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

A Multimodal Real-Time MRI Articulatory Corpus of French for Speech Research

Ioannis K. Douros[1,2], Jacques Felblinger[2,4], Jens Frahm[3], Karyna Isaieva[2], Arun A. Joseph[3], Yves Laprie[1], Freddy Odille[2,4], Anastasiia Tsukanova[1], Dirk Voit[3], Pierre-André Vuissoz[2]

[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
[2]Université de Lorraine, INSERM U1254, IADI, F-54000 Nancy, France
[3]Biomedizinische NMR, MPI für biophysikalische Chemie, 37070 Göttingen, Germany
[4]Université de Lorraine, INSERM, CIC-IT 1433, CHRU de Nancy, F-54000 Nancy, France

ioannis.douros@loria.fr, karyna.isaieva@univ-lorraine.fr, yves.laprie@loria.fr, freddy.odille@inserm.fr, anastasiia.tsukanova@loria.fr, pa.vuissoz@chru-nancy.fr

Abstract

In this work we describe the creation of ArtSpeechMRIfr: a real-time as well as static magnetic resonance imaging (rtMRI, 3D MRI) database of the vocal tract. The database contains also processed data: denoised audio, its phonetically aligned annotation, articulatory contours, and vocal tract volume information, which provides a rich resource for speech research. The database is built on data from two male speakers of French.

It covers a number of phonetic contexts in the controlled part, as well as spontaneous speech, 3D MRI scans of sustained vocalic articulations, and of the dental casts of the subjects. The corpus for rtMRI consists of 79 synthetic sentences constructed from a phonetized dictionary that makes possible to shorten the duration of acquisitions while keeping a very good coverage of the phonetic contexts which exist in French. The 3D MRI includes acquisitions for 12 French vowels and 10 consonants, each of which was pronounced in several vocalic contexts. Articulatory contours (tongue, jaw, epiglottis, larynx, velum, lips) as well as 3D volumes were manually drawn for a part of the images.

Index Terms: speech corpus, speech production, speech synthesis, 3D MRI data, real-time MRI data, multi-modal database, French language

1. Introduction

In recent years there has been an increasing interest in data that include audio-articulatory recordings. Such data have several applications and help improve results in the field of speech production [1], speech recognition [2], etc.

Magnetic Resonance Imaging (MRI) provides an unequalled means of observing the vocal tract during speech production due to the possibility of covering the whole vocal tract with a very good geometric precision and doing several acquisitions without known health hazard as long as the conditions of use of the MRI are respected. This paper reports on the design of a 3D static and 2D real-time dynamic MRI database. Dynamic data can be used for studying continuous speech, i.e. read speech and spontaneous speech as well, while static 3D data allows all the cavities of the vocal tract to be measured and numerical simulations to be carried out for a single well-articulated sound. While it is relatively easy to find audio-articulatory databases in English [3, 4], there is no free similar data in French.

Despite of the visibility of the MRI images by themselves, sometimes it may be complex to extract the information about the form and position of the articulators. In ArtSpeechMRIfr, part of the images was semi-automatically processed, and corresponding segmentation or contouring maps are present in the database. In addition, while there are 3D MRI databases with audio recordings, in most cases the sound was recorded not at the same time with the MRI recording, as it happens in our case.

The paper is organised as follows. In section 2 the acquisition setups are described while the content of the database is presented in section 3. In section 4 some possible applications of these data are presented, and finally in section 5 we conclude by giving our future plans for the extension of the database, and by suggesting some research perspectives.

2. Data acquisition

The acquisition was carried out in two parts: the 2D real-time MRI data (rtMRI) were recorded at Max Planck Institute in Göttingen, Germany, while 3D static data (3D MRI) was recorded at Nancy Hospital, France.

2.1. Subjects

The selected subjects are 2 adult male French native speakers speaking French. Subject 1 (S_1) is male, 32 years old, 180 cm tall and 65 kg, while subject 2 (S_2) is male, 35 years old, 182 cm tall and 74 kg.

2.2. 2D data

Our rtMRI dataset was recorded on a Siemens Prisma-fit 3T scanner (Siemens, Erlangen, Germany). We used radial RF-spoiled FLASH sequence [5] with TR $=$ 2.02 ms, TE $=$ 1.28 ms, FOV $=$ 19.2 $\times$ 19.2 cm, flip angle $=$ 5 degrees, and slice thickness is 8 mm. Pixel bandwidth is 1600 Hz/pixel. Image resolution is 136 $\times$ 136. The acquisition time varied from 34 sec to 90 sec, mostly about 60 sec. We followed the protocol described in [6]. Images are recorded at a frame rate of 55 frames per second with the algorithm presented in [5].

