g, t	[p] and [b], frame with minimum inter-lip distance; [k],[d], [g] and [t], midpoint
Fricatives s, v	midpoint

4.6. Determination of Critical Articulators

Critical articulator identification requires the grand statistics, characterizing the distribution, for each variable, along the whole data; and the phone statistics, representing the distribution of the variable, for each phone, considering the phone data selection. Table 2, to the right, summarizes the different statistics computed to initialize the method, adopting the notation as in Jackson and Singampali [19]. Critical articulator identification was performed taking articulator properties (e.g., d , degree and l , location, for the constrictions) independently—the 1D case—for example, TBd for the constriction degree at the tongue body, or combining them—the 2D case.

Table 2. Summary of the computed statistics for each landmark and corresponding notation as in Reference [19].

Grand Stats	Not.	Comment
grand mean	M	all selected frames
grand variance	Σ	all selected frames
total sample size	N	Spk 8458: 870; Spk 8460: 750; Spk 8545: 853; Spk All: 2473;
corr. matrix	R^*	keeping statistically significant and strong correlations ($r_{ij} > 0.2$ and $\alpha = 0.05$)
Phone Stats	Not.	Comment
mean	μ^ϕ	frames selected for each phone
variance	Σ^ϕ	frames selected for each phone
sample size	v^ϕ	variable among phones
corr. matrix	R^ϕ	not attending to significance and module

The 1D correlation matrices for the articulators (e.g., considering TBl and TTd, etc.), given the size of our data set, was computed considering correntropy, as proposed in Rao et al. [40]. Bivariate correlations (i.e, taking both properties of each articulator together) were computed through canonical correlation analysis [19,41]. For the grand correlation matrices, adopting the criteria proposed in Reference [19], only statistically significant ($\alpha = 0.05$) correlation values above 0.2 were kept, reducing the remaining ones to zero. The computed data statistics were used to initialize the critical articulator analysis method and 1D and 2D analysis was performed, for each speaker, returning a list of critical articulators per phone.

Additionally, we wanted to assess how the method would work by gathering the data for the three speakers to build a “normalized” speaker. This would enable the consideration of a larger dataset potentially providing a clearer picture on the overall trends for the critical articulators for an EP speaker disentangled from specific speaker characteristics. To that effect, we normalized the articulator

data, for each speaker, based on the variation ranges, for each variable component, computed over all selected data samples, and considered this gathered data as a new speaker following a similar analysis methodology. This is, naturally, a very simple normalization method, but considering the overall results obtained in previous work [25], it was deemed appropriate for this first assessment including three speakers.

To determine the list of critical articulators, the analysis method requires establishing a stopping threshold, Θ_C . If the threshold is low, then, a large number of critical articulators will potentially appear, for each sound. Higher thresholds will potentially result in shorter (possibly empty) lists being determined for some of the sounds. The impact of changing the threshold does not affect all the sounds in the same way and relates with the amount of data available and with its diversity, i.e., a certain threshold may yield an empty list for some sounds and a long list for others (i.e., including articulators of less importance). The founding work of Jackson and Singampalli [26], serving as grounds for what we present here, have described this aspect and, as we have also observed, in previous works, this threshold is variable among speakers.

As in our previous work, we defined a stopping threshold, Θ_C , for each of the speakers (including the normalized speaker), as the highest possible value that would ensure that each phone had, at least, one critical articulator. This resulted in the inclusion of less important articulators for some of the phones, but avoided that phones with a smaller amount of data had no results.

5. Results

The conducted analysis has two main outcomes: the correlation between tract variables and the determination of their criticality for the production of different sounds. Additionally, each of these aspects can be analysed from the 1D perspective, where each dimension of the variables is taken independently, for example, lip protrusion and aperture are considered two variables, or the 2D perspective, with each variable as a bidimensional entity. In relation with previous work, particularly Silva et al. [25], these results are obtained for a larger number of data samples, consider a different definition of the tract variable for the velum, and include data for a new speaker, also influencing the normalized speaker data.

5.1. Tract Variable Correlation

Figure 6 shows the 1D correlation matrix for the different tract variable components, for the three speakers and the normalized speaker (All). Overall, two sets of mild correlated variable dimensions appear, although differently for the considered speakers: LIP protrusion and aperture; and TT constriction degree and location. As expected, by changing how the velar tract variable is represented, moving from the x and y coordinates of a landmark (as tested by the authors of Reference [25]) into the velopharyngeal and orovelar passages, has made the correlation between them almost disappear.

The consideration of additional data, including more contexts for each phone, had an effect on the results for speaker 8460, when comparing to what was previously observed [25], since the mild correlation between the TB and the LIPS and TT has disappeared. Interestingly, speaker 8545 (along with the normalized speaker) does not show any relevant correlation between any of the variable dimensions.

Table 3 shows the canonical correlation [19,41] computed among the different tract variables.

Overall, this table provides information about how the different tract variables correlate and further confirms what was observed for the 1D analysis. While the observed correlations are small/mild, a corpus with a greater phonetic richness would lower them further. For instance, if the corpus included the lateral /l/, this would probably further emphasize the independence of the TT towards the TB.

