CAN STATIC VOCAL TRACT POSITIONS REPRESENT ARTICULATORY TARGETS IN CONTINUOUS SPEECH? MATCHING STATIC MRI CAPTURES AGAINST REAL-TIME MRI FOR THE FRENCH LANGUAGE

Anastasiia Tsukanova[1], Ioannis K. Douros[1,2], Anastasia Shimorina[1], Yves Laprie[1]

[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France,

[2]Université de Lorraine, INSERM U1254, IADI, F-54000 Nancy, France,
anastasiia.tsukanova@inria.fr, ioannis.douros@loria.fr, anastasia.shimorina@loria.fr, yves.laprie@loria.fr

ABSTRACT

This paper uses mediosagittal slices of a static magnetic resonance imaging (MRI) dataset capturing the blocked articulation of vowels and of consonants that anticipate /a, i, u, y/ and a variety of other vowels to study the presence and distinctness of these deliberately taken articulatory targets in real-time MRI recordings. The study investigates whether such articulatory targets are actually attained in fluent speech, how marked they are, and what factors influence the degree of similarity between a given articulatory target and the actual vocal tract shape. To quantify the similarity, we use structural similarity, Wasserstein distance, and SIFT measure. We analyze the amplitude and timing of the observed similarity peaks across different phonetic classes and speech types (spontaneous versus not). We show that although real-time speech involves shapes quite similar to the static data, there is a great intra- and inter- speaker variability.

Keywords: articulatory targets, coarticulation, speech production

1. INTRODUCTION

Speech is produced through a complicated process that involves high cognitive skills including semantic and syntactic language operation, both automatic and consciously refined motor control, and perceptual feedback processing. Articulatory evidence can be a window into different facets of all of these sub-tasks. Thus, it is of utmost importance to further enhance our understanding of speech production.

One of the central issues in the mechanics of speech production is the notion of a smallest unit or some other elementary component. From the auditory perspective, it is the phoneme: the smallest semantically distinguishing unit. From the articulatory point of view, however, each and every sound we utter is a result of the filter that is formed by the vocal tract, whose shape is a compound effect of coordinating the articulatory movements. Given that each articulator seemingly follows its own timing and has considerable degrees of freedom in space, it is easy to see how studying the organization of speech movements is no trivial task.

A major branch of modeling these motions consists in decomposing speech into articulatory gestures (articulatory phonology [2], motor primitives [8, 13]). However, when dealing with overlapping gestures, we face the disadvantage that, at least as of now, the ground truth is not available; every muscle contraction and every bend at a joint can be integrated into various gesture elements, and it is up to the model only to stay consistent about determining where the boundaries of those gestures are.

From this standpoint, it is interesting to consider an alternative view on what guides speech phenomena: targets. They can be found in a virtual task space (task dynamics [16, 17]), or they can literally be specific positions of the vocal tract [4]. It is this line of thought that the present work follows, chosen due to the benefits of the methodological clarity and high applicability, e.g. in articulatory speech synthesis [6, 18, 19], which in turn will be capable of serving as further evidence for the underpinnings of speech production [11].

To investigate whether it is possible to use some kind of static articulatory representations as points of reference for the dynamics of speech, we naturally need articulatory data. Generally speaking, this choice is a tradeoff between the frequency in time of the acquired information and its spatial richness, with such techniques as electromagnetic articulography (EMA) falling at one end of the spectrum, and magnetic resonance imaging (MRI) at the other. Fortunately, recent technological advances have boosted real-time MRI (RT-MRI) to such a level that it allows us to alleviate the gravity of the choice between the two. Currently, RT-MRI is argued to be one of the most promising sources of articulatory information for speech production research [3, 12].

[4] employed targets as the vocal tract configurations attained at the middle of the duration of each phoneme, which is when it is most stable. However, that study recognized the need to allow for contextually modified targets to capture coarticulation. [1] worked in that direction and differentiated targets according to their vocalic context, following the ideas of [10]; the static MRI dataset that was put in use in this study is quite similar to that of [1], only composed for French.

