*INTERSPEECH 2019*
September 15–19, 2019, Graz, Austria

# Exploring Critical Articulator Identification from 50Hz RT-MRI Data of the Vocal Tract

*Samuel Silva*[1], *António Teixeira*[1], *Conceição Cunha*[2], *Nuno Almeida*[1], *Arun Joseph*[3], *Jens Frahm*[3]
[1]DETI/IEETA, University of Aveiro, Portugal
[2]IPS, LMU Munich, Germany
[3]Max-Planck-Institut für Biophysikalische Chemie, Germany
sss@ua.pt, ajst@ua.pt

## Abstract

The study of the static and dynamic aspects of speech production can profit from technologies such as electromagnetic midsagittal articulography (EMA) and real-time magnetic resonance (RTMRI). These can improve our knowledge on which articulators and gestures are involved in producing specific sounds and foster improved speech production models, paramount to advance, e.g., articulatory speech synthesis. Previous work, by the authors, has shown that critical articulator identification could be performed from RTMRI data of the vocal tract, with encouraging results, by extending the applicability of an unsupervised statistical identification method previously proposed for EMA data. Nevertheless, the slower time resolution of the considered RT-MRI corpus (14 Hz), when compared to EMA, potentially influencing the ability to select the most suitable representative configuration for each phone — paramount for strongly dynamic phones, e.g., nasal vowels —, and the lack of a richer set of contexts – relevant for observing coarticulation effects –, were identified as limitations. This article addresses these limitations by exploring critical articulator identification from a faster RTMRI corpus (50 Hz), for European Portuguese, providing a richer set of contexts, and testing how fusing the articulatory data of two speakers might influence critical articulator determination.

**Index Terms**: critical articulators, speech production model, real-time magnetic resonance

## 1. Introduction

The development and improvement of speech production models fosters improvements in speech technologies, such as speech synthesis [1], and can, in turn, serve to test new theories and further increase our understanding of speech production [2]. One of the main aspects posing challenges is the study of coarticulation, to understand how different speech organs interact with each other. This is particularly important to improve articulatory speech synthesis [3] or audiovisual synthesis [4], in which lip and tongue movement need to abide by specific timings to attain realism.

Regarding coarticulation, Articulatory Phonology [5, 6] proposes that, for each phone, there are three types of articulators: (1) those that are critical, resisting to context and having a coarticulatory effect on neighbour phones; (2) those that depend on the critical articulators due to an anatomic link; and (3) those that are redundant and suffer no particular constraint. For instance, producing /p/ necessarily involves lip closure, but the tongue is free to move. In consequence, the lips are critical articulators and the tongue is redundant. For alveolar sounds, as /t, d/, the tongue tip is the main articulator, but the tongue dorsum is anatomically linked and is responsible for a second movement in /r, l/.

A variety of technologies can provide data for static and dynamic studies of speech production (e.g., real-time magnetic resonance, RTMRI [7], and electromagnetic articulography, EMA), supporting the study of the relevance (criticality) and timings of each articulator for attaining specific linguistic goals [8, 9]. However, the acquisition of this data require access to expensive devices, the processing is complex, and a posteriori labeling is very time consuming. Therefore, most of the works analyse only a very reduced set of speakers. The need for a systematic quantitative assessment advises tackling these matters through data-driven approaches, preferably unsupervised, to avoid the time consuming annotation, errors, and inconsistencies associated with manual correction. In this regard, the community has made an effort to contribute with data-driven approaches to extract and analyse the features of interest [10, 11, 12, 13, 14].

On the specific subject of articulator criticality, a few authors have proposed data-driven methods, e.g., [9, 15, 16, 17, 18]. In a previous work [19], the authors have shown that critical articulator identification could be performed from RTMRI data of the vocal tract by extending the applicability of a method proposed for EMA data by Jackson et al. [20]. The results encouraged further exploration of the method and several aspects were identified as possible limitations and deemed relevant for improvement, in future studies: (1) the reduced size of the corpus and its phonetic representativeness; (2) a strong bias towards oral and nasal vowels, the corpus original purpose; (3) the reduced number of contexts; and (4) a low time resolution (14Hz), when compared to EMA (filtered to 100Hz), possibly entailing the selection of a representative frame, for each phone, which is not the most adequate (e.g., not the highest curvature of the tongue blade, for /l/) due to a lack of enough time resolution. Additionally, the method was applied to each speaker separately, but its application to the full data, at once, might help to more clearly identify critical articulators disentangled from specific speaker characteristics.

