INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Critical articulators identification from RT-MRI of the vocal tract

Samuel Silva[1], *António Teixeira*[1]

[1]DETI / IEETA, University of Aveiro (Aveiro, Portugal)

sss@ua.pt, ajst@ua.pt

Abstract

Several technologies, such as electromagnetic midsagittal articulography (EMA) or real-time magnetic resonance (RT-MRI), enable studying the static and dynamic aspects of speech production. The resulting knowledge can, in turn, inform the improvement of speech production models, e.g., for articulatory speech synthesis, by enabling the identification of which articulators and gestures are involved in producing specific sounds.

The amount of data available from these technologies, and the need for a systematic quantitative assessment, advise tackling these matters through data-driven approaches, preferably unsupervised, since annotated data is scarce. In this context, a method for statistical identification of critical articulators has been proposed, in the literature, and successfully applied to EMA data. However, the many differences regarding the data available from other technologies, such as RT-MRI, and language-specific aspects create a challenging setting for its direct and wider applicability.

In this article, we address the steps needed to extend the applicability of the proposed statistical analyses, initially applied to EMA, to an existing RT-MRI corpus and test it for a different language, European Portuguese. The obtained results, for three speakers, and considering 33 phonemes, provide phonologically meaningful critical articulator outcomes and show evidence of the applicability of the method to RT-MRI.

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

1. Introduction

The identification of which articulators and gestures are involved in producing specific sounds is important to inform the development and improvement of speech production models. Advancing these 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].

Our interest in these aspects is triggered by our research on speech production for European Portuguese (EP) [3, 4], and by our subsequent work applying speech production knowledge for articulatory speech synthesis [5] and audiovisual speech synthesis [6]. One of the main challenges posed by these contexts concerns coarticulation, important, for example, to attain realistic lip and tongue movement in audiovisual speech synthesis. Regarding coarticulation, Articulatory Phonology [7, 8] 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.

Several technologies, such as electromagnetic midsagittal articulography (EMA) or real-time magnetic resonance (RT-MRI) [9], provide data to study the static and dynamic aspects of speech production from which the relevance (criticality) and timings of each articulator for attaining specific linguistic goals can be inferred [10, 11]. However, the sheer amount of data made available by these technologies, and the need for a systematic quantitative assessment, advise tackling these matters through data-driven approaches, preferably unsupervised, since annotated data is scarce. This has motivated the community at large [12, 13, 14], and our team in particular [15, 16], to adopt quantitative approaches to extract and analyse relevant features from the available data. Among these approaches, a few authors have specifically addressed critical articulator identification through data-driven methods (e.g., [11, 17, 18, 19, 20]) and one of the methods that provides an interesting and robust approach is proposed by Jackson et al. [21] and applied to EMA data. The authors consider a large set of articulatory data from EMA to build statistical models for the movement of each articulator. Then, the data for particular phones is compared with the overall models and the critical articulators for each phone are identified.

In view of this method, we considered that its approach might also be suitable to determine critical articulators from RT-MRI speech production data, and specially relevant considering the importance of RT-MRI for speech studies [22] and its differentiating characteristics towards EMA, such as non-invasiveness and whole vocal tract view. However, the many differences regarding the kind of data available from EMA and RT-MRI and the potential influence of language-specific aspects create a challenging setting for the direct applicability of the method. Comparing both technologies, RT-MRI potentially presents several challenges that might hinder identical performance: (1) RT-MRI has a much lower sampling rate and far less phone repetitions; and (2) while, for EMA, flesh points (the pellets) are fixed, for the RT-MRI data we need to rely on unsupervised landmark selection that will necessarily encompass some variability.

The work presented here describes the steps devised to address these challenges and extend the applicability of the statistical analyses proposed by [21], for the determination of critical articulators – originally applied to EMA data –, to data extracted from midsagittal real-time MRI of the vocal tract. Overall, the obtained results, for three speakers, and considering 33 phonemes, provide phonologically meaningful critical articulator outcomes and show evidence of the applicability of the method to RT-MRI.

The remainder of this article is organized as follows: section 2 describes the main aspects of the considered methods; section 3 presents the results obtained for three speakers of European Portuguese and section 4 discusses them; finally, section 5 draws some conclusions and presents routes for future developments.

http://dx.doi.org/10.21437/Interspeech.2017-742

2. Methods

In brief, the method proposed by Jackson et al. [21] considers the data of the EMA landmarks, 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). Then, critical articulator identification is based on the distances between the grand and phone probability distribution functions. for each phone.