The aim of the paper is to look for such static, frozen articulatory targets in RT-MRI data and give an interpretation of their presence or absence. The objective is to employ measures that are proven to be efficient in computer vision to compare the static and dynamic MRI datasets, and to draw conclusions from the dynamics and distributions of these metrics.

2. DATASETS AND METHODS

2.1. MRI corpora

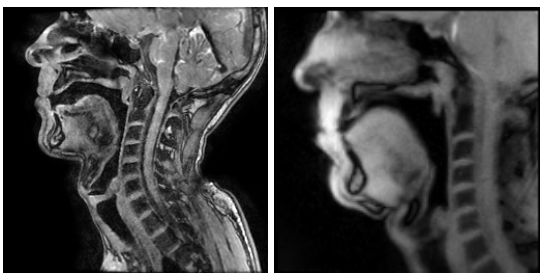
The static MRI dataset consisted of 86 mid-sagittal images collected in a 3D mode with a GE Signa 3T machine with an 8-channel neurovascular coil array. The acquisition used a custom modified Enhanced Fast Gradient Echo (EFGRE3D, TR 3.12 ms, TE 1.08 ms, matrix 256x256x76, spatial resolution 1.02x1.02x1.0 mm^3). The speaker was to show the position that he would have to attain to produce a particular sound. For vowels, that is the position when the vowel would be at its clearest if the subject were phonating. For consonant-vowel (CV) syllables, that is the blocked configuration of the vocal tract, as if the subject were about to start pronouncing it before a coming vowel V. There were 13 vowels, 71 CV syllables and 2 semi-vowels in the final dataset. This covers all main phonemes of the French language, but not in all contexts. Each consonant was recorded in the context of at least the three cardinal vowels /a, i, u/ and /y/, which is strongly protruded in French; some more vocalic contexts were included as well. The speaker of this dataset will be referred to as speaker A.

The RT-MRI data consisted of 9 real-time 1 min-long acquisitions of both speakers A and B (one recording of spontaneous speech per speaker, the rest not), following the protocol described in [9] (the sampling frequency 55 Hz, images of 189x189 pixels representing the midsagittal section that is 8 mm wide), including both the articulatory information (RT-MRI itself) and speech.

The task for non-spontaneous speech was to read out sentences that came from a phonetically balanced corpus as well as syllables that imitated the static MRI corpus (for example, where the static corpus treated the positions for /p(i)/, /p(a)/, /p(y)/, the dynamic corpus would require the speaker to produce "*pis, pas, pu*" which are the dynamic implementation of these configurations). Fig. 1 shows the static and dynamic examples of /p(i)/.

The task for spontaneous speech involved a spontaneous, unrehearsed answer to the question "What do you think of the healthcare system in France".

Figure 1: Static (left) and dynamic (right) recordings of /p(i)/ articulation by speaker A.



The speech was denoised to reduce the noise of the MRI machine that is dominant in the recording and then manually transcribed. The text of the transcription was subsequently treated by eLite HTS tool [14] to retrieve phonetic labels, and those were force aligned with HTK [22], using Merlin as frontend [21].

2.2. Comparing static and dynamic images

When matching the images of these two datasets, one has to face several issues:

(1) The resolution and quality of the images is not the same: 256x256 pixels against 189x189. Furthermore, MRI is very sensitive to movement, resulting in a considerable amount of blurring in the dynamic images.

(2) The images do not depict exactly the same areas of the subjects' vocal tracts, nor do the subjects take exactly same positions. Moreover, three years passed between acquiring the static and dynamic datasets, resulting in some minor physical changes in speaker A; and naturally, there are differences between speakers A and B.

(3) Static acquisitions may produce shapes that will never be observed in dynamic data since they involve no phonation and there are phonemes whose sustained imitation of articulation is either difficult or impossible (liquids due to their dynamic nature [5]; stops, whose burst is a result of pressure building up in the vocal tract; it is difficult to control nasality).

(4) The static dataset is rather small and should not be expected to cover all the images in the dynamic dataset.

(5) While being larger, the dynamic dataset still remains relatively small as far as speech resources go. When breaking down into specific contexts, phoneme classes, syntactic structures, or speaking styles, data sparsity quickly becomes an issue.