8458	LIPSa	LIPSp	TTCd	TTCl	TBCd	TBCl	Vp	Vt
LIPSa	1.00	0.52	0.25	0.00	0.00	0.00	0.00	0.31
LIPSp	0.52	1.00	0.00	0.00	0.00	0.00	0.00	0.41
TTCd	0.25	0.00	1.00	0.59	0.00	0.33	0.00	0.30
TTCl	0.00	0.00	0.59	1.00	0.22	0.31	0.00	0.27
TBCd	0.00	0.00	0.00	0.22	1.00	0.48	0.25	0.00
TBCl	0.00	0.00	0.33	0.31	0.48	1.00	0.00	0.32
Vp	0.00	0.00	0.00	0.00	0.25	0.00	1.00	0.00
Vt	0.31	0.41	0.30	0.27	0.00	0.32	0.00	1.00

8460	LIPSa	LIPSp	TTCd	TTCl	TBCd	TBCl	Vp	Vt
LIPSa	1.00	0.50	0.38	0.27	0.36	0.29	0.00	0.00
LIPSp	0.50	1.00	0.48	0.32	0.00	0.46	0.00	0.29
TTCd	0.38	0.48	1.00	0.53	0.29	0.43	0.21	0.22
TTCl	0.27	0.32	0.53	1.00	0.33	0.27	0.30	0.00
TBCd	0.36	0.00	0.29	0.33	1.00	0.40	0.00	0.39
TBCl	0.29	0.46	0.43	0.27	0.40	1.00	0.00	0.00
Vp	0.00	0.00	0.21	0.30	0.00	0.00	1.00	0.38
Vt	0.00	0.29	0.22	0.00	0.39	0.00	0.38	1.00

8545	LIPSa	LIPSp	TTCd	TTCl	TBCd	TBCl	Vp	Vt
LIPSa	1.00	0.25	0.35	0.24	0.23	0.00	0.00	0.26
LIPSp	0.25	1.00	0.23	0.25	0.00	0.00	0.30	0.38
TTCd	0.35	0.23	1.00	0.31	0.00	0.33	0.00	0.33
TTCl	0.24	0.25	0.31	1.00	0.00	0.00	0.00	0.00
TBCd	0.23	0.00	0.00	0.00	1.00	0.31	0.00	0.25
TBCl	0.00	0.00	0.33	0.00	0.31	1.00	0.00	0.00
Vp	0.00	0.30	0.00	0.00	0.00	0.00	1.00	0.00
Vt	0.26	0.38	0.33	0.00	0.25	0.00	0.00	1.00

All	LIPSa	LIPSp	TTCd	TTCl	TBCd	TBCl	Vp	Vt
LIPSa	1.00	0.32	0.32	0.00	0.00	0.00	0.00	0.00
LIPSp	0.32	1.00	0.00	0.00	0.00	0.00	0.22	0.28
TTCd	0.32	0.00	1.00	0.42	0.00	0.34	0.00	0.21
TTCl	0.00	0.00	0.42	1.00	0.24	0.24	0.00	0.00
TBCd	0.00	0.00	0.00	0.24	1.00	0.41	0.00	0.00
TBCl	0.00	0.00	0.34	0.24	0.41	1.00	0.00	0.00
Vp	0.00	0.22	0.00	0.00	0.00	0.00	1.00	0.00
Vt	0.00	0.28	0.21	0.00	0.00	0.00	0.00	1.00

Figure 6. Correlation among the different components of the considered tract variables (1D correlation) for the three speakers 8458, 8460 and 8545, and for the speaker gathering the normalized data. **Tract variables for 1D correlation:** LIPSa: lip aperture; LIPSp: lip protrusion; TTCd: tongue tip constriction distance; TTCl: tongue tip constriction location; TBCd: tongue body constriction distance; TBCl: tongue body constriction location; Vp: velar port distance; Vt: orovelar port distance.

Table 3. Canonical correlation for pairs of the chosen tract variables (2D canonical correlation) for the three speakers (8458, 8460, and 8545) and for the speaker gathering the normalized data.

	8458	8460	8545	All				
$\rho_{LIPS,TTC}$	0.26	0.00	0.51	0.00	0.43	0.00	0.37	0.00
$\rho_{LIPS,TBC}$	0.00	0.00	0.47	0.31	0.25	0.00	0.00	0.00
$\rho_{LIPS,V}$	0.44	0.00	0.37	0.00	0.50	0.00	0.37	0.00
$\rho_{TTC,TBC}$	0.36	0.00	0.47	0.00	0.33	0.00	0.42	0.00
$\rho_{TTC,V}$	0.33	0.21	0.38	0.25	0.34	0.00	0.32	0.00
$\rho_{TBC,V}$	0.32	0.30	0.49	0.00	0.36	0.00	0.30	0.00

Tract variables for 2D correlation: LIPS: lips; TTC: tongue tip constriction; TBC: tongue body constriction; V: velopharyngeal and orovelar ports.

5.2. Critical Articulators

Table 4 shows the 1D analysis of critical articulators, for each phone, speaker, and the normalized speaker. Please note that, for the sake of space economy, the name of the tract variables is simplified by removing the T and C, for example, TBCl, becomes Bl. Since each variable dimension is treated independently, it provides a finer grasp over which particular aspect of the variable is most critical, for example, if it is the constriction degree or its location. For the sake of space, for those phones with a list of critical articulators longer than four, only the first four are presented. Considering that the order of the articulators is important, the remaining elements of the list were judged less important to provide any further elements for discussion.