2.3. 3D data

The 3D MRI data was recorded at Nancy Central Regional University Hospital under the approved medical protocol "METHODO" (ClinicalTrials.gov Identifier: NCT02887053).

Subject S_2's data was recorded on a General Electric Signa HDxt 3T scanner (GE healthcare, Chicago, Illinois, United States). We used 3D FGRE (TR $=$ 3.12 ms, TE = 1.084, FOV

http://dx.doi.org/10.21437/Interspeech.2019-1700

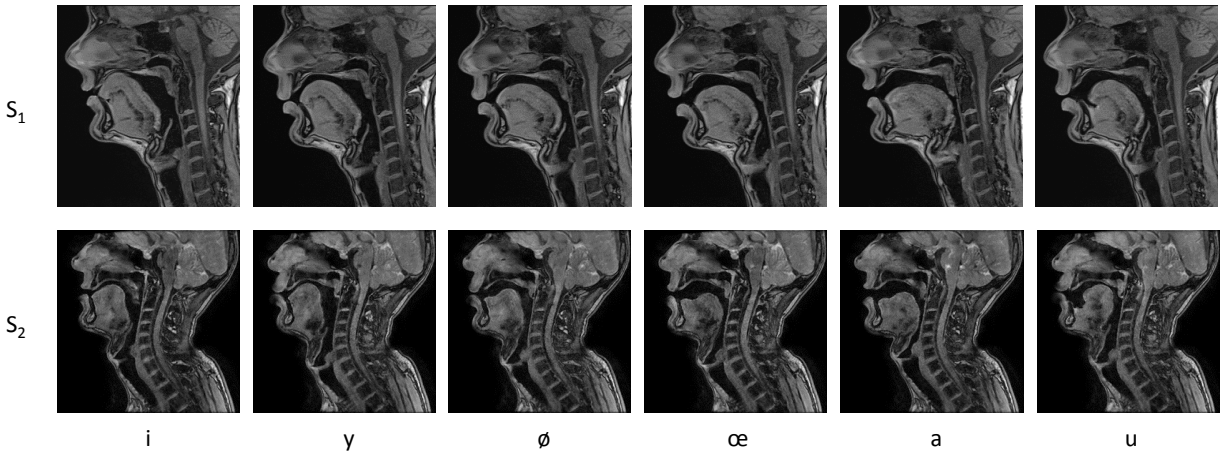

Figure 1: *Mid-sagittal slice of the static 3D images of subjects S_1 (upper line) and S_2 (lower line) for several of the French vowels.*

$= 26 \times 26$ cm, flip angle $= 10$ degrees) for the acquisition. Scan slice thickness is 2 mm, spacing between slices is 1 mm and pixel bandwidth is 488 Hz/pixel. Acceleration factor is 2. The image resolution is 256×256 with 76 slices. Duration of one acquisition is 12.7 seconds. Examples of mid-sagittal cuts of this data are shown on figures 1 and 2.

Subject S_1's data was recorded on a Siemens Prisma 3T scanner (Siemens, Erlangen, Germany). We used 3D VIBE (TR $= 3.57$ ms, TE $= 1.43$, flip angle $= 9$ degrees) for the acquisition. Acceleration factor is iPAT $= 3$. Scan slice thickness is 1.2 mm, FOV $= 22 \times 20$ cm and pixel bandwidth is 445 Hz/pixel. Data is divided in two parts.

In the first part, audio was recorded just before the MRI scan started, and the subject was asked to keep the same articulatory position without phonation for 15 seconds, i.e. during the acquisition. The image resolution is 256×232 with 120 slices. One example of this type of acquisition is shown in Figure 1.

In the second part, audio was recorded simultaneously with the MRI acquisition. Duration of the acquisition was both 7 seconds and 15 seconds. The image resolution is 320×290 with 36 slices and 256×232 with 120 slices, respectively. Some data from 15 seconds acquisitions are presented on Figure 2.