In this article, we follow up on previous work, further exploring the potential of the critical articulator determination method. We innovate by considering a new RT-MRI corpus for European Portuguese, tackling some of the limitations enunciated above, namely by providing a larger sample size, increased number of contexts, and higher time resolution (50Hz), and by performing critical articulator determination by fusing normalized data for multiple speakers.

The remainder of this article is organized as follows: section 2 provides a presentation of the main aspects of the adopted methods, namely describing the considered corpus, and the selected data and considered tract variables/landmarks. Then, section 3 presents the main results for the determined critical articulators considering two speakers of European Portuguese and these are discussed in section 4. Finally, section 5 presents the

Copyright © 2019 ISCA

http://dx.doi.org/10.21437/Interspeech.2019-2897

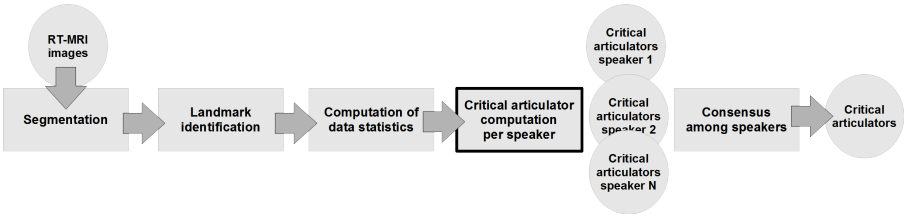


Figure 1: *Main steps contributing to the statistical identification of critical articulators from RT-MRI.*

main conclusions and ideas for further evolutions.

## 2. Methods

The method previously adopted for determining critical articulators from RT-MRI data [19], inherited from the method proposed by Jackson et al. [20], which considers vocal tract landmarks (mimicking the position of the EMA pellets), as representative of the articulators, selects landmark samples, at the midpoint of each phone, and uses the selected data to compute several statistics concerning: (1) the whole landmark data (the grand statistics), used to build the models for each landmark (articulator); and (2) the data for each phone (phone statistics). Critical articulator identification is then obtained by analysing the distances between the grand and phone probability distributions. Figure 1 depicts the main stages required to perform this analysis for an RTMRI corpus, as described in what follows.

### 2.1. RT-MRI Corpus and Acquisition

The analysed materials consisted of one syllable words starting with labiodental fricative or two syllable words starting either with bilabial stop or nasal followed by all stressed oral and nasal vowels as well as nasal diphthongs and the oral counterparts. The target words were embedded in one of three carrier sentences alternating the verb as follows (Diga 'Say'—ouvi 'I heard'— leio 'I read') as in 'Diga **pote**, diga **pote** baixinho' ('Say **pot**, Say **pot** gently'). All sentences were read from a computer screen and presented in randomized order. So far, this corpus has been recorded from sixteen native speakers (8m, 8f) of EP.

To deploy all the methods for this new RT-MRI corpus and confirm their applicability (since aspects such as segmentation variability might have an impact on the outcomes), this article explores the data for two of the male speakers (8458 and 8460);

RT-MRI recordings were conducted at the Max Planck Institute for biophysical Chemistry, Göttingen, Germany, using a 3 Tesla Siemens Magnetom Prisma Fit MRI System equipped with high performance gradients (Max ampl=80 mT/m; slew rate = 200 T/m/s). A standard 64-channel head coil was used with a mirror mounted on top of the coil. Real-time MRI measurements were based on a under-sampling method, in which radial FLASH acquisitions are combined with nonlinear inverse reconstruction (NLINV) providing images at high spatial and temporal resolutions [21]. Further advancements in this technique allow acquisitions at 50 fps.