Regarding the determination of critical tract variables, Table 5 presents the results obtained for each speaker (columns 8458, 8460 and 8545) and for the normalized data (column ALL). As with the 1D case, the analysis was performed considering a conservative stopping threshold, Θ_C , to avoid the appearance of phones without, at least, one critical articulator. Note, nevertheless, that the order of the articulators is meaningful, starting from the one more strongly detected as critical. The rightmost column, shows the characterization of EP sounds based on the principle of Articulatory Phonology as reported by Oliveira [7] to be considered as a reference for the analysis of the obtained results.

Table 4. Critical articulators for the different phones and speakers. Each component of the considered tract variables is considered an articulator (1D analysis). For the sake of brevity, for phones yielding a list of more than four critical articulators, only four articulators are presented. The order of the articulators reflects their determined importance. For the sake of space economy, in the tract variable listing the T and C were omitted, for example, TBCd became Bd.

ph	spk 8458	spk 8460	spk 8545	spk All
e	Tl	B1 Tl	Tl	Vt Tl Lp
a	Tl B1 La Lp	B1 Lp Tl La	Tl B1	Tl Lp Vt Bd
e	Tl B1	Tl B1 Lp La	Vp Tl	Tl B1 Lp Vt
ε	Vp Tl	B1 Lp Tl La	Vp Tl	B1 Vp Tl Vt
i	B1 Tl Bd Td	B1 Tl Bd Vp	Vp Tl	B1 Td Lp Bd
o	Vp Tl	Vp B1	Vp B1	Vp Tl
ɔ	B1 Tl Td Lp	B1 Tl Lp Bd	Vp Tl	B1 Bd Lp Tl
u	Tl Td B1	Vp Tl	Vp Tl	Tl Vt La
ẽ ^{start}	Tl Lp Td B1	Vp B1	Tl Vp Lp B1	Vt Tl Bd
ẽ ^{mid}	Tl Lp B1 La	Vp Tl	Tl B1 Vp Lp	Tl Vt Bd Td
ẽ	Tl Lp Td B1	Lp B1 Tl Bd	Tl B1 Vp Lp	Tl Vt Bd
ẽ ^{start}	Tl Lp B1	Tl B1 Lp	Vp Tl	Tl Lp
ẽ ^{mid}	Tl Lp B1 Vp	Tl B1 Lp Vp	Tl B1	Tl Vt Lp Td
ē	B1 Tl Vp Lp	Tl Vp Lp B1	Tl B1	Lp Tl B1 Vt
ĩ ^{start}	Tl B1 Bd Vt	Tl B1 Lp Vp	B1 Tl	B1 Tl Bd Td
ĩ ^{mid}	Bd Tl Vt Lp	Tl B1 Vp Bd	B1 Tl Vp Lp	B1 Tl Bd Td
ĩ	Bd Tl Vt Lp	Tl B1 Vp Bd	Tl B1 Vp Lp	B1 Tl Td Bd
ō ^{start}	La Tl Lp B1	B1 Tl Td Lp	Tl Td Vp	Bd Tl Vt
ō ^{mid}	B1 Tl La Lp	B1 Tl Vp Lp	Tl B1 Vp Td	Bd B1 Tl Vt
ō	Tl B1 La Lp	B1 Tl Vp Lp	Tl Vp B1	Bd Tl Lp
ũ ^{start}	La Td Tl B1	Tl Vt B1 Vp	B1 Tl Td Vp	Td Vt Tl Bd
ũ ^{mid}	Tl Vt B1 Td	Tl Vt B1 Lp	Tl Vp Td	Vt Tl La Td
ũ	Vt B1 Lp Tl	B1 Tl Vp Lp	Vp Tl Td	La Vt Lp
m	La Tl Vp Lp	Vp Tl B1 Lp	Tl Vp La Lp	La Tl Vt
n	Td Tl Vp	Td Tl Vp B1	Vp Td Tl	Td B1 Lp Vp
ŋ	Vp Tl Td	Vp Tl B1 Lp	Vp Tl Td	Tl Vp
p	La Lp Tl	Vp La	Vp Tl	La Tl Vp
b	Tl La Lp B1	Vp Tl	Tl B1 La	La Bd Tl
t	Td Tl	Td Tl B1 Lp	Td Tl Vp B1	Td Vt
d	Tl Td B1	Vp Tl	Tl Td Vp B1	Td Tl
k	Vp Bd	Vp Tl	Vp Tl	Vp Bd
g	Bd Tl B1	Vp Tl	Vp Tl	Vp B1 Vt Bd
s	Vp Td Tl Vt	Vp Tl	Vp Td	Vp Td Bd Lp
v	Vp Tl La Lp	Vp Tl	Tl Vp	Vp Tl La Bd
r	Td Tl	Td Tl	Vp Tl	Lp Td Vt Bd

1D Articulator: Lips: La aperture Lp protrusion; Tongue tip constriction: Td distance Tl location; Tongue body constriction: Bd distance B1 location; Velar: Vp pharyngeal port Vt orovelar port.

