Speaker's Articulatory Strategy Analysis: Theoretical Framework and Preliminary Experiment

Antoine Serrurier

Clinic for Phoniatrics, Pedaudiology, and Communication Disorders, University Hospital and Medical Faculty of the RWTH Aachen University, Aachen, Germany
Email: aserrurier@ukaachen.de

Abstract

During speech production, the vocal tract is articulated to achieve specific articulatory-acoustic goals. The generated signal conveys concomitantly phonetic information to be understood and speaker-specific characteristics. In particular, the speaker has a specific anatomical morphology that influences the articulatory strategy that he/she implements. This study proposes a theoretical framework to decompose the articulations into morphology and articulatory strategy constituents. The articulatory strategy constituent is further decomposed into a speaker-independent sub-constituent, related to the articulation of the phoneme, and a speaker-specific sub-constituent, related to the idiosyncratic articulatory strategy of the speaker for the considered phone. This framework is tested on 12 speakers articulating [a i u] to analyse whether the speaker-specific articulatory strategy tends to increase the size of the vowel triangle in the F1-F2 formant plane for enhanced intelligibility.

1 Introduction

In the process of speech production, the acoustic signal is generated by the speaker by adapting the shape of the vocal tract. The speaker controls the position and shape of the articulators, such as the tongue and the lips, so as to achieve the articulatory-acoustic goals necessary to generate the desired signal. The speech signal is at the same time robust enough to be understood by any listener unequivocally and specific enough to embed the speaker's characteristics in it. This article intends to provide a theoretical framework to analyse the speaker-specific characteristics.

The speaker-specific characteristics can be the consequence of different factors [1], such as the anatomic morphology, thereafter the *morphology*, the social background, the accent, the phonetic context, the adaptation to the listener, the idiosyncratic articulatory strategy, etc. In a recent work [2], it has been emphasised that the morphology, i.e. '*the intrinsic position, size, and shape of [a speaker's] organs or articulators irrespective of the articulatory task*' [2] forms the basis of all speaker's articulation on which the speaker does not have any control. The speaker has therefore to adapt her/his articulatory strategy, i.e. '*the displacement and deformation of the articulators*' [2] to achieve the desired acoustic-articulatory goals.

The current article intends to propose a practical implementation of this vision to separate in a speaker's articulation the morphology from the articulatory strategy. This will allow the analysis of the articulatory strategy independently of the morphology component. The concept of morphology of the vocal tract, including the soft tissues, has been introduced in a previous article [2]. The current article relies on these results for the definition and calculation of the morphology.

Removing the contribution of the morphology component in a speaker articulation can be seen as a sort of speaker normalisation, a technique traditionally used in multi-speaker studies (e.g. [3, 4]). The originality of the current approach lies in the normalisation of the sole morphology part of an articulation, i.e. the part of the articulation not related to the speech task.

In this context, the contribution of this study has three objectives: (1) propose a theoretical framework for the separation of the morphology and articulatory strategy, (2) propose a method for the analysis of the speaker's articulatory strategy and (3) design a first proof-of-concept experiment to illustrate how to manipulate these concepts.

As emphasised above, the articulatory strategy of a speaker has to comply with her/his morphology to achieve the articulatory-acoustic goals. Morphology and articulatory strategy are therefore intrinsically linked. To cope with this complexity, the overall method relies on the definition of artificial articulations that do not exist in reality but are useful abstract representations. To provide an overview, the morphology is defined as an artificial articulation independent of the speech task. The articulatory strategy is defined as the difference between the speaker's articulation and her/his artificial morphology articulation. The articulatory strategy component is then further decomposed into two sub-components: (1) a speaker-independent component corresponding to the phonetic task and (2) a speaker-specific component, corresponding to the distinct articulatory strategy of the speaker for the realisation of the phonetic task. The method is then tested on the three vowels [a i u] to analyse whether the speaker-specific components of the articulatory strategy contributes at increasing the distinction from other vowels by expanding the size of the vowel triangle in the F1-F2 plane.