To use the same analysis method for RT-MRI data, we need to define which landmarks are considered, what data samples are selected, for each phone, and compute the grand and phone statistics used to initialize the method. A description of the main aspects concerning these steps is presented in what follows.

2.1. RT-MRI Corpus

Since, prior to a large new investment in corpora acquisition, it is recommended to assess the potential of method applicability, an existing corpus was selected. The considered RT-MRI corpus [4] was primarily acquired for the study of European Portuguese oral and nasal vowels and included a few sequences covering other sounds such as nasal consonants, fricatives, stops, and laterals. The first column of Table 1 presents a list of the classes and phones included in the corpus.

Acquisition, at 14 images/s, was performed at the midsagittal plane of the vocal tract following the protocols described in [4]. Data was acquired for three female speakers (SV, CM, CO), aged between 21 and 33, phonetically trained, with no history of hearing or speech disorders.

Audio was recorded synchronously with the RT-MRI images, inside the MR scanner, at a sampling rate of 16 kHz, using a fiberoptic microphone, and manually annotated using the software tool Praat[1].

The segmentation of the vocal tract outlines was performed based on active appearance models, as described in [15], followed by manual review to detect and manually correct any major segmentation issue, mainly at the lips, for bilabials, and at the tongue tip for laterals and nasal consonants.

2.2. Landmark Identification

The method by Jackson et al. [21] is general enough to support any set of landmarks, but we chose, at this stage, to replicate, as best as possible, the overall position of the flesh points (i.e., the pellet position) considered for the EMA data used in [21]. Figure 1 illustrates the location chosen for each landmark, as representative of each articulator.

We defined criteria for positioning each landmark and then applied them, unsupervised, to the whole database. 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), each of the additional landmarks (tongue blade, TB, and tongue dorsum, TD) were located at fixed distances from TT, measured along the tongue outline and kept constant throughout the sequences. Finally, the velum landmark (V), since the velum region is prone to exhibit image artefacts and placing the landmark at velum tip or along the soft palate surface, although possible, is prone to a lot of variability, we opted for placing the velum landmark on the interior soft palate wall.

[1] http://www.fon.hum.uva.nl/praat/

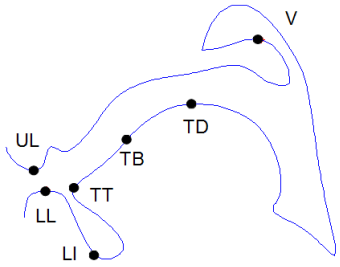

Figure 1: *Landmarks depicted over a sample vocal tract outline for speaker SV.*

Table 1: *Summary of phones in the corpus and criteria used for selecting the representative frame.*

phone (SAMPA)	criterion
oral vowels 6, a, e, E, i, o, O, u	midpoint
nasal vowels 6˜, e˜, i˜, o˜, u˜	for each, three classes were created, taking the first, middle, and final frames
nasal consonants m, n, J	[m], frame with minimum inter-lip distance; [n] and [J], midpoint
laterals l	frame with highest tongue blade curvature [24]
stops p, k, t, b,	[p] and [b], frame with minimum inter-lip distance; [k] and [t], midpoint
taps 4	midpoint
fricatives f, s	[f], frame with minimum inter-lip distance; [s], midpoint

For the lower incisor (LI, as representative of jaw rotation), since the teeth are not visible in RT-MRI, we settled for an approximation by choosing the point, from the segmented contour, that should appear in the required region.

Note that critical articulator identification was performed in two different ways: (1) taking landmark coordinates (x and y) independently, the 1D case, for example ULx for the x coordinate of the upper lip ; or (2) combining them, the 2D case.

2.3. Articulatory Data Selection

For the selection of the representative time frame (sample), for each phone, from the annotated interval, an automated selection method was applied and, differently from Jackson et al. [21], we did not always select samples at phone midpoints. Instead, we paid attention to particularities of each sound, as described in Table 1. In fact, Jackson et al. [23] suggest, for alveolar obstruents, that an improved criterion for sample selection could improve results.

Additionally, for nasal vowels, considering that the literature [25, 3, 26, 27] shows evidence that nasal vowels have different stages, we were interested in assessing if any difference would arise when computing the critical articulators at different times. Therefore, each nasal vowel was taken as three new "pseudo-phones", represented by the first, middle and last frame of the annotated interval and named, respectively, [vowel]_I, [vowel]_M and [vowel]_F.