Synchronized audio was 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. All volunteers provided informed written consent and filled an MRI screening form in agreement with institutional rules, prior recordings. They were compensated for their participation and none of them reported language, speech or hearing problems.

The segmentation of the vocal tract outlines was performed considering the method described in Silva et al. [13], followed by manual revision to detect and manually correct any major segmentation issue, e.g., at the lips, for bilabials.

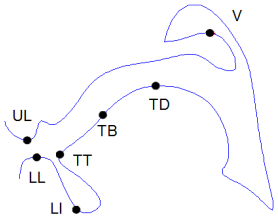


Figure 2: *Illustration of the landmark points selected over the vocal tract outline. These landmarks, considering the $x$ and $y$ coordinates for 1D analysis, or joining them, for 2D analysis, were used for critical articulator determination.*

### 2.2. Landmark Positioning

As in our previous work, we chose landmark positioning in accordance to what is considered for the EMA data, by Jackson et al. [20], i.e., an approximation to the EMA pellet positions. Figure 2 illustrates the location chosen for each landmark, as representative of each articulator.

The positioning of each landmark was determined by an unsupervised method, following a set of predetermined criteria, from the segmented vocal tract outlines, for both speakers. For the upper and lower lips (UL and LL), we consider the highest and lowest point, respectively, of the corresponding lip. For the points located on the tongue surface, besides the tongue tip (TT), the tongue blade (TB) and tongue dorsum (TD) landmarks were placed at fixed distances from TT, measured along the tongue outline, analogous to what happens with the EMA pellets. Regarding the velum landmark (V), since the velum region is prone to exhibit image artefacts, potentially entailing a high degree of variability, in the segmentation, we opted for placing the velum landmark on the interior soft palate wall.

To have a landmark providing data correlated with jaw rotation, and since the teeth are not visible in RT-MRI, we considered the region of the vocal tract contour located were the base of the teeth should be (lower incisor, LI).

### 2.3. Articulatory Data Selection

For the selection of the representative frame, for each phone, an automated selection method was applied by considering the different characteristics of each sound, as described in Table 1.

While in [20], Jackson et al. considered the midpoint for all sounds, in [22] they suggest, for alveolar obstruents an improved criterion for sample selection could improve the results,

thus hinting on the relevance of pursuing such approach.

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

| **phone (SAMPA)** | **criterion** |
|---|---|
| **oral vowels**<br>6, a, e, E, i, o, O, u | midpoint |
| **nasal vowels**<br>6˜, e˜, i˜, o˜, u˜ | for each, three classes were created, taking the first, middle, and final frames |
| **nasal consonants**<br>m, n | [m], frame with minimum inter-lip distance; [n], midpoint |
| **stops**<br>p, k, t, b, | [p] and [b], frame with minimum inter-lip distance; [k] and [t], midpoint |

Finally, since the literature [23, 24, 25, 26] shows evidence for a dynamic structure of the nasal vowels, with different stages, we were also interested in studying if any difference would arise when computing the critical articulators, at different timepoints, along the vowel. Therefore, each nasal vowel was included as three "pseudo-phones", represented by the beginning, middle and final frame of the annotated interval and named, respectively, [vowel]_B, [vowel]_M and [vowel]_F.

### 2.4. Computation of Data Statistics

The pivotal step for the critical articulator determination is the computation of the statistics: the grand statistics, characterizing the distribution of positions, for each landmark, along the whole data; and the phone statistics, representing the distribution of positions of each landmark, for each phone, considering the data selection as per the criteria in Table 1. Table 2 summarizes the different statistics that need to be computed to initialize the method, following the notation of Jackson et al. [20]. Critical articulator identification was performed taking landmark coordinates ($x$ and $y$) independently – the 1D case – for example ULx for the $x$ coordinate of the upper lip, or combining them – the 2D case.

The 1D correlation matrices for the landmarks (e.g., considering $TB_x$ and $TT_y$, etc.), given the size of our data set, was computed considering correntropy, as proposed in Rao et al. [27]. Bivariate correlations (i.e, taking both coordinates of each landmark together) were computed through canonical correlation analysis [28, 20]. For the grand correlation matrices, adopting the criteria proposed in [20], only statistically significant ($\alpha = 0.05$) correlation values above 0.2 were kept, reducing the remaining ones to zero.