2 Material & Methods

2.1 Data

Magnetic Resonance Imaging (MRI) recordings from twelve French native speakers, five females and seven males, have been considered. The speakers, in a supine position in the MRI scanner, were instructed to sustain a phoneme for about seven seconds during which a midsagittal image of the vocal tract was recorded. For the current study, a subset consisting of the three images corresponding to the task of sustaining the three cardinal vowels /a i u/ was considered. For each image, the vocal tract contours were manually outlined, aligned on a speaker reference coordinate system and finally on a single cross-speaker reference coordinate system attached to cranium landmarks. The articulator contours are restricted to the sections constitutive of the vocal tract, but their extension to non-vocal tract sections are included for illustration purposes only. These data form a subset of the data presented in [2], where further details are provided on the speakers, the acquisition process, the

ISBN: 978-3-8007-6164-7

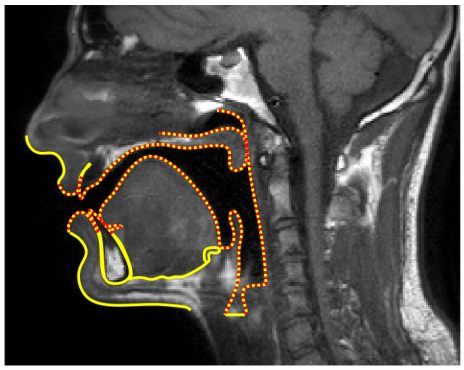

Figure 1: MRI image superimposed with the vocal tract contours (dotted red lines) and same contours also extended to non-vocal tract sections (solid yellow lines).

image characteristics, and the contour extraction and alignment process. An example of an image and the vocal tract contours is provided in Figure 1.

Altogether, the data consist of 12 speakers $\times$ 3 articulations $\times$ 1044 points $\times$ 2 x-y coordinates. Formally, an articulation A is a matrix of point coordinates of size $(n \times 2)$, where $n = 1044$ is the number of points.

2.2 Method

The principle of the method relies in identifying in an articulation the component related to the morphology of the speaker and the components related to her/his articulatory strategy. As detailed in the introduction, these two components are intrinsically linked. Practically, these components can, however, be seen mathematically as additive components: The morphology forms the common basis to all speaker's articulations and the difference between an articulation and this common morphology basis can be considered as the speaker's articulatory strategy to produce the articulation. The following sections describe technically how this theoretical approach can be implemented.

2.3 Morphology

In a recent work [2], we proposed a morphological model of the vocal tract. This relies on the determination for a speaker of a so-called morphological average-articulation representing the morphological common background of all speaker's articulations. Briefly, this is calculated as the mean articulation of a set of articulations representative of the articulatory repertoire of the speaker. In practise, it has been shown that a set of articulations comprising the oral vowels and some consonants in one or several vowel contexts are enough for this purpose. Only the velum, usually presenting an unbalanced bimodal distribution on this set of articulations (upper position for an oral phoneme, lower position for a nasal phoneme), is computed slightly differently as the intermediary position between the upper and lower positions.

The result is a vocal tract contour that looks like an articulation and has all the characteristics of an articulation but is an artificial articulation. It is therefore referred to as *morphological average-articulation*, shortly *average-articulation*, to keep the distinction with the real articulations produced by the speaker. It is represented for one speaker of the study by the black contour on the Figure 2. Formally, the morphological average-articulation M is a matrix of point coordinates of size $(n \times 2)$.

2.4 Articulatory Strategy

Once the morphological average-articulation of a speaker has been determined, her/his articulatory strategy S for an articulation can be calculated as the simple subtraction of this average-articulation from the articulation:

$$S = A - M$$

The difference between these two contours is illustrated for one articulation of one speaker in Figure 2: The articulation A is represented in blue, the average-articulation M in black and the difference between these two, S, in red and green areas (left figure).

This component can itself be seen as the addition of two elements: The strategy common to all speakers to produce the considered phoneme and the speaker-specific marginal strategy component that the speaker implements in addition to achieve her/his particular articulation. Indeed, across-speaker strategies are well understood in the field of phonetics [5]. For instance, /i/ is achieved by bringing the tongue in a frontward and upper position and /m/ is achieved by closing the lips and opening the velopharyngeal port. The idea in the current method is to make the distinction between this common strategy across speakers and the speaker-specific strategy for this articulation. The common phoneme strategy is referred to in this study as the *phoneme strategy* and the speaker-specific marginal articulation strategy as the *speaker marginal strategy*:

$$S = PS + SMS$$

where PS is the phoneme strategy and SMS the speaker marginal strategy. Note that unlike the articulations and average-articulations, the strategy matrices represent differences between articulations and/or average-articulations and cannot be represented as articulations. The two levels of strategy can also be seen as a first gross articulatory setting towards the phonetic goal and a fine tuning from the speaker to achieve her/his specific goals.