We chose to stay as rigorous in our approach as possible and to have a coherent measure between each of the static images and each of the dynamic images. We cut out the rectangular of the vocal tract and resized the resulting images to 84x82 pixels. Then three metrics known to perform well in image processing and computer vision feature extraction were used: structural similarity (SSIM) [20], Earth mover's distance (EMD) [15] (Wasserstein distance on the histograms of the images), and scale-invariant feature transform (SIFT) [7]. Figure 2 shows the behavior of these metrics on one of the recordings.

EMD measures the difference between two probability distributions, calculating the work it would take to transform one of them into the other. When applied to pixel intensity histograms, it produces a measure of image similarity. If $f_{i,j}$ is the optical flow between clusters p_i and q_j and $d_{i,j}$ is the ground distance between them,

$$(1) \quad EMD(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j} d_{i,j}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j}}$$

Lower values of EMD mean more similar images.

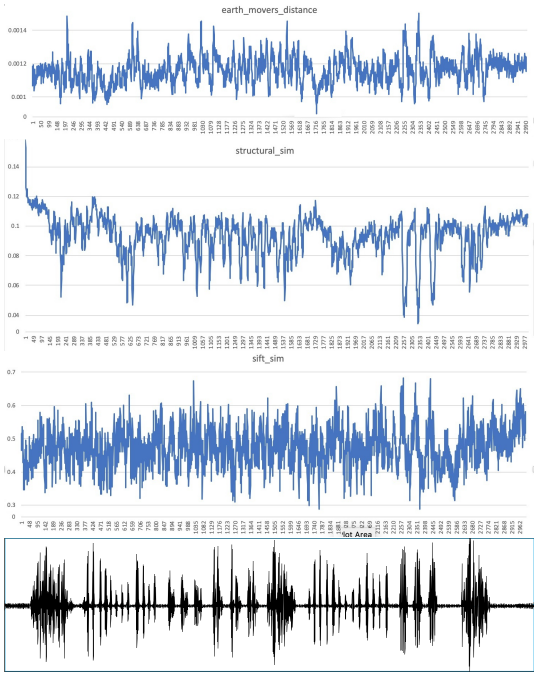
SSIM is a measure that originally quantified perceived image degradation when given an original image and its compressed version, but can be used to quantify similarity between any two images. It is calculated on windows of the image. SSIM between two windows x and y is a ratio that depends on the windows' averages, variances and the covariance. Its values range from -1 to 1, 1 standing for identical images.

SIFT is a feature detection algorithm. It matches keypoints that agree on the object and its location, scale, and orientation. We perform a ratio test that rejects all feature matches whose distance ratio is greater than 0.7:

$$(2)\quad SIFT(x,y) = \frac{|matches(x,y) : match.dist < 70|}{|matches(x,y)|}$$

Since it is a proportion of all matches, its values range from 0 to 1, and the greater the value, the closer two images are.

Figure 2: The distance between the static image /f(i)/ and all the dynamic images in one of the one-minute long sequences by speaker A, aligned with the power spectrum of the recording (lowermost): 1) Earth mover's distance (values from 0 to 1, the lower, the more similar); 2) structural similarity (from -1 to 1, the greater, the more similar); 3) SIFT (from 0 to 1, the greater, the more similar).



It should be noted that while articulation is a very smooth process, where nothing happens in jerks, the behavior of these metrics is not. Therefore we find it imprudent to look for patterns one image at a time, for example, at the image in the center of a given phoneme. Instead, we resolve ourselves to use averaging and look for general patterns.

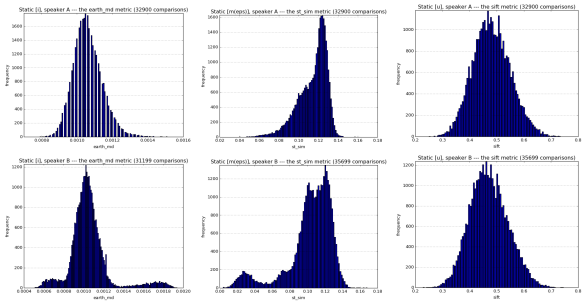
3. EXPERIMENTS

The calculated distances were aggregated by speakers, by phonemes, by phonemes in vocalic context (what vowel V is anticipated in the dynamic dataset according to the phonetic labeling) and by speaker styles (spontaneous and not).