Table 5. Critical articulators for the different phones and speakers. Each tract variable is considered as an articulator (2D analysis). The order of the different articulators, for each phone, reflects their importance. The two rightmost columns present the determined critical articulators gathering the normalized data for all speakers (spk All) and a characterization of EP sounds based on the principle of Articulatory Phonology as found in Oliveira [7]. For the sake of space economy, in the tract variable listing the T and C were omitted, for example, TBC became B.

ph	spk 8458	spk 8460	spk 8545	spk All	CO [7]
e	B	B L V	L	L	B
a	T	L	V L	L V T B	B
e	B T	B L T V	B L	B L	B
ε	B T L	B T L V	B L	B V L	B
i	B T	B T L	B T	B V	B
o	V L T	B L T	B L	L B	B L
ɔ	B T	B L T V	B L	B V T	B L
u	B L T V	L B T	L B T	L B	B L
ē ^{start}	T L V	B T L	T B V	B T	—
ē ^{mid}	T L B	L B T	B T L V	T	—
ē	T L V	B L V	B T L V	B T	B V
ē ^{start}	L	V	L B T	L	—
ē ^{mid}	T L	V	V L	T	—
ē	B V L	B L V T	V T L	L B	B V
ī ^{start}	B T V	B L T	B T L	B	—
ī ^{mid}	B T V	B L V T	B T L	B V	—
ī	B T V	B T V L	B T L V	B V	B V
ō ^{start}	B L V	B T L	L B	B T L	—
ō ^{mid}	L B V	B T L	B L	B T L	—
ō	B L V	B L V	B L	B T L	B L V
ū ^{start}	B V L T	V T B L	B L T	T V	—
ū ^{mid}	B V L	V L T B	L B T	L	—
ū	B V L	B L V T	L V	L	B L V
m	L	L V	L	L V T	L V
n	T B V	T B V L	T V	T V L	T V
ŋ	V	V	V	V L	V
p	L	L T	L T	L V	L V
b	L	L B T	B V L	L V	L V
t	T	T V L	V T B	T V B	T V
d	T	T	V T	T	T V
k	V B L T	B T L	B	B	B V
g	B V T	B T L	B L	V	B V
s	V T B L	B T L	T L B	V	T B V
v	V L	L T B	V L T	V	L V
r	T	T V	T	T V	T

Constrictions at: L: Lips T: Tongue Tip B: Tongue Body V: Velum.

6. Discussion

When comparing our preliminary work [25] and the results presented here, several aspects are worth noting. First, a novel speaker was considered (8545, in the third column of Tables 4 and 5) and the obtained results are consistent with those for the previously analysed speakers; second, the larger number of data samples considered for speaker 8460, entailing a larger number of samples per phone and including more phonetic contexts, turned some of the results more consistent with those of speaker 8458, as previously hypothesized [25]; and third, the consideration of one additional speaker for the normalized speaker, did not disrupt the overall previous findings for the critical variable (articulator) analysis.

Concerning the 1D correlation, among the different variable dimensions (see Figure 6), the variables are, overall, more decorrelated than in previous approaches considering landmarks over the vocal tract (e.g., see Silva et al. [24]) and has been further improved by the novel representation for the velar data considered in this work. The larger amount of data, in comparison with our first testing of the tract variable aligned with Articulatory Phonology [25], resulted in an even smaller number of correlations. Speaker 8545, along with the normalized speaker, do not show any correlation worth noting.

The mild/weak correlation observed for the lips (protrusion vs aperture) and tongue tip constriction (location vs degree) are, probably, due to a bias introduced by the characteristics of the considered corpus. Regarding the tongue tip, mild correlations between TTCI and TTCd may appear due to the fact the the strongest constrictions happen, typically, at the highest location angle.

The correlations observed, in our previous work (refer to Figure 7), for speaker 8460, between the lips and the tongue body and tongue tip have disappeared with the larger number of data samples considered, as hypothesized [25].

8458	LIPSa	LIPSp	TTCd	TTCI	TBCd	TBCI	Vy	Vx
LIPSa	1.00	0.59	0.26	0.28	0.00	0.00	0.00	0.00
LIPSp	0.59	1.00	0.36	0.00	0.00	0.00	0.00	0.00
TTCd	0.26	0.36	1.00	0.37	0.00	0.37	0.00	0.00
TTCI	0.28	0.00	0.37	1.00	0.29	0.27	0.30	0.34
TBCd	0.00	0.00	0.00	0.29	1.00	0.58	0.31	0.23
TBCI	0.00	0.00	0.37	0.27	0.58	1.00	0.00	0.00
Vy	0.00	0.00	0.00	0.30	0.31	0.00	1.00	0.88
Vx	0.00	0.00	0.00	0.34	0.23	0.00	0.88	1.00

8460	LIPSa	LIPSp	TTCd	TTCI	TBCd	TBCI	Vy	Vx
LIPSa	1.00	0.69	0.43	0.00	0.23	0.40	0.00	0.00
LIPSp	0.69	1.00	0.64	0.00	0.00	0.61	0.00	0.00
TTCd	0.43	0.64	1.00	0.00	0.23	0.61	0.00	0.00
TTCI	0.00	0.00	0.00	1.00	0.39	0.35	0.00	0.00
TBCd	0.23	0.00	0.23	0.39	1.00	0.45	0.00	0.00
TBCI	0.40	0.61	0.61	0.35	0.45	1.00	0.00	0.00
Vy	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.97
Vx	0.00	0.00	0.00	0.00	0.00	0.00	0.97	1.00