### 2.5. Identification of Critical Articulators

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. Considering the variability observed between speakers, we adopted two approaches to obtain a consensus: (1) our previous approach [19], weighting each articulator based on its position on the list, for each phone and speaker. For instance, an articulator in the first place weights 7 and, in the second place, 6. Adding the weights for each articulator, from all speakers, for each phone, the consensus is the list of articulators reaching a total weight above 10 (maximum of 14). Additionally, we wanted to assess how the method would work by gathering the data for both speakers in a "normalized" speaker. To that effect,

Table 2: *Summary of computed statistics for each landmark and corresponding notation as in [20].*

| **Grand statistics** | Notation | Comment |
|---|---|---|
| grand mean | $M$ | all selected frames |
| grand variance | $\Sigma$ | all selected frames |
| total sample size | $N$ | Spk 8458: 1492; Spk 8460: 785 |
| corr. matrix | $R^*$ | keeping statistically significant and strong correlations ($r_{ij} > 0.2$ and $\alpha = 0.05$) |
| **Phone statistics** | Notation | Comment |
| mean | $\mu^\phi$ | frames selected for each phone |
| variance | $\Sigma^\phi$ | frames selected for each phone |
| sample size | $v^\phi$ | variable among phones |
| corr. matrix | $R^\phi$ | without attending to significance and module |

Table 3: *Grand correlation matrices for speakers 8458 considering the $x$ and $y$ coordinates for each landmark separately (1D analysis).*

| 8458 | ULy | ULx | LLy | LLx | LIy | LIx | TTy | TBy | TDy | TTx | TBx | TDx | Vy | Vx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ULy | 1.00 | -0.50 | -0.28 | -0.26 | 0.00 | 0.00 | 0.36 | 0.29 | 0.00 | 0.21 | 0.21 | 0.00 | 0.00 | 0.00 |
| ULx | -0.50 | 1.00 | 0.53 | 0.43 | 0.00 | 0.38 | -0.50 | -0.54 | -0.43 | -0.47 | -0.47 | -0.43 | 0.00 | 0.00 |
| LLy | -0.28 | 0.53 | 1.00 | 0.57 | 0.63 | 0.43 | 0.00 | 0.00 | 0.00 | -0.25 | -0.27 | -0.25 | 0.00 | 0.00 |
| LLx | -0.26 | 0.43 | 0.57 | 1.00 | 0.38 | 0.48 | -0.24 | 0.00 | 0.00 | -0.31 | -0.31 | -0.28 | 0.25 | 0.00 |
| LIy | 0.00 | 0.00 | 0.63 | 0.38 | 1.00 | 0.41 | 0.20 | 0.23 | 0.20 | 0.00 | 0.00 | 0.00 | 0.29 | 0.00 |
| LIx | 0.00 | 0.38 | 0.43 | 0.48 | 0.41 | 1.00 | 0.00 | 0.00 | -0.24 | -0.33 | -0.39 | -0.39 | 0.61 | -0.57 |
| TTy | 0.36 | -0.50 | 0.00 | -0.24 | 0.20 | 0.00 | 1.00 | 0.56 | 0.44 | 0.74 | 0.63 | 0.56 | 0.00 | 0.00 |
| TBy | 0.29 | -0.54 | 0.00 | 0.00 | 0.23 | 0.00 | 0.56 | 1.00 | 0.92 | 0.47 | 0.65 | 0.62 | 0.00 | 0.00 |
| TDy | 0.00 | -0.43 | 0.00 | 0.00 | 0.20 | -0.24 | 0.44 | 0.92 | 1.00 | 0.53 | 0.73 | 0.75 | 0.00 | 0.00 |
| TTx | 0.21 | -0.47 | -0.25 | -0.31 | 0.00 | -0.33 | 0.74 | 0.47 | 0.53 | 1.00 | 0.93 | 0.90 | 0.00 | 0.00 |
| TBx | 0.21 | -0.47 | -0.27 | -0.31 | 0.00 | -0.39 | 0.63 | 0.65 | 0.73 | 0.93 | 1.00 | 0.99 | -0.27 | 0.27 |
| TDx | 0.00 | -0.43 | -0.25 | -0.28 | 0.00 | -0.39 | 0.56 | 0.62 | 0.75 | 0.90 | 0.99 | 1.00 | -0.26 | 0.26 |
| Vy | 0.00 | 0.00 | 0.00 | 0.25 | 0.29 | 0.61 | 0.00 | 0.00 | 0.00 | 0.00 | -0.27 | -0.26 | 1.00 | -0.88 |
| Vx | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | -0.57 | 0.00 | 0.00 | 0.00 | 0.00 | 0.27 | 0.26 | -0.88 | 1.00 |