Practically, the phoneme strategy is obtained by calculating the mean articulatory strategy over all speakers for the given phoneme, i.e. the difference between the articulation and the morphological average-articulation. An intermediary artificial articulation, referred to in the rest of the manuscript as the *speaker phoneme-articulation*, can then be calculated as the addition of the morphological average-articulation with the phoneme strategy. It can be seen as the articulation that the speaker would produce if he/she was simply implementing the average articulatory strategy. Note that this articulation is speaker-specific as it embeds the speaker morphology. This phoneme-articulation is represented by the red contour on the Figure 2 and the difference between the two contours of the middle figure represents the phoneme strategy. The speaker marginal strategy is then obtained by calculating the difference between the effective articulation of the speaker (in blue in the Figure 2) and her/his phoneme-articulation (in red). This strategy is represented by the difference between the two contours on the right of Figure 2. It represents the articulatory strategy that the speaker implements on top of the phoneme strategy to achieve her/his articulation.

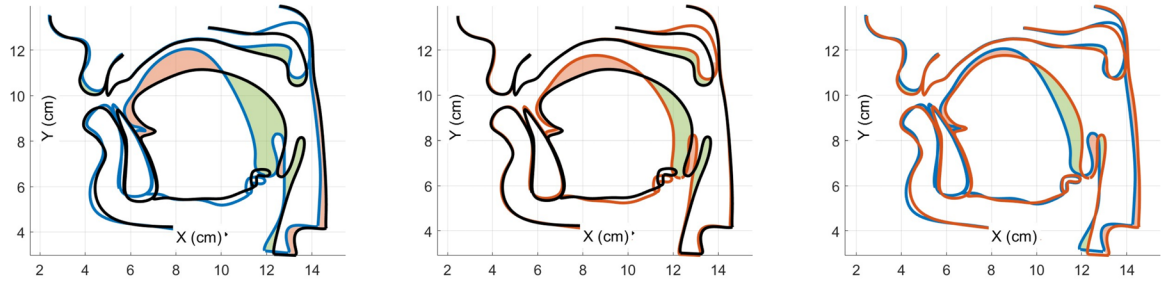

Figure 2: Articulations and artificial articulations for a speaker producing [i]. Left: Morphological average-articulation M (black) and articulation A (blue); the difference between the two contours, represented by the red and green areas, corresponds to the articulatory strategy S of the speaker. Middle: Morphological average-articulation M (black) and speaker phoneme-articulation (red); the difference between these two contours corresponds to the speaker-independent phoneme strategy PS for [i]. Right: Articulation A (blue) and speaker phoneme-articulation (red); the difference between the two contours corresponds to the speaker-specific marginal articulation strategy SMS. The red and green area colors on the three figures symbolise the positive and negative differences between the contours.

In summary, an articulation can be considered as the addition of three elements:

- The phoneme-independent morphology of the speaker
- The speaker-independent articulatory strategy for the phoneme
- The speaker-specific phoneme-specific marginal articulatory strategy

2.5 Experiment

The purpose of the experiment is to use the proposed theoretical framework to carry out a first analysis of the articulatory strategies to illustrate how to manipulate these concepts. Considering that the analysis of the morphology was the focus of a previous study [2] and that the phoneme strategies are nowadays rather well-understood [5], this study concentrates on the speaker marginal strategy. In addition, our objective being to characterise the speaker rather than the language, we focus in this experiment on the speaker-specific component of the articulatory strategy.

Vowels are primarily characterised by their formants, in particular the two first formants F1 and F2. One can therefore reasonably assume that one of the primary goals of a speaker to be understood unequivocally is to maximise the distance in the F1-F2 plane between the vowels. Of course, in real speech, this goal might be competing with other conflicting goals, such as for instance minimising the articulatory movements. The resulting articulation is ultimately a trade-off between all the constraints. Nevertheless, the current experiment aims at analysing whether the marginal strategy tends to increase the distance between vowels in the F1-F2 plane.

For simplicity, only the cardinal vowels [a i u] were considered. Indeed, these vowels form the extremes of the vowel triangle and enhancing their distinction from the other vowels consist naturally in moving them in an outward direction from the triangle centre. The experiment hypothesis is thus that the speaker marginal strategy tends to move the F1-F2 formants of the [a i u] vowels in an outward direction from the [a i u] triangle in the F1-F2 plane.