Table 1 shows the mean and standard deviation of each metric across the entire volume of speech by speakers A and B. Overall there is no linear relationship between the metrics: correlation $r(SIFT,EMD) = -0.009$ for A, 0.096 for B; $r(SSIM,EMD) = -0.742$ for A, -0.511 for B; $r(SSIM,SIFT) = -0.131$ for A, -0.303 for B.

None of the metrics followed the normal distribution

Figure 3: The overall distribution of the EMD metric for the distance between the static capture /i/ and all the dynamic captures for speakers A (up) and B (down) (leftmost pair); structural similarity, /mɛ/ (middle pair); SIFT, /u/ (rightmost pair).



(EMD: Shapiro-Wilk's test (ShW): statistic 0.822, p-value 0.000, D'Agostino's K^2 test (DA): 2429273.281, p-value 0.000; SIFT: ShW: statistic 0.967, p-value 0.000, DA: 4736.912, p-value 0.000; SSIM: ShW: statistic 0.853, p-value 0.000, DA: 1635032.574, p-value 0.000), thus, in order to determine whether the observed differences between distances are statistically significant, we needed to use non-parametric tests.

Table 1: Means (E) and standard deviations (SD) of the image similarity metrics, speakers A and B

	$E(EMD) \pm SD(EMD)$	$E(SSIM) \pm SD(SSIM)$	$E(SIFT) \pm SD(SIFT)$
A	0.001 ± 0.00009	0.113 ± 0.017	0.429 ± 0.0766
B	0.001 ± 0.00019	0.099 ± 0.024	0.416 ± 0.0717

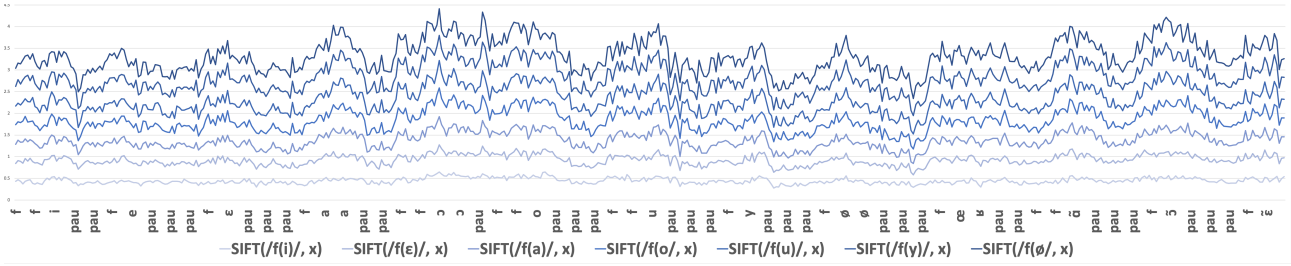
3.1. Analyzing the speakers

The EMD metric consistently showed visible variation in histogram shape across speakers. While speaker A's distribution had a shape that resembled a skewed normal distribution, speaker B's data displayed smaller hills to the left and to the right of the main distribution body (see Figure 3, below). This could be due to multiple reasons: 1) the static data of speaker A being a better fit for the dynamic data of speaker A than for the dynamic data of speaker B; 2) the metric picking up on the presence of multiple speaking strategies in speaker B (for all phonemes, unlikely); 3) it could be the case that the phonetic labeling algorithm struggled with speaker B more than with speaker A.

As for SSIM, the histograms of speaker B displayed much more pronounced hills in the less similar ranges of the similarity measure values than those of A (see Figure 3, middle).