we normalized the landmark data, for each speaker, based on the variation ranges, for each landmark, computed over the entire corpus, and considered this gathered data as a new speaker following a similar analysis methodology.

## 3. Results

Table 3 presents the correlation table for the 1D analysis for all articulators, for speaker 8458. The matrix for speaker 8460 shows a similar pattern and is not shown for the sake of space. A notable aspect, also observed in our previous work, for a different corpus, is the appearance of correlation "clusters", namely for the tongue (TT, TB, TD), the lips and the velum, although with a less clear distinction as the ones observed in the work of Jackson et al. [20] for EMA data. Differently from [20], but in agreement with our previous work, there is a correlation between the $x$ and $y$ coordinates of the tongue. A significant correlation is present between $TT_y$ and $TD_y$, for both speakers. Albeit small, its significance is probably due to the lack of additional occurrences of phones, such as /l/, in this corpus, which would more strongly evidence the independence between the tongue tip and tongue dorsum $y$ movements. These matrices and those for 2D supported the determination of the 1D and 2D critical articulators. For the sake of brevity, we will solely report and discuss the outcomes for 2D critical articulator determination. Table 4 presents the full list of critical articulators resulting from the 2D analysis for each of the speakers, for the "normalized" speaker – obtained by gathering the normalized data of both speakers –, and a consensus voting considering the

Table 4: *Critical articulator identification for a list of phones present in the analysed corpus, considering each speaker (columns 8458 and 8460), gathering the normalized data for both speakers (ALL), and by consensus (see text).*

| ph | spk 8458 | spk 8460 | spk ALL | consensus |
|---|---|---|---|---|
| 6 | TD V TB TT | V LL LI TT TB TD | V TD | V TD |
| a | TT TB TD LL LI | TT UL LL TB LI | TT | TT |
| e | TD TB LL V TT LI UL | TB TD LL TT LI UL | TD TB LL UL V | TD TB LL |
| E | TD TB V TT LI LL | LL TD TB V LI TT UL | TD TT | TD TT TB |
| i | TD TB TT V LL UL | LI TD TB V TT UL LL | V TD TB LI | TD TB V LI |
| o | V LL TT TD | V TD TB TT UL LI LL | TB TD TT | TD TT V TB |
| O | TB TD TT V LL UL | V TT TB TD UL LI LL | TT V | TT V TB |
| u | TD LL TB TT LI UL | V TB TD LI TT LL UL | TB | TB TD |
| 6~B | TB LL UL TT TD | TB TT TD LI UL LL V | TB | TB |
| 6~M | LI TT TB LL TD UL | TB TD LI LL UL TT | TB | TB LI |
| 6~F | TD LI TB LL UL TT V | TB TD LI LL UL TT | TB TT | TB TD LI |
| a~F | LI TT LL UL | TT UL LI TD | LI | LI TT |
| e~B | TB TD TT | V | TB | TB |
| e~M | TB TD UL LL | | TB | TB |
| e~F | TB TD UL LL | | TB | TB |
| i~B | TD TB TT LL UL V LI | TB TD LL LI TT V | TD TB V LI UL | TD TB |
| i~M | TD TB TT UL LL LI V | TB TD LL LI TT V | TD TB V LI UL | TD TB |
| i~F | TD TB TT LI UL V | TB TD LL LI TT V UL | TD TB V | TD TB |
| o~B | LL LI UL | V LL LI TT | LL TD TT TB UL | LL LI |
| o~M | LL TB UL TD LI | V LL TD TT TB UL | LL UL TB TD | LL TB UL TD |
| o~F | LL UL LI TD TB | TD LL TB | LL TD TT TB UL | LL TD TB |
| u~B | TD LL TB LI UL TT V | TT TB TD UL LL LI | TB | TB TD |
| u~M | TD LI TB UL LL TT | TT TB TD UL LI LL | TB LI | TB LI TD |
| u~F | TD LI TB UL LL TT | TB UL TD LI LL | TB LI | TB LI TD |
| d | TT UL TB | TB TT LI TD UL LL | TT TD TB | TT TB |
| g | TB TD LI UL TT V | LL V TD UL TB LI TT | TB TD V UL | TB TD V UL |
| p | LL TD TT TB UL LI | LL V UL TD | LL TD | LL TD |
| t | TT | TB TD LL TT | | TT |
| k | V TT TB LI TD LL UL | LI TB TD TT V LL UL | V TB | TB V LI |
| m | LL LI TB TD TT UL | LI LL TD | LL UL TD LI TB | LL LI TD |
| n | TB LI TD TT LL | V TD TB UL LI LL | TB TD | TB TD |