2.4. Sound recording

Audio is recorded at a sampling frequency of 16 kHz inside the MRI scanner by using a FOMRI III optoacoustics fibre-optic microphone. The subject wears earplugs to be protected from the noise of the scanner, but is still able to communicate orally with the experimenters via an in-scanner intercom system. Since the sound is recorded at the same time with the MRI acquisition, there is additional noise in the audio signal. In order to de-noise it, we used the de-noising algorithm proposed in [7].

2.5. Transcription of the continuous speech corpus

Text alignment was done with Astali [8], which can exploit an optional pronunciation dictionary if some words do not exist in the default lexicon. The transcription procedure is based on the guidelines described in [9].

This procedure has some limitations. In particular, since the bilabial trill /β/ and the alveolar trill /r/ do not belong to the French language, they are not recognized by French SAMPA alphabet. Therefore, they had to be mislabeled: such sequences as /aβa/ and /ara/ had to be transcribed as /aba/ or /ava/ and /ara/.

A phenomenon that turned out to be impossible to represent within the framework of SAMPA was stops with a long closure due to gemination, occurring in sequences like "*crabes bagarreurs*", /krabˈba.ga.ʁœʁ/: the first /b/ has no audible release and is followed by the next /b/. Their transcriptions were decided on a case-by-case basis.

3. Database description

The objective of the database is to enable the exploration and modeling of coarticulation phenomena. It is thus necessary to get a good geometric description of the whole vocal tract and to get running speech which exhibits how the global geometry of the vocal tract evolves over time during speech production. So far, despite the presence of techniques for dynamic 3D MR acquisition [10], time and spatial resolution of such images is still quite low. The corpus construction strategy therefore consists of collecting a number of static configurations of the vocal tract corresponding to sustained vowels, or blocked CV articulations in 3D on the one hand and running speech in 2D (in the mid-sagittal plane) on the other hand.

The images must provide the greatest possible variability of articulatory shapes since we use these shapes to build an articulatory model. Besides, the teeth are not visible on the MRI images, and therefore it is necessary to merge these data with a numerical scan of the subject's dental cast. This requires additional MRI volumes to derive the position of teeth which are not visible. In our case, there were three: one by pressing the tongue against the upper teeth, especially in the area of the incisors, one by pressing the tongue against the lower teeth, and one with the lower and upper incisors in contact. The tongue is clearly visible on MR images, and pressing it against the teeth makes them appear in negative.

3.1. Real-time 2D data

The analysis of coarticulation requires a good coverage of all phonetic contexts which can appear in French. Reading specifically prepared sentences meets this objective. For some specific

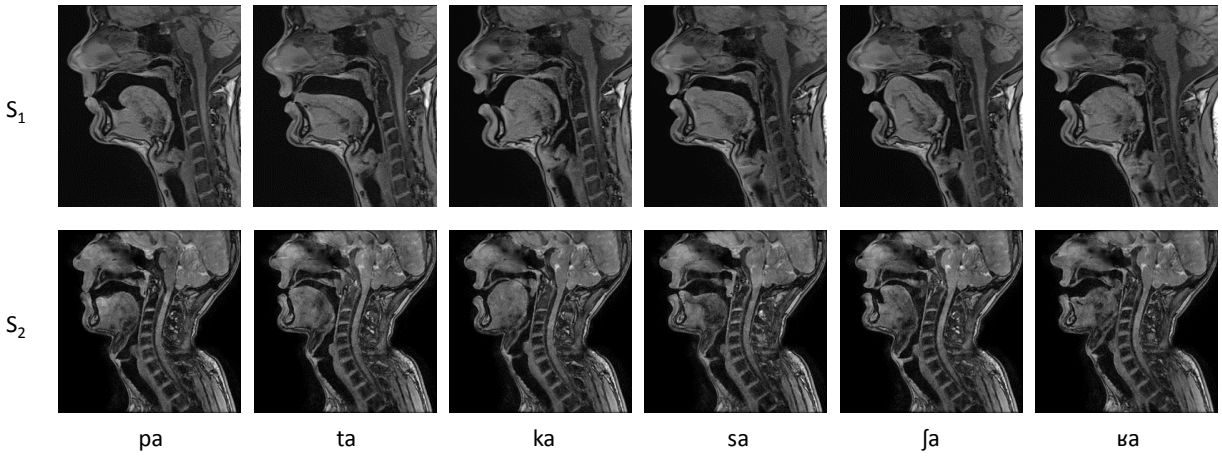

Figure 2: *Mid-sagittal slice of the static 3D images of subjects S_1 (upper line) and S_2 (lower line) for some of the French consonants. Consonant was pronounced in context of the following vowel.*

issues, and to remove distant contextual effects, nonsense words (for instance a selected few CV or complex consonant clusters followed by a vowel) are better and are thus included in the database.