Contrary to EMD and SSIM, the shape of distribution of SIFT proves to not be extremely sensitive to the change of speakers. However, overall distributions $SIFT_A(/static_ph/,x)$ and $SIFT_B(/static_ph/,x)$ do differ (Table 1, images of A being more similar to static images than those of B; the difference is statistically significant: Kolmogorov-Smirnov statistic 0.346277, p-value 0.0, and Mann-Whitney statistic 351926033.0, p-value

Figure 4: Stacked plots of SIFT between seven static images of /f/ (from bottom and the lightest color up to the darkest: /f(i), f(ɛ), f(a), f(o), f(u), f(y), f(ø)/) and the images of the dynamic sequence /fi, fe, fɛ, fa, fɔʁ, fo, fu, fy, fø, fœʁ, fɑ̃, fɔ̃, fɛ̃/



0.0).

All distance distributions equivalent between speaker A and speaker B are different (Kolmogorov-Smirnov and Mann-Whitney tests). Static images that depict the same consonant, but in different vocalic context may produce the same distribution within the same speaker, though it may not always be the case (/r(y)/ and /r(i)/ of A, /r(i)/ and /r(ɛ)/ and /r(y)/ and /r(o)/ of B producing same distributions in EMD and SIFT, though different in SSIM).

3.2. Phoneme comparisons

Then we aggregated distances by phonemes and analyzed the distances that appeared. Table 2 shows a part of the confusion matrix that we have. It demonstrates that EMD and SSIM similarity peaks do not seem to have any phonological grounds. When retrieving *N* smallest or greatest distances to a given static phoneme, they turn out to be associated to the same dynamic phonemes. SIFT also makes mistakes, but they are explicable: vowels stay classified as phonemes that do not involve complete obstruction of the airway (the measure looks out for the presence or absence of contact), possibly not paying enough attention to the extent of how close the articulators are supposed to come to each other (e.g. both /u/ and /o/ are back rounded vowels, and the difference between them is that /u/ is close and /o/ is close-mid; /w/ is a labial approximant, the closest consonantal equivalent of /u/).

Table 2: Confusion matrix: the rows are static captures of vowels, the columns their best fits in the dynamic data—for A, for B

Ph	EMD	SSIM	SIFT
/a/	/z,i/–/oe,ŋ/	/g,oe/–/ø,ʒ/	/ɔ̃,õe/–/l,ɔ̃/
/i/	/z,n/–/oe,ŋ/	/g,i/–/ø,ŋ/	/ʃ,y/–sil,/i/
/u/	/z,ŋ/–/oe,s/	/g,oe–/ø,ʒ/	/w,o/–/u,w/

Furthermore, SIFT can treat consonants, which require a higher degree of precision. The dynamic /f/ and /ʃ/ are on average the closest phonemes to all versions of the static /f/ and /ʃ/ respectively, and /ʒ/ for /s/. The static /j(i)/ matches /j/ and /ɥ/, and /w(u)/ to /w/ and /o/. While being reasonably good for fricatives and approximants, SIFT has a greater trouble with stops and liquids: the confusions involve not achieving the required constriction (as in the dynamic /ʃ/ being the closest to the static /l(i),l(ɛ),l(a)/) or mishandling nasalization (/b/ as the closest to the static /m(a),m(ɛ)/).

Figure 4 shows the temporal evolution of SIFT of the static /f(i), f(ɛ), f(a), f(o), f(u), f(y), f(ø)/) over the dynamic sequence /fi, fe, fɛ, fa, fɔʁ, fo, fu, fy, fø, fœʁ, fɑ̃, fɔ̃, fɛ̃/ by A. The plots show an increment at /f/ and go down at the vowel and pause. Distinguishing the anticipated vowel by the position of /f/ alone, however, seems impossible.

Speaker B represents more inexplicable similarity peaks, and spontaneous speech even more so. We suppose that the reason for this is either the greater mismatch between speakers A and B, or speaker B's peculiarity. It is due to mention another possible source of curious phenomena in the entire dataset: phonetic label files. Annotated with an automatic tool and force aligned with a signal whose denoising corrupts a part of speech information, they are prone to errors and should ideally be corrected by hand.

4. CONCLUSION

We investigated the matches for static articulatory targets in an RT-MRI dataset. As we identified them, they vary both within a single speaker and across speakers. Out of all metrics, SIFT used most of its domain of values and gave the most interpretable results. It captures best approximants and fricatives, while struggling at precision for articulators being in contact (stops) or wide apart (vowels).