lists for both speakers. All computations considered a convergence threshold ($\theta_c$) of 1.7. While a higher value would yield shorter critical articulator lists, for each phone, we chose to keep the value used in our previous work, to enable comparison.

## 4. Discussion

As reported by Jackson et al. [20], and as observed in our previous work [19], the velum (V) appeared as critical for the oral vowels rather than their nasal congeneres. Coherently, for the later, V also appears as critical during the first stage – oral phase – of the nasal vowels (e.g., 6~_B, i~_B). This hints that the velum is in a well defined fixed position (closed) at the start of the vowel, but its position at the middle and at the end is not as definite, eventually as a result of context influencing the transition to the nasal phase. Additionally, consistent with Articulatory Phonology based descriptions of EP phones, available from Oliveira [29], and confirming the results reported by Jackson et al. [20], the method consistently identifies the tongue blade and tongue dorsum as critical articulators for vowels. The generally lower number of critical articulators identified for the vowels (Tab. 4: ALL and consensus), when compared with our previous work, might be a result of the broader set of contexts present in the corpus, enabling a clearer identification of what is truly critical. Regarding other phones, for [p], LL and UL appear, for both speakers, and V appears for speaker 8460. Similarly, LL and UL appear for [m], confirming its bilabial nature. Notably, [k] and [g] share several critical articulators with a prominence for V followed by TD/TB (see speaker ALL).

For the results obtained by gathering the normalized data of both speakers, some of the critical articulators that were observed for both speakers, e.g. UL for [p] or TT for [t], did not stand for the normalized speaker. This is potentially due to the simple normalization method used and motivates further work, in this regard.

## 5. Conclusions

Taken together, the work presented here and our earlier work to study critical articulator determination from RT-MRI, also profiting from our proposals in vocal tract outline segmentation and analysis [13, 14] establish promising grounds for more strongly investing in evolving several aspects of this method and its application to the novel RT-MRI corpus of 16 EP speakers. The adopted approach of gathering the data for both speakers, after normalization, to see beyond the variability between speakers, in order to grasp what might be truly critical, rather than speaker specific approaches, provided interesting results, sometimes similar to our previous consensus approach, but avoiding its empirical nature.

The way the different vocal tract configurations were defined, considering a set of landmarks mimicking the position of the flesh points for EMA (i.e., pellets positions) left room for further exploring how this might affect the determination of critical articulators. Since the method by Jackson et al. [20] is general enough to support any set of landmarks, the exploration of other track variables, e.g., constriction degree and location, seems an important next step and is, currently, under way.

## 6. Acknowledgements

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

# 7. References


[1] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PLoS ONE*, vol. 8, no. 4, pp. 1–17, 04 2013.