However, coarticulation is often more pronounced in the case of spontaneous speech which is less controlled and gives rise to stronger articulatory adjustments. Similarly, there is intra-speaker variability, and it is thus interesting to add several repetitions of the sentences, or at least some of them. In previous work [11] we had worked with X-ray films [12] and so we added some of these sentences to compare both techniques of acquisition.

Each of these aspects gave rise to a part of the dynamic corpus of speech, which is described below.

In speech synthesis, the classical way of constructing a corpus consists of adding sentences from a vast written corpus, newspaper for instance, so as to enrich the linguistic coverage iteratively. Despite the efficiency of the construction algorithm each sentence contributes to a limited number of new phonetic contexts. To prevent a very long recording, the corpus design strategy consisted of constructing sentences by hand from a phonetized dictionary so as to add the expected phonetic contexts, i.e. those not present in the existing sentences. The dictionary is the phonetized version of the French Morphalou lexicon [13] which provides 620.000 flexed forms [14].

We used several levels of criteria to guide the manual construction of new sentences from words. After the insertion of each sentence the first level of criteria evaluated is the number of VV for all the vowels, the number of CV for C in /ptkfsʃlʁ/ and V in /i, a, u/ plus /y/, the number of VC with C as a coda and C in /l, ʁ, n, m/ and V in /i, a , u, y, e, ɛ, o, ɔ/, the consonant clusters C1C2V with C1 in /ptkbdgf/, C2 in /ʁl/ and V in /a, i, u, y/ (the other CCV following the same pattern with /sʃv are rare in French), and 15 complex consonant clusters (at least a sequence of 3 consonants, between two vowels). Except for those clusters and with very few exceptions all the contexts appear within words to avoid the effect of prosodic boundaries.

This first level of criteria covers the very heart of the corpus in terms of mandatory phonetic contexts. We wanted well constructed sentences of French and therefore words do not corresponding to the targets contexts were added. They provide new contexts, and in particular contexts with vowels outside the set of cardinal vowels plus /y/. VCV are counted by taking into account groupings of close vowels. There are 6 groups of vowels (/i,e/, /ɛ,a/, /u, o, ɔ/, /y, ø/, /oe,ə/ and nasal vowels /ã, õ, ɛ̃, œ̃/. This provides a second level of evaluation which guides the choice of words required to build well constructed sentences.

In total this corpus is made up of 79 sentences offering a very good coverage of all the phonetic contexts in French. Even if these sentences are sometimes a little bit curious they remain perfectly readable. There are only two non French words (”cartoons” and ”squaw”) but they can be easily pronounced by French speakers. One objective was to enable the comparison with an old X-ray database of 15 very short utterances (groups of 3 short words) [15]. In total those sentences represent 138 phonemes, and less than 30 seconds of reading were included in the corpus.

Then the corpus was augmented with sentences frequently used in phonetics “La bise et le soleil…” sequence [16], some sentences from a corpus from [17] and six sequences with trills that do not belong to the French language: /aβa, iβi, uβu, ara, iri, uru/.

To study non-spontaneous speech, each sentence from the core part of the corpus was to be uttered at least three times. Sentences are presented to the speaker in random order.

For investigation of spontaneous speech, each subject was presented with randomly ordered prompts to talk about for a minute. They covered everyday topics: “What do you like in your work?”, “Speak about your last trip anywhere”, “Speak about a film or a book that had a lasting impression on you” (20 topics in total). Despite having hesitations in speech, both subjects had enough to say to fill the allowed minute. In total, the database contains 2h17m of speech recorded in the mid-sagittal plane by rtMRI.

3.2. Static 3D data

We made acquisitions for vowels and blocked /CV/ articulations. For vowels the subjects are instructed to phonate the vowel before the acquisition noise starts. They had to stop the phonation just before the acquisition starts and keep the same articulation during the acquisition. Asking subjects to phonate the vowels allows them to adjust the articulation. In

order to get similar articulations between speakers we asked them to phonate the vowel in the context /pV/, with a very long vowel. We chose /p/ because all the /pV/ correspond to a French word (except /pɔ/). For consonant articulations subjects were instructed to choose the articulatory position that would allow them to produce the expected /CV/. We accepted /p,t,k,f,s,ʃ,l,ʁ,m,n/.