Figure 7. Correlation matrices for previous results [25] considering Articulatory Phonology aligned tract variables for two of the speakers also considered in this work (8458 and 8460). In this previous work, we considered less data samples per speaker and represented the velum by the *x* and *y* coordinates of a landmark positioned at its back. Please refer to Figure 6 for the corresponding matrices obtained in the current work. **Tract variables for 1D correlation (previous work):** LIPSa: lip aperture; LIPSp: lip protrusion; TTCd: tongue tip constriction distance; TTCI: tongue tip constriction location; TBCd: tongue body constriction distance; TBCI: tongue body constriction location; Vx: velar landmark *x*; Vy: velar landmark *y*.

6.1. Individual Tract Variable Components

The analysis of critical articulators treating each tract variable dimension as an independent variable is much more prone to being affected by the amount of data samples considered [19]. Therefore, while a few interesting results can be observed for some phones and speakers, some notable trends are not phonologically meaningful, such as the tongue tip constriction location (Tl) appearing prominently for the nasal vowels. Because the normalized speaker considers more data, some improvements are expected here when compared to the individual speakers and, indeed, it shows several promising results. Therefore, our discussion will mostly concern the normalized data. At a first glance, the tongue body (Bl and Bd) appears as critical in a prominent position for many of the vowels, as expected. The lip aperture (La) appears as critical for all bilabials segments (/p/, /b/, and /m/). The tongue tip constriction degree (Td) appears for the alveolars /n/, /t/ (with Vt) and /d/, the latter also with Tl, which seems to assert tighter conditions for the tongue tip positioning for the /d/. The velopharyngeal passage (Vp) appears as critical for the velar sounds /k/ (with Bd), /g/ (with Bl, Vt, and Bd), probably because some reajustments in the soft palate region preceding the velum. Also labiodental /v/ (with La) and for ɱ, which makes sense, since later concerns the nasal tail.

Concerning the lips, it is solely lip aperture (La) that appears as critical for /u/ and its nasal congenere and lip protrusion (Lp) appears across several of the vowels. This might be a similar effect to what we have previously observed for the velum: an articulator may appear as critical for those cases when it will be in a more fixed position during the articulation of a sound. The velum, for

instance, tends to appear more prominent for oral vowels since, at the middle of their production, it is closed, while, at the end of a nasal vowel, it can be open to different extents. Therefore, Lp may appear as critical not because the sound entails protrusion, but because the amount of observed protrusion throughout the different occurrences does not vary much.

Given the restricted number of speakers and occurrences, one aspect that seems interesting and should foster further future analysis, is the appearance of the orovelar (Vt) and not the velopharyngeal (Vp) passage as critical for nasal vowels. This does not diminish the velum opening, but points out that the extent of the orovelar passage is more stable across occurrences and, hence, more critical. Additionally, it is also relevant to note that for /*ü*/ and its oral congenere, the tongue body constriction does not appear as critical as happens, for example, with /*õ*/. Since the velopharyngeal passage and the tongue body constriction do not appear as critical—only the orovelar passage—this may hint that any variation of velar aperture, across occurrences, is compensated with tongue adjustments to keep the oral passage [42,43]. Also of note is the absence of Vt for the more fronted vowel /*ĩ*/ and its oral congenere. Given the fronted position of the tongue, Vt is large and more variable since its variation is not as limited, as for the back vowels, by velar opening.

One example that shows a different behavior between the tongue and velum is /*g*/, where both Vp and Vt are determined as critical and coincide with Bl and Bd, hinting that both the velum and tongue body are in a very fixed position along the occurrences of /*g*/. A similar result can also notably be observed for /*k*/.

Overall, Tl is still widely present (as with the individual speakers), mostly not agreeing with current phonological descriptions for EP and should motivate further analysis considering more data (speakers and phonetic contexts) and different alternatives for the computation of the tongue tip constriction.

6.2. Critical Tract Variables

Overall, and considering that the corpus is prone to strong coarticulation effects, the obtained results strongly follow our preliminary results presented in Silva et al. [25] and are mostly in accordance with previous descriptions considering Articulatory Phonology [7].

The TB is determined as the most critical articulator, for most vowels, in accordance to the descriptions available in the literature. The appearance of V, as critical articulator for some oral vowels, earlier than for nasals, is aligned with previous outcomes of the method [19,23,24]. This is probably due to a more stable position of V at the middle of oral vowels (the selected frame) than at the different stages selected for the nasal vowels for which it appears, mostly, in the fourth place, eventually due to the adopted conservative stopping criteria, to avoid phones without any reported critical articulator. It is also relevant to note that, for instance, if some of the nasal vowels are preceded by a nasal consonant it affects velum position during the initial phase of the vowel, which will have an incomplete movement towards closure [44]. This might explain why V does not appear as critical in the first frame (*start*) of some nasal vowels (typically referred as the oral stage [45]) since the velum is not in a stable position. The lips correctly appear, with some prominence for the back rounded vowels /*u*/ and /*o*/ and their nasal congeneres, but the appearance of this articulator for unrounded low vowels, probably due to the limitations of the corpus, does not allow any conclusion for this articulator.