To test this hypothesis, the following experiment has been carried out:

1. For each speaker, the morphological average-articulation has been calculated following the method proposed in [2].
2. For each vowel [a i u], the phoneme strategy has been calculated as the mean articulatory strategy over all speakers, i.e. the mean of the differences between the articulation and the morphological average-articulation:

$$PS_k = \frac{1}{N}\sum_{i=1}^{N}(A_{ik} - M_i)$$

   where k represents the vowels [a], [i] or [u], N is the number of speakers, M_i is the morphological average-articulation of the speaker i, A_{ik} the articulation k of the speaker i and PS_k the phoneme strategy for the vowel k.
3. For each vowel of each speaker, the speaker phoneme-articulation has been calculated as the sum of the morphological average-articulation of the speaker with the phoneme strategy of the vowel:

$$A'_{ik} = M_i + PS_k$$

   where A'_{ik} is the phoneme-articulation for the speaker i and the vowel k. It represents the articulation that the speaker would produce if he/she did not implement any speaker marginal strategy for this phoneme.
4. For all articulations A_{ik} and phoneme-articulations A'_{ik}, the acoustic transfer function has been calculated by simulating plane acoustic wave propagation in the vocal tract [6]. For that purpose, the vocal tract is represented in a series of elementary tubes whose cross-sectional areas are derived from the midsagittal vocal tract contours following an α-β model [7]. Note that the minimal cross-sectional area has been set to 0.1 cm^2 to ensure wave propagation and formant values. The transfer function is calculated using an electrical equivalent [8] and the frequency values of the two first formants are extracted. Further details are provided e.g. in [2]. The formants of the phoneme-articulations are denoted in the following as the *phoneme-formants*, as opposition to the *formants*, corresponding to the real articulations.
5. The formants and phoneme-formants are plotted in the F1-F2 plane and analysed to test the hypothesis. In addition, the areas of the triangles formed by the [a i u] formants in the F1-F2 plane are calculated and compared to those of the phoneme-triangles formed by the [a i u] phoneme-formants.

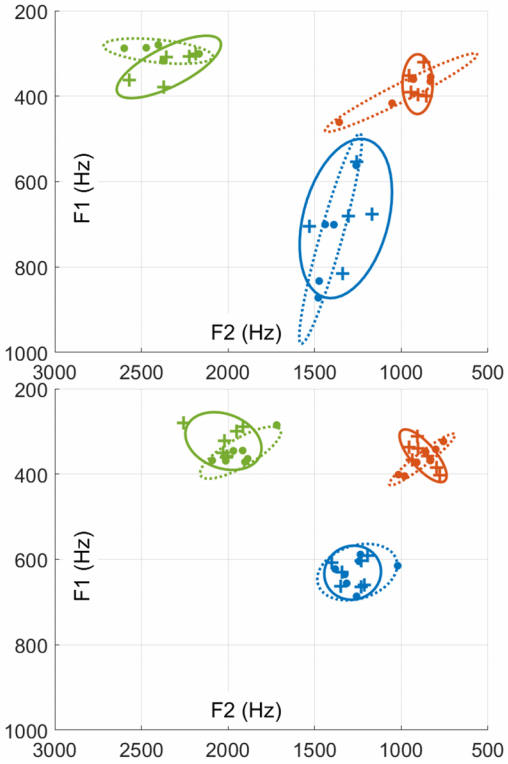

Figure 3: Formants ('+') and phoneme-formants (dots) together with their standard-deviation ellipses (resp. solid and dashed lines) in the F1-F2 plane for the vowels [a] (blue), [i] (green) and [u] (red) for the female (top) and male (bottom) speakers.

3 Results

The formants and phoneme-formants in the F1-F2 plane are visible in Figure 3. We observe in general that the formants do not tend to be more outward than the phoneme-formants. The areas of the triangles formed by the [a i u] formants in the F1-F2 plane and those of the phoneme-triangles are visible in Figure 4. Again, the triangles do not show larger areas than the phoneme-triangles.

4 Discussion & Conclusion

In general, the results do not confirm the initial hypothesis: The formants of the [a i u] vowels do not tend to be more outwards than the phoneme-formants in the F1-F2 plane. In other words, it is not observed that the speaker marginal strategy tends to increase the distance of the [a i u] formants to the other formants. It suggests as a consequence that, for the data considered in the study, the speaker marginal strategy is driven by other motivations than maximising the formant distance in the F1-F2 plane. As mentioned in the introduction, these motivations can be as varied as the morphology, the social background, the accent, the phonetic context, etc.