Table 2: *Summary of computed statistics for each landmark and corresponding notation as in [21].*

Grand statistics	Notation	Comment
grand mean	M	all selected frames
grand variance	Σ	all selected frames
total sample size	N	CO: 629; SV: 786; CM: 655
corr. matrix	R^*	keeping statistically significant and strong correlations ($r_{ij} > 0.2$ and $\alpha = 0.05$)
Phone statistics	Notation	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	without attending to significance and module

2.4. Computation of Data Statistics

Table 2 summarizes the different statistics that need to be computed to initialize the method, following the same notation as in Jackson et al. [21]. All values mentioned were computed for each of the landmarks. The grand statistics were derived from all the samples collected for each of the phones following the criteria described in Table 1. As for the phone statistics, all occurrences of a particular phone were considered.

Landmark correlation matrices were computed for the grand and phone statistics. Regarding 1D correlation (i.e., considering the x and y coordinates separately), and given the size of our data set, we hypothesized that using a different method for its computation might improve the results. In this context, we considered the computation of correntropy, as proposed in Rao et al. [28], but no relevant differences were found and normal correlation was kept. Bivariate correlations for the 2D articulatory data (i.e, taking both coordinates of each landmark together) were computed through canonical correlation analysis [29]. For the grand correlation matrices, only statistically significant ($\alpha = 0.05$) correlation values above 0.2 were kept, reducing the remaining ones to zero, as in [21].

2.5. Critical Articulator Identification

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 some variability observed for the critical articulators identified for each speaker's data, we were interested in getting an overall idea of the results, for discussion. As a first approximation, we weighted 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 a list of articulators reaching a total weight above 10 (maximum of 21).

3. Results

Table 3 depicts the correlation matrices for the 1D analysis for speaker SV. This matrix, a representative example of the other speakers' matrices, presents some resemblance to those presented by Jackson et al. although showing less well defined articulator groups, particularly when it comes to the lips and jaw groups. As observed by [21], there is no correlation between TTy and TDy, showing independence of the vertical movement

Table 3: *Grand correlation matrix for speaker SV considering the coordinate for each landmark separately (1D analysis).*

SV	ULy	ULx	LLy	LLx	LIy	LIx	TTy	TBy	TDy	TTx	TBx	TDx	Vy	Vx
ULy	1.00	0.29	0.00	0.00	0.00	0.00	0.34	0.53	0.00	0.54	0.65	0.52	0.22	0.33
ULx	0.29	1.00	0.00	0.40	0.00	0.00	0.00	0.28	0.34	0.00	0.00	0.23	0.25	0.26
LLy	0.00	0.00	1.00	0.56	0.23	0.47	0.00	0.33	0.35	0.00	0.00	0.00	0.00	0.00
LLx	0.00	0.40	0.56	1.00	0.00	0.61	0.00	0.46	0.00	0.22	0.36	0.23	0.00	0.00
LIy	0.00	0.00	0.23	0.00	1.00	0.53	0.00	0.00	0.24	0.00	0.00	0.00	-0.23	0.00
LIx	0.00	0.00	0.47	0.61	0.53	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TTy	0.34	0.00	0.00	0.00	0.00	0.00	1.00	0.41	0.00	0.67	0.36	0.00	-0.26	0.00
TBy	0.53	0.28	0.33	0.46	0.00	0.00	0.41	1.00	0.42	0.35	0.79	0.66	0.27	0.26
TDy	0.00	0.34	0.35	0.00	0.24	0.00	0.00	0.42	1.00	-0.24	0.00	0.42	0.22	0.00
TTx	0.54	0.00	0.00	0.22	0.00	0.00	0.67	0.35	-0.24	1.00	0.72	0.58	0.00	0.00
TBx	0.65	0.00	0.00	0.36	0.00	0.00	0.36	0.79	0.00	0.72	1.00	0.91	0.26	0.31
TDx	0.52	0.23	0.00	0.23	0.00	0.00	0.00	0.66	0.42	0.58	0.91	1.00	0.35	0.38
Vy	0.22	0.25	0.00	0.00	-0.23	0.00	-0.26	0.27	0.22	0.00	0.26	0.35	1.00	0.90
Vx	0.33	0.26	0.00	0.00	0.00	0.00	0.00	0.26	0.00	0.00	0.31	0.38	0.90	1.00

between the tongue tip and dorsum and there are strong correlations for the tongue in x and weaker in y. Differences in our matrices, when compared to Jackson et al.'s, which are worth mentioning, concern the correlation between ULy and the tongue, the correlation among the x and y coordinates of the tongue, and a (albeit mild) correlation between the velum and TBx and TDx.

The determination of 1D critical articulators, and in line with what is shown in the correlation matrices, yielded lists of critical articulators that we deemed less interesting to discuss, at this stage. For that reason, and for the sake of space we will mostly concentrate on the presentation and discussion of the outcomes for the 2D analysis.