Regarding consonants, for /*d*/, /*t*/, /*s*/ and /*r*/, as expected, T is identified as the most critical articulator, although, for /*s*/, it disappears in the normalized speaker. For bilabials, /*p*/, /*b*/ and /*m*/ correctly present L as the most critical articulator, and this is also observed for /*v*/, along with the expected prominence of V, except for speaker 8460. For /*m*/, V also appears, along with L, as expected. For /*p*/, the tongue tip appears as critical, for two of the speakers, probably due to coarticulatory reasons, but disappears in the normalized speaker which exhibits L and V, as expected. For /*k*/, V and TB are identified as the most critical articulators. Finally, *ŋ*, which denotes the nasal tail, makes sense to have V as critical. The appearance of L, in the normalized speaker is unexpected, since it does not appear for any of the other speakers.

By gathering the normalized data, for the three speakers, in speaker ALL, the method provided lists of critical articulators that are, overall, more succinct, cleaner, and closer to the expected outcomes, when compared to the literature [7], even considering a simple normalization method.

This seems to point out that the amount of considered data has a relevant impact on the outcomes. While this is expectable, the amount of data seems to have a stronger effect than in previous approaches using more variables [24], probably due to the fewer number of dimensions representing the configuration for each phone.

These good results, obtained with a very simple normalization approach, gathering the data for three speakers, may hint on how the elected tract variables are not strongly prone to the influence of articulator shape differences, among speakers, as was the case when we considered landmarks over the tongue. Instead, they depict the outcomes of the relation between parts of the vocal tract, for example, the tongue and hard palate (constriction). Nevertheless, some cases where the normalized speaker failed to follow the trend observed for the individual speakers, as alluded above, for example, for η , hint on the need to further improve the data normalization method.

7. Conclusions

Continuing the quest for data-driven methods to enable the determination of critical gestures for EP, this paper adopts a vocal tract configuration description aligned with Articulatory Phonology and, considering tract data obtained from midsagittal RT-MRI, presents the analysis of tract variable criticality for EP sounds. Overall, taking into consideration that the corpus was not specifically designed for the analysis of articulator criticality, since, for instance, some EP sounds and contexts are not present, the obtained results are already very interesting.

Following on the results presented here and on the experience gathered throughout, there are several aspects that elicit further attention. First of all, during the preparation of the methods considered in this work for the computation of the tract variables, some informal experiments revealed that slight variations, for example, for how lip protrusion is computed (such as, the leftmost point of both lips versus the middle point of maximum lip constriction) can result in slight variations in how the lips appear for some sounds and speakers. This would entail small improvements, for some phones and/or speakers and worst for others. While we kept the method considering the leftmost point of both lips, to enable a direct comparison with previous work [25], and since it already presents good results, this aspect is worth a more systematic exploration.

As already mentioned, another aspect that requires further research is the method used to perform speaker data normalization so it can be considered for the normalized speaker. Since some critical articulator results aligned with known descriptions of EP sounds and observed for all speakers disappeared in the normalized speaker, it is paramount to gather further understanding regarding the causes and improve the method accordingly.

One aspect that can improve the results and how they are interpreted is to have a full report regarding the considered contexts, for each sound and speaker. This would enable a clearer idea about speaker idiosyncrasies versus those aspects influenced by different amounts of data for a particular context, for example, among speakers. A part of this effort was already put in place, for the work presented here, for example, to obtain all available contexts for each phone, but a more automated analysis and report of these aspects is needed.

Finally, the method adopted here pertains a static analysis of critical articulators, i.e, it is based on the selection of a representative frame for each sound (e.g., the middle frame for oral vowels). For the nasal vowels, we split them in three key stages to understand if any notable differences arise, among them, but it seems relevant that this is further explored, for all sounds, to achieve an analysis of criticality over time. In this context, the audio signal, which has only been considered for annotating the data, so far, can be an important asset to explore a multimodal approach to these matters [46].

Author Contributions: Conceptualization, S.S. and A.T.; methodology, S.S. and A.T.; software, S.S. and N.A.; validation, S.S., N.A. and C.C.; formal analysis, S.S.; investigation, S.S., C.C., N.A., A.T., J.F. and A.J.; resources, A.T., C.C., S.S., J.F. and A.J.; data curation, S.S., C.C. and N.A.; writing—original draft preparation, S.S., A.T. and C.C.; writing—review and editing, A.T., S.S. and C.C.; visualization, S.S. and N.A.; supervision, A.T. and S.S.; project administration, S.S. and A.T.; funding acquisition, A.T., S.S. and C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research is partially funded by the German Federal Ministry of Education and Research (BMBF, with the project ‘Synchonic variability and change in European Portuguese’, 01UL1712X), by IEETA Research Unit funding (UID/CEC/00127/2019), 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).