The current experiment tested only one possible hypothesis for the speaker strategy on very limited data. This was intended to provide an example of use of the theoretical articulatory strategy analysis presented in this article. Further analyses on extensive data are naturally needed in the future.

The current article presents an original theoretical framework to analyse the speaker articulatory strategy. In this framework, a speaker articulation is considered as the sum of the speaker-specific morphology articulation, the speaker-independent phoneme articulatory strategy and the speaker-specific phoneme-specific marginal articulatory strategy. Naturally, this is a theoretical representation, as the speaker does not accumulate independently these three elements but forms an articulation as a whole. The morphological average-articulation and the phoneme-articulation are therefore artificial articulations that do not exist in reality, which can sometimes lead to some lack of realism. For instance, it can happen that the tongue contour crosses unrealistically the palate contour in a phoneme-articulation. It can happen for speakers with much flatter palates than the other speakers. This is not a problem as the articulatory 'correction' is captured by the speaker marginal strategy which will tend to bring back the tongue downward, indicating rightfully that (1) the speaker has a specific articulatory strategy different from the rest of the speakers and (2) this strategy is, at least partly, driven by morphological constraints. Although based on artificial representations, this approach allows to derive meaningful information for the speaker's articulatory strategy. This is also the reason why a minimum cross-section area had been imposed in the calculation of the acoustic transfer function of the phoneme-articulations, to ensure the calculation of formants related to the articulation even in case of occlusion. Further analyses are necessary to address properly this issue in the future.

The experiment, intended to illustrate how to manipulate the concepts presented in the article, relies on three vowels sustained several seconds by twelve speakers. This small subset can evidently not be representative of the speakers' articulatory strategies. Further analyses with extensive dynamic data, such as real-time MRI, are necessary to characterise accurately the speaker articulatory strategy.

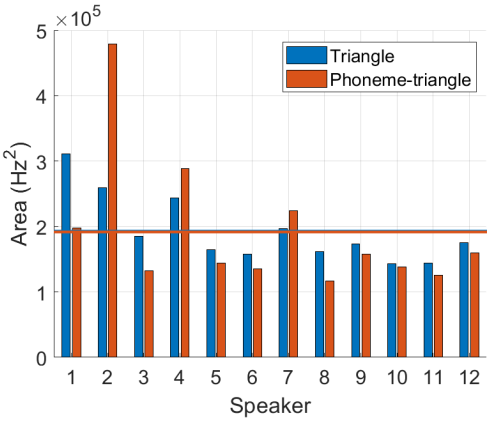

Figure 4: Areas of the [a i u] triangles and phoneme-triangles for all speakers. The solid horizontal lines represent the average areas.

Acknowledgements

The author is very grateful to P. Badin and T. Sawallis for providing the data and to L. Lamalle for helping at recording them. The data recording has been partially funded by the French ANR (Grant No. ANR-08-EMER-001-02 "ARTIS"). The IRMaGe MRI facility was partly funded by the program "Investissement d'Avenir" run by the "Agence Nationale de la Recherche" – Grant "Infrastructure d'avenir en Biologie Santé" – ANR-11-INBS-0006.

References

[1] P. Ladefoged and D. E. Broadbent, "Information conveyed by vowels," *The Journal of the Acoustical Society of America*, vol. 29, no. 1, pp. 98–104, 1957.

[2] A. Serrurier and C. Neuschaefer-Rube, "Morphological and acoustic modeling of the vocal tract," *The Journal of the Acoustical Society of America*, vol. 153, pp. 1867–1886, mar 2023.

[3] M. Hashi, J. R. Westbury, and K. Honda, "Vowel posture normalization," *The Journal of the Acoustical Society of America*, vol. 104, no. 4, pp. 2426–2437, 1998.

[4] C. Geng and C. Mooshammer, "How to stretch and shrink vowel systems: Results from a vowel normalization procedure," *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3278–3288, 2009.

[5] P. Ladefoged and K. Johnson, *A Course in Phonetics*. Cengage Learning, 2011.

[6] G. Fant, *Acoustic Theory of Speech Production*. The Hague: Mouton, 1960.

[7] A. Soquet, V. Lecuit, T. Metens, and D. Demolin, "Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with MRI," *Speech Communication*, vol. 36, no. 3-4, pp. 169–180, 2002.

[8] J. L. Flanagan, *Speech Analysis Synthesis and Perception*. Springer-Verlag, Berlin, Heidelberg, New-York, 1972.