For the 2D analysis, the results for the critical articulators identified for each phone are shown in Table 4, per speaker. The three first columns present the full lists of critical articulators by considering a convergence threshold (θ_c) of 1.7 to all speakers.

Overall, and despite differences among speakers in the position occupied by each articulator, the number of critical articulators identified for each phone is similar. The consensus column shows the weighted (among all speakers) critical articulator list and is a simple first approximation to serve the depiction of the overall behaviour.

4. Discussion

Despite the notable differences of the RT-MRI data towards EMA (corpus size, phonetic richness, sample rate), the results obtained show the method as capable to extract critical articulators for phones of EP from RT-MRI.

Regarding the 1D analysis, the obtained results might be more affected than the 2D results by the size and nature of the corpus and one cannot rule out that landmark positioning might require further attention to improve this aspect.

Regarding 2D analysis, to assess the meaningfulness of the results, the list of critical articulators identified for each phone (from the consensus column in Table 4) was compared to Articulatory Phonology based descriptions of EP phones, available from Oliveira [30] (shown in the rightmost column of Table 4). Note that these phonetic descriptions consider different variables, namely tongue body (TBd) and lip aperture (LA) and protrusion (LP). For the sake of simplicity, in our discussion we take our TB and TD as encompassed in TBd and UL+LL as providing information on LA and LP.

Overall, the outcomes of the method are in good agreement with Oliveira's proposal [30]. For vowels, there is a strong presence, in the identified critical articulators, of TB and TD, hence, tongue body (TBd), in accordance with the phonetic descrip-

Table 4: *List of critical articulators resulting from 2D analysis, for each speaker, their weighted consensus, and the phonetic description for EP phonemes taken from Oliveira [30].*

ph	spk SV	spk CM	spk CO	consensus	Oliveira [30]
6	V TD TB	TD TB	TD TB LL TT	TD TB	TBd
a	LL	LL V	TT TD UL	LL	TBd
E	TB V TD TT UL LI LL	TB TD UL TT V LL	V TD TB TT LL	TB TD V TT	TBd
i	TT TB TD UL	TT TB TD UL	TT TB TD LL UL LI	TT TB TD UL	TBd
o	TD TB LI LL UL	TD UL TB LL TT LI V	TB TD UL LL TT V LI	TD TB UL LL	TBd LP LA
O	V TB TD TT UL LL	UL TT LL LI TB TD	V TB TD TT LI LL UL	TB TT V TD UL	TBd LP LA
u	UL	UL LL	UL TD LL	UL LL	TBd LP LA
6̃ I	TD TB TT LL LI UL	TB TT TD UL	TT TD TB LL LI	TT TB TD	TBd
6̃ M	TD TB V LI TT UL LL	TT TB UL TD LI V	TT TB TD V LL LI	TB TT TD V	TBd V
6̃ F	TD TB TT V LI LL	TB TT TD UL LI	TT TB TD LL LI V	TB TT TD	TBd V
ẽ I	TB TD TT UL LI V	TB TD TT UL	TB TT TD UL LL	TB TD TT UL	TBd
ẽ M	TD TB TT UL V LL	TB TD TT V UL LI	TD TB TT LL V LI UL	TD TB TT	TBd V
ẽ F	TB TD TT V UL LI	TB TT TD UL LL LI V	TD TB TT LL V UL LI	TB TD TT	TBd V
ĩ I	TB TD TT UL V	TT TB TD	TT TB TD LL UL	TT TB TD	TBd
ĩ M	TB TD V TT UL LI	TT TB TD V UL LI	TB TD TT V UL LL LI	TB TD TT V	TBd V
ĩ F	TB V TD TT UL LI	TT TB TD V LI	TT TD TB V	TT TB TD V	TBd V
õ I	TB TT UL LL V LI TD	TB TT TD LI UL LL	TD TT TB LL LI UL	TB TT TD	TBd LP LA
õ M	TB V UL LL TD TT LI	TT TB LL UL LI TD	TD UL TT TB LI	TB UL TT TD	TBd LP LA V
õ F	TB TT V UL LL LI TD	TB TT LL TD UL LI	TT UL LL TB TD LI	TT TB UL LL	TBd LP LA V
ũ B	UL TB TT TD LI V LL	TB TT TD UL LL LI	TT TB TD LL UL	TB TT UL TD	TBd LP LA
ũ M	TB TD UL TT LI V	TT TB TD LI UL LL	TT LL TB TD UL LI	TT TB TD UL	TBd LP LA V
ũ F	TB TD TT UL LI V	TB TT TD UL LL LI	TB TT TD LL V UL	TB TT TD	TBd LP LA V
m	V	LI LL V	V LL UL LI	V LL LI	LA V
n	TT TB LL UL	V UL TB LI TT TD	TT LI V TB LL UL	TT TB UL V	TT V
J	TB TT TD V LL UL	TB TT TD V LI UL LL	TB V TT TD LL UL LI	TB TT TD V	TBd V
l	TT V	TT UL V LI	TT TB V TD UL LI LL	TT V	TT TBd
p	V	UL	UL V	UL V	LA V
k	V TT TB UL LL LI	TT UL TB TD LL LI V	TB TT V UL LI	TT TB UL V	TBd V
t	TB V TT UL	UL TD LL LI TB V TT	TB TT UL LI TD V LL	TB UL TT	TT V
b	TT TB V TD UL LI	TB TT TD LL V UL	V TB LL TD LI TT UL	TB TT V TD	LA V
4	V TB TT LL UL TD LI	LL TD TT UL TB LI V	UL V TT LL	LL TT UL V	TT
s	V TT	UL TT LL TB TD LI	TT LI LL TD TB UL	TT	TT TBd V
f	TD LL LI V TB UL	LL TT UL LI TD TB	V LL UL TT TB	LL UL V	LA LP V