Acknowledgments: We thank all the participants for their time and voice and Philip Hoole for the scripts for noise suppression.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Marin, S.; Pouplier, M. Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling model. *Mot. Control* **2010**, *14*, 380–407. [[CrossRef](#)] [[PubMed](#)]
2. Cunha, C. Portuguese lexical clusters and CVC sequences in speech perception and production. *Phonetica* **2015**, *72*, 138–161. [[CrossRef](#)] [[PubMed](#)]
3. Cunha, C. Die Organisation von Konsonantenclustern und CVC-Sequenzen in zwei portugiesischen Varietäten. Ph.D. Thesis, Ludwig Maximilian University of Munich, München, Germany, July 2012.
4. Xu, A.; Birkholz, P.; Xu, Y. Coarticulation as synchronized dimension-specific sequential target approximation: An articulatory synthesis simulation. In Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia, 5–9 August 2019.
5. Alexander, R.; Sorensen, T.; Toutios, A.; Narayanan, S. A modular architecture for articulatory synthesis from gestural specification. *J. Acoust. Soc. Am.* **2019**, *146*, 4458–4471. [[CrossRef](#)] [[PubMed](#)]
6. Silva, S.; Teixeira, A.; Orvalho, V. Articulatory-based Audiovisual Speech Synthesis: Proof of Concept for European Portuguese. In Proceedings of the Iberspeech 2016, Lisbon, Portugal, 23–25 November 2016; pp. 119–126.
7. Oliveira, C. From Grapheme to Gesture. Linguistic Contributions for an Articulatory Based Text-To-Speech System. Ph.D. Thesis, University of Aveiro, Aveiro, Portugal, 2009.
8. Martins, P.; Oliveira, C.; Silva, S.; Teixeira, A. Velar movement in European Portuguese nasal vowels. In Proceedings of the Iberspeech 2012, Madrid, Spain, 21–13 November 2012; pp. 231–240.
9. Scott, A.D.; Wylezinska, M.; Birch, M.J.; Miquel, M.E. Speech MRI: Morphology and function. *Phys. Med.* **2014**, *30*, 604–618. [[CrossRef](#)] [[PubMed](#)]
10. Lingala, S.G.; Sutton, B.P.; Miquel, M.E.; Nayak, K.S. Recommendations for real-time speech MRI. *J. Magn. Reson. Imaging* **2016**, *43*, 28–44. [[CrossRef](#)]
11. Silva, S.; Teixeira, A. Unsupervised Segmentation of the Vocal Tract from Real-Time MRI Sequences. *Comput. Speech Lang.* **2015**, *33*, 25–46. [[CrossRef](#)]
12. Labrunie, M.; Badin, P.; Voit, D.; Joseph, A.A.; Frahm, J.; Lamalle, L.; Vilain, C.; Boë, L.J. Automatic segmentation of speech articulators from real-time midsagittal MRI based on supervised learning. *Speech Commun.* **2018**, *99*, 27–46. [[CrossRef](#)]
13. Lammert, A.C.; Proctor, M.I.; Narayanan, S.S. Data-driven analysis of realtime vocal tract MRI using correlated image regions. In Proceedings of the Eleventh Annual Conference of the International Speech Communication Association, Chiba, Japan, 26–30 September 2010.
14. Chao, Q. Data-Driven Approaches to Articulatory Speech Processing. Ph.D. Thesis, University of California, Merced, CA, USA, May 2011.
15. Black, M.P.; Bone, D.; Skordilis, Z.I.; Gupta, R.; Xia, W.; Papadopoulos, P.; Chakravarthula, S.N.; Xiao, B.; Segbroeck, V.M.; Kim, J.; et al. Automated evaluation of non-native English pronunciation quality: Combining knowledge-and data-driven features at multiple time scales. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015.