[2] H. Nam, V. Mitra, M. Tiede, E. Saltzman, L. Goldstein, C. Y. Espy-Wilson, and M. Hasegawa-Johnson, "A procedure for estimating gestural scores from natural speech." in *Proc. Interspeech*, Makuhari, Japan, 2010, pp. 30–33.

[3] A. Teixeira, L. Silva, R. Martinez, and F. Vaz, "SAPWindows – towards a versatile modular articulatory synthesizer," in *Proc. IEEE Workshop on Speech Synthesis*, Santa Monica, CA, USA, Sept 2002, pp. 31–34.

[4] S. Silva, A. Teixeira, and V. Orvalho, "Articulatory-based audiovisual speech synthesis: Proof of concept for European Portuguese," in *Proc. Iberspeech*, Lisbon, Portugal, 2016, pp. 119–126.

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

[6] N. Hall, "Articulatory phonology," *Language and Linguistics Compass*, vol. 4, no. 9, pp. 818–830, 2010.

[7] 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.

[8] L. Goldstein and M. Pouplier, "The temporal organization of speech," in *The Oxford Handbook of Language Production*, M. A. Goldrick, V. Ferreira, and M. Miozzo, Eds. Oxford University Press, 2014, pp. 210 – 227.

[9] J. Kim, A. Toutios, S. Lee, and S. S. Narayanan, "A kinematic study of critical and non-critical articulators in emotional speech production," *The Journal of the Acoustical Society of America*, vol. 137, no. 3, pp. 1411–1429, Mar 2015.

[10] 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.

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

[12] 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.

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

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

[15] A. Sepulveda, G. Castellanos-Domínguez, and R. C. Guido, "Time-frequency relevant features for critical articulators movement inference," in *Proc. 20th European Signal Processing Conference (EUSIPCO)*, Aug 2012, pp. 2802–2806.

[16] G. Ananthakrishnan and O. Engwall, "Important regions in the articulator trajectory," in *Proc. ISSP*, Strasbourg, France, 2008, pp. 305–308.

[17] V. Ramanarayanan, M. V. Segbroeck, and S. S. Narayanan, "Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories," *Computer Speech & Language*, vol. 36, pp. 330–346, 2016.

[18] A. Prasad and P. K. Ghosh, "Information theoretic optimal vocal tract region selection from real time magnetic resonance images for broad phonetic class recognition," *Computer Speech & Language*, vol. 39, pp. 108 – 128, 2016.

[19] S. Silva and A. J. Teixeira, "Critical articulators identification from rt-mri of the vocal tract." in *INTERSPEECH*, 2017, pp. 626–630.

[20] 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.

[21] 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.

[22] P. Jackson and V. Singampalli, "Coarticulatory constraints determined by automatic identification from articulograph data," in *Proc. ISSP*, Strasbourg, France, 2008, pp. 377–380.

[23] G. Feng and E. Castelli, "Some acoustic features of nasal and nasalized vowels: A target for vowel nasalization," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3694–3706, 1996.

[24] A. Teixeira and F. Vaz, "European portuguese nasal vowels: an EMMA study." in *Proc. Interspeech*, Aalborg, Denmark, 2001, pp. 1483–1486.

[25] C. Oliveira and A. Teixeira, "On gestures timing in European Portuguese nasals," in *Proc. ICPhS*, Saarbrücken, Germany, 2007.

[26] P. Martins, C. Oliveira, S. Silva, and A. Teixeira, "Velar movement in European Portuguese nasal vowels," in *Proc IberSpeech*, 2012, pp. 231–240.

[27] M. Rao, S. Seth, J. Xu, Y. Chen, H. Tagare, and J. C. Príncipe, "A test of independence based on a generalized correlation function," *Signal Processing*, vol. 91, no. 1, pp. 15–27, 2011.

[28] R. A. Johnson and D. W. Wichern, *Applied multivariate statistical analysis*, 6th ed. Pearson Prentice Hall, 2007.

[29] C. Oliveira, "From grapheme to gesture. linguistic contributions for an articulatory based text-to-speech system," Ph.D. dissertation, University of Aveiro, 2009.