Possible solutions would be to mask images and work only on the parts that represent speech-related areas with a method such as in [4]. One could model specific articulators, such as the tongue and the lips. Some effort should also be dedicated to correcting the phonetic annotation to be sure to draw conclusions from relevant data.

Further work could be to integrate this information into force alignment of RT-MRI data or to find the best images to serve as articulatory targets in speech synthesis.

5. REFERENCES

[1] Birkholz, P. 2013. Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PloS one* 8(4), e60603.

[2] Browman, C. P., Goldstein, L. 1992. Articulatory phonology: An overview. *Phonetica* 49(3-4), 155–180.

[3] Engwall, O. 2000. Are static MRI measurements representative of dynamic speech? results from a comparative study using MRI, EPG and EMA. *Sixth International Conference on Spoken Language Processing*.

[4] Lammert, A. C., Quatieri, T. F., Shadle, C. H., Narayanan, S. S. 2017. Speed accuracy tradeoffs in speech production. Technical report MIT LIncoln Laboratory Lexington United States.

[5] Laprie, Y., Elie, B., Tsukanova, A., Vuissoz, P.-A. 2018. Centerline articulatory models of the velum and epiglottis for articulatory synthesis of speech. *Eusipco*.

[6] Lin, Q. 1991. Speech production theory and articulatory speech synthesis. *The Journal of the Acoustical Society of America* 90(4), 2203–2203.

[7] Lowe, D. G. 1999. Object recognition from local scale-invariant features. *Computer vision, 1999. The proceedings of the seventh IEEE international conference on* volume 2. Ieee 1150–1157.

[8] Mussa-Ivaldi, F. A., Gantchev, N., Gantchev, G. 1999. Motor primitives, force-fields and the equilibrium point theory. *From Basic Motor Control to Functional Recovery. Academic Publishing House" Prof. M. Drinov", Sofia, Bulgaria* 392–398.

[9] Niebergall, A., Zhang, S., Kunay, E., Keydana, G., Job, M., Uecker, M., Frahm, J. 2013. Real-time mri of speaking at a resolution of 33 ms: Undersampled radial flash with nonlinear inverse reconstruction. *Magnetic Resonance in Medicine* 69(2), 477–485.

[10] Öhman, S. E. 1967. Numerical model of coarticulation. *The Journal of the Acoustical Society of America* 41(2), 310–320.

[11] Perrier, P. 2017. What goals for articulatory speech synthesis? *The 11th International Seminar on Speech Production*.

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

[13] Ramanarayanan, V., Van Segbroeck, M., Narayanan, S. S. 2016. Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. *Computer speech & language* 36, 330–346.

[14] Roekhaut, S., Brognaux, S., Beaufort, R., Dutoit, T. 2014. eLite-HTS: Un outil TAL pour la génération de synthèse hmm en français. *Démonstration aux Journées d'étude de la parole (JEP)*.

[15] Rubner, Y., Tomasi, C., Guibas, L. J. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision* 40(2), 99–121.

[16] Saltzman, E., Kelso, J. 1987. Skilled actions: a task-dynamic approach. *Psychological review* 94(1), 84.

[17] Saltzman, E. L., Munhall, K. G. 1989. A dynamical approach to gestural patterning in speech production. *Ecological psychology* 1(4), 333–382.

[18] Toutios, A., Sorensen, T., Somandepalli, K., Alexander, R., Narayanan, S. S. 2016. Articulatory synthesis based on real-time magnetic resonance imaging data. *INTERSPEECH* 1492–1496.

[19] Tsukanova, A., Elie, B., Laprie, Y. 2017. Articulatory speech synthesis from static context-aware articulatory targets. *ISSP 2017-11th International Seminar on Speech Production*.

[20] Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13(4), 600–612.

[21] Wu, Z., Watts, O., King, S. 2016. Merlin: An open source neural network speech synthesis system. *Proc. SSW, Sunnyvale, USA*.

[22] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., others, 2002. The HTK book. *Cambridge university engineering department* 3, 175.