16. Silva, S.; Teixeira, A. Quantitative systematic analysis of vocal tract data. *Comput. Speech Lang.* **2016**, *36*, 307–329. [[CrossRef](#)]
17. Kim, J.; Toutios, A.; Lee, S.; Narayanan, S.S. A kinematic study of critical and non-critical articulators in emotional speech production. *J. Acoust. Soc. Am.* **2015**, *137*, 1411–1429. [[CrossRef](#)]
18. Sepulveda, A.; Castellanos-Domínguez, G.; Guido, R.C. Time-frequency relevant features for critical articulators movement inference. In Proceedings of the 20th European Signal Processing Conference (EUSIPCO), Bucharest, Romania, 27–31 August 2012.
19. Jackson, P.J.; Singampalli, V.D. Statistical identification of articulation constraints in the production of speech. *Speech Commun.* **2009**, *51*, 695–710. [[CrossRef](#)]
20. Ananthkrishnan, G.; Engwall, O. Important regions in the articulator trajectory. In Proceedings of the 8th International Seminar on Speech Production (ISSP'08), Strasbourg, France, 8–12 December 2008; pp. 305–308.
21. Ramanarayanan, V.; Segbroeck, M.V.; Narayanan, S.S. Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. *Comput. Speech Lang.* **2016**, *36*, 330–346. [[CrossRef](#)] [[PubMed](#)]
22. Prasad, A.; Ghosh, P.K. Information theoretic optimal vocal tract region selection from real time magnetic resonance images for broad phonetic class recognition. *Comput. Speech Lang.* **2016**, *39*, 108–128. [[CrossRef](#)]
23. Silva, S.; Teixeira, A.J. Critical Articulators Identification from RT-MRI of the Vocal Tract. In Proceedings of the INTERSPEECH, Stockholm, Sweden, 20–24 August 2017.
24. Silva, S.; Teixeira, A.; Cunha, C.; Almeida, N.; Joseph, A.A.; Frahm, J. Exploring Critical Articulator Identification from 50Hz RT-MRI Data of the Vocal Tract. In Proceedings of the INTERSPEECH, Graz, Austria, 15–19 September 2019. [[CrossRef](#)]
25. Silva, S.; Cunha, C.; Teixeira, A.; Joseph, A.; Frahm, J. Towards Automatic Determination of Critical Gestures for European Portuguese Sounds. In Proceedings of the International Conference on Computational Processing of the Portuguese Language, Vora, Portugal, 2–4 March 2020.
26. Jackson, P.J.; Singampalli, V.D. Statistical identification of critical, dependent and redundant articulators. *J. Acoust. Soc. Am.* **2008**, *123*, 3321. [[CrossRef](#)]
27. Goldstein, L.; Byrd, D.; Saltzman, E. The role of vocal tract gestural action units in understanding the evolution of phonology. In *Action to Language via Mirror Neuron System*; Cambridge University Press: Cambridge, UK, 2006; pp. 215–249.
28. Browman, C.P.; Goldstein, L. Gestural specification using dynamically-defined articulatory structures. *J. Phon.* **1990**, *18*, 299–320. [[CrossRef](#)]
29. Proctor, M. Towards a gestural characterization of liquids: Evidence from Spanish and Russian. *Lab. Phonol.* **2011**, *2*, 451–485. [[CrossRef](#)]
30. Recasens, D. A cross-language acoustic study of initial and final allophones of/l. *Speech Commun.* **2012**, *54*, 368–383. [[CrossRef](#)]
31. Teixeira, A.; Oliveira, C.; Barbosa, P. European Portuguese articulatory based text-to-speech: First results. In Proceedings of the International Conference on Computational Processing of the Portuguese Language, Aveiro, Portugal, 8–10 September 2008.
32. Saltzman, E.L.; Munhall, K.G. A dynamical approach to gestural patterning in speech production. *Ecol. Psychol.* **1989**, *1*, 333–382. [[CrossRef](#)]
33. Nam, H.; Goldstein, L.; Saltzman, E.; Byrd, D. TADA: An enhanced, portable Task Dynamics model in MATLAB. *J. Acoust. Soc. Am.* **2004**, *115*, 2430. [[CrossRef](#)]
34. Oliveira, C.; Teixeira, A. On gestures timing in European Portuguese nasals. In Proceedings of the ICPhS, Saarbrücken, Germany, 6–10 August 2007, pp. 405–408.
35. Almeida, N.; Silva, S.; Teixeira, A.; Cunha, C. Collaborative Quantitative Analysis of RT-MRI. In Proceedings of the 12th International Seminar on Speech Production (ISSP), Providence, RI, USA, 14–18 December 2020.
36. Uecker, M.; Zhang, S.; Voit, D.; Karaus, A.; Merboldt, K.D.; Frahm, J. Real-time MRI at a resolution of 20 ms. *NMR Biomed.* **2010**, *23*, 986–994. [[CrossRef](#)]
37. Browman, C.P.; Goldstein, L. Some notes on syllable structure in articulatory phonology. *Phonetica* **1988**, *45*, 140–155. [[CrossRef](#)]
38. Feng, G.; Castelli, E. Some acoustic features of nasal and nasalized vowels: A target for vowel nasalization. *J. Acoust. Soc. Am.* **1996**, *99*, 3694–3706. [[CrossRef](#)] [[PubMed](#)]

39. Teixeira, A.; Vaz, F. European Portuguese nasal vowels: An EMMA study. In Proceedings of the Seventh European Conference on Speech Communication and Technology, Aalborg, Denmark, 3–7 September 2001.
40. Rao, M.; Seth, S.; Xu, J.; Chen, Y.; Tagare, H.; Príncipe, J.C. A test of independence based on a generalized correlation function. *Signal Process.* **2011**, *91*, 15–27. [[CrossRef](#)]
41. Johnson, R.A.; Wichern, D.W. *Applied Multivariate Statistical Analysis*; McGraw-Hill: New York, NY, USA, 2007.
42. Cunha, C.; Silva, S.; Teixeira, A.; Oliveira, C.; Martins, P.; Joseph, A.A.; Frahm, J. On the Role of Oral Configurations in European Portuguese Nasal Vowels. In Proceedings of the INTERSPEECH 2019, Graz, Austria, 15–19 September 2019. [[CrossRef](#)]
43. Carignan, C. Covariation of nasalization, tongue height, and breathiness in the realization of F1 of Southern French nasal vowels. *J. Phon.* **2017**, *63*, 87–105. [[CrossRef](#)]
44. Teixeira, A.; Vaz, F.; Príncipe, J.C. Nasal Vowels After Nasal Consonants. In Proceedings of the 5th Seminar on Speech Production: Models and Data, Bavaria, Germany, 1–4 May 2000.
45. Parkinson, S. Portuguese nasal vowels as phonological diphthongs. *Lingua* **1983**, *61*, 157–177. [[CrossRef](#)]
46. Mitra, V.; Nam, H.; Espy-Wilson, C.Y.; Saltzman, E.; Goldstein, L. Retrieving tract variables from acoustics: A comparison of different machine learning strategies. *IEEE J. Sel. Top. Signal Process.* **2010**, *4*, 1027–1045. [[CrossRef](#)]

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).