tions, except for [a] and [u]. For posterior vowels ([o], [O] and [u]) the lips are correctly identified as critical. The TT appears for front vowels ([E] and [i]) as in [21]. For nasal vowels, differences between initial and medial frames appear several times, but the velum is not often identified in all medial and final configurations (or in opposition to the initial frames), as we would expect, although it appears often for speaker SV.

For the nasal consonants, TB and UL appear differently from the description, for [n], and TT and TD for [J]. For the lateral [l], V appears as extra, which might make sense, and the method fails the detection of additional tongue articulators, possibly explained by a varying degree of velarization. For stops, TB and UL appear extra for [t], TT for [k] and TB, TT and TD for [b], possibly due to the reduced number of contexts; for [t] it failed to detect V and for [b], UL. As reported in Jackson et al. [21] only one of the lips (UL) was chosen for [p]. For the tap, the method captured the intervention of TT, while LL, UL and V appear as extra. Since only one context was considered (the corpus only included 'caro' [ka4u]), some of the extras might be a consequence of that (LL and UL). Nevertheless, this phoneme lacks a detailed characterization and the role of V cannot be safely disregarded. Additionally, taps are complex and highly dynamic sounds, for which additional time resolution would be helpful. For the fricative [s], the method failed the detection of V and TB/TD and, for [f], the method successfully identified all of the expected critical articulators.

Considering the mixed results for the criticality of the velum, these can arise from issues concerning the definition of the corresponding landmark (possibly hinted by the 1D correlations with TDx and TBx), but since the role of the velum for EP sounds and, in general, has deserved insufficient attention, the theory leaves room for reasonable doubt and this should motivate further enquiries on this matter.

Several limitations affect the results such as the size of the considered corpus, its bias towards oral and nasal vowels and the reduced number of contexts for the consonants. The lower frequency of the RT-MRI (14 Hz) data, when compared to EMA (filtered to 100Hz) might also entail the selection of a frame that is not the ideal (e.g, not the highest curvature of TB for [l]).

5. Conclusions

In this paper, a statistical method for identification of critical articulators, previously proposed and tested for EMA data, was applied to a small RT-MRI corpus to characterize a set of 33 EP phones. Despite the small size of the considered RT-MRI database, and the challenges related to the selection of landmarks, the obtained results are quite interesting and phonologically meaningful to encourage further pursuing this approach.

Considering the results reported here, additional steps should consider improvements to landmark definition and testing with larger corpora.

6. Acknowledgements

Samuel Silva is funded by grant SFRH/BPD/108151/2015 from FCT. Research partially funded by IEETA Research Unit funding (UID/CEC/00127/2013.) and Marie Curie Actions IRIS (ref. 610986, FP7-PEOPLE-2013-IAPP).

7. References

[1] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PLoS ONE*, vol. 8, no. 4, pp. 1–17, 04 2013. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0060603

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

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

[5] 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.

[6] 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.

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

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

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

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.

[22] A. Toutios and S. S. Narayanan, "Advances in real-time magnetic resonance imaging of the vocal tract for speech science and technology research," *APSIPA Transactions on Signal and Information Processing*, vol. 5, p. e6, 2016.

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

[24] C. Smith, "Complex tongue shaping in lateral liquid production without constriction-based goals," in *Proc. ISSP*, Cologne, Germany, 2014, pp. 413–416.

[25] 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.

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

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

[28] 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.

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

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