The ArtSpeechMRIfr covers:

- all French vowels /i,e,ɛ,y,ø,œ,ɔ,o,u,ɑ̃,õ,ɛ̃/ with a single acquisition for /œ̃, ɛ̃/ since the vast majority of French speakers no longer realizes the contrast between both vowels [18]. Despite the precautions taken to ensure that the articulation of vowels inside an MRI machine is as close as possible to natural speech, subjects generally reduce the aperture, which is therefore underestimated in the models built from those images. For this reason we asked subjects to record an extra vowel which is a "very open" /a/, i.e. similar to a vowel that would be articulated with a loud voice. Examples of the mid-sagittal slices for vowels are shown on Fig. 1.
- /p,t,k,f,s,ʃ,l,ʁ,m,n/ followed by vowels /i,a,u/ as a minimal set of CV. According to the subject, his judgment about his immobility (or the closeness with the target) which could require some acquisitions to be repeated, for vowels this minimal set can be extended. Extensions consist of adding other intermediate vowels. Examples of some consonant articulations are given on Fig. 2.

4. Applications

4.1. Articulatory modeling

The articulatory model is intended to generate the geometric shape of the vocal tract from a small number of parameters corresponding to the speech articulators. It is a key component of an articulatory synthesizer and the challenge is to design a model that can generate all the possible forms of the vocal tract that can appear during speech production. Our model takes into account the links between the articulators so as to find out the intrinsic deformation factors for each of the articulators by applying Principal Component Analysis (PCA) on articulatory contours extracted from static images [19]. This model has been improved several times, most recently for elongated articulators [20], i.e. epiglottis and uvula. Usually, the input of PCA are the contours of the target articulator for all the static MRI images of one speaker. However, due to delineation errors, direct application of PCA in this case leads to unrealistic PCA components, and especially irrelevant swelling deformation modes. The new model relies on the application of the PCA to the central line of these two articulators. The model improved in this way gives much better results on the epiglottis and uvula. This model built from the ArtSpeechMRIfr database's static images can be tested on the database's dynamic images. The contours of the articulators were carefully extracted semi-automatically or by hand, and corrected if necessary, for about 500 images, which allows the accuracy of the articulatory model to be assessed.

4.2. Acoustic simulations

To carry out synthesis, several simplifications are generally made to keep a reasonable calculation time. The first consists in making the hypothesis of a plane wave propagating through the vocal tract which allows the analogy with an electrical transmission line to be made. The second consists in moving from three-dimensional volumes to mid-sagittal slices. The existence in ArtSpeechMRIfr of 3D data together with the acoustic signal enables the impact of those simplifications on acoustics to be explored.

4.2.1. Comparison between various types of simulations

We investigated five French vowels represented by 3D MRI [21]. Two types of simulations were performed : acoustic simulations that use 2D or 3D data without any hypothesis on the wave propagation on the one hand, and electrical simulations using the mid-sagittal slice with the approximation of the transverse area. The results demonstrate fairly good agreement between 2D and 3D but show that the electrical simulations have a significant impact on some formants.

4.2.2. Impact of approximation at the level of velum and epiglottis

The epiglottis and uvula add a certain complexity to the shape of the vocal tract. For example, the space between the epiglottis and the tongue varies over time and it is therefore important to know whether or not it should be taken into account in simulations and articulatory models. The mid-sagittal slice from the 3D data was used to investigate this issue [22].

4.3. Comparison between static and dynamic data

The existence of static and dynamic data makes it possible to know to what extent 3D static data (of very good quality) can approach 2D dynamic data (of lower quality). This question arises as soon as an articulatory model is developed or when the objective is to reconstruct better quality dynamic images by exploiting static images. It has been shown that there is a high intra-speaker variability [23] and that there are some dynamic images, which are far from static images, that cannot be approximated with an articulatory model. It is therefore necessary to include some dynamic images in addition to static images to improve the articulatory model.

5. Future plans

This database has been built with the intention of being able to be used for many complementary applications. Containing static 3D data offers the possibility of performing acoustic experiments, for example using 3D printed models. It is also possible to study various types of articulatory representations/models and further examine the role and coordination of the articulators in speech production. Finally, research in articulatory-acoustic mapping and automated recognition systems can also benefit from this database since the quality of the denoised speech is well above that of other real-time MRI databases. We plan to extend the database by including more subjects (male and female), providing 2D dynamic MRI data with higher frame rate and improved the sound. Finally, we intend to add more processed material like delineations of the speech articulators.

6. Acknowledgments

Research supported by the project ArtSpeech of ANR (Agence Nationale de la Recherche), France. The authors would like to thank INSERM, FEDER and Région Lorraine for their supports.

7. References

[1] B. Elie and Y. Laprie, "Simulating alveolar trills using a two-mass model of the tongue tip," *Journal of the Acoustical Society of America*, vol. 142, no. 5, 2017. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01525882

[2] I. K. Douros, A. Katsamanis, and P. Maragos, "Multi-view audio-articulatory features for phonetic recognition on rtmri-timit database," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5514–5518.

[3] S. Narayanan, E. Bresch, P. K. Ghosh, L. Goldstein, A. Katsamanis, Y. Kim, A. Lammert, M. Proctor, V. Ramanarayanan, and Y. Zhu, "A multimodal real-time mri articulatory corpus for speech research," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[4] T. Sorensen, Z. I. Skordilis, A. Toutios, Y.-C. Kim, Y. Zhu, J. Kim, A. C. Lammert, V. Ramanarayanan, L. Goldstein, D. Byrd *et al.*, "Database of volumetric and real-time vocal tract mri for speech science." in *INTERSPEECH*, 2017, pp. 645–649.

[5] 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.

[6] A. Niebergall, S. Zhang, E. Kunay, G. Keydana, M. Job, M. Uecker, and J. Frahm, "Real-time mri of speaking at a resolution of 33 ms: Undersampled radial flash with nonlinear inverse reconstruction," *Magnetic Resonance in Medicine*, vol. 69, no. 2, pp. 477–485, 2013.

[7] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 4, pp. 1118–1133, 2012.

[8] D. Fohr, O. Mella, and D. Jouvet, "De l'importance de l'homogénéisation des conventions de transcription pour l'alignement automatique de corpus oraux de parole spontanée," in *8es Journées Internationales de Linguistique de Corpus (JLC2015)*, 2015.

[9] S. M. Strassel, D. Miller, K. Walker, and C. Cieri, "Shared resources for robust speech-to-text technology," in *8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland, September 1-4, 2003*, 2003. [Online]. Available: http://www.isca-speech.org/archive/eurospeech_2003/e03_1609.html

[10] Y. Lim, Y. Zhu, S. G. Lingala, D. Byrd, S. Narayanan, and K. S. Nayak, "3d dynamic mri of the vocal tract during natural speech," *Magnetic resonance in medicine*, vol. 81, no. 3, pp. 1511–1520, 2019.

[11] Y. Laprie, R. Sock, B. Vaxelaire, and B. Elie, "Comment faire parler les images aux rayons x du conduit vocal ?" in *Actes du Congrès Mondial de la Linguistique Fran caise*, Berlin, Jul. 2014.

[12] B. Vaxelaire, "Etude comparee des effets des variations de debit-lent, rapide-surles parametres articulatoires, a partir de la cineradiographie (sujets francais)," Ph.D. dissertation, Strasbourg 2, 1993.

[13] [Online]. Available: http://www.cnrtl.fr/lexiques/morphalou/LMF-Morphalou.php

[14] L. Romary, S. Salmon-Alt, and G. Francopoulo, "Standards going concrete: from lmf to morphalou," in *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*. Association for Computational Linguistics, 2004, pp. 22–28.

[15] R. Sock, F. Hirsch, Y. Laprie, P. Perrier, B. Vaxelaire, G. Brock, F. Bouarourou, C. Fauth, V. Hecker, L. Ma, J. Busset, and J. Sturm, "DOCVACIM an X-ray database and tools for the study of coarticulation, inversion and evaluation of physical models," in *The Ninth International Seminar on Speech Production - ISSP'11*, Canada, Montreal, 2011.

[16] I. P. Association *et al.*, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[17] U. Musti, A. Toutios, V. Colotte, and S. Ouni, "Introducing visual target cost within an acoustic-visual unit-selection speech synthesizer," in *Proceedings of AVSP*, Volterra, Italy, 2011.

[18] I. Maddieson, *Patterns of sounds (Cambridge Studies in Speech Science and Communication)*. Cambridge university press, 1984.

[19] Y. Laprie and J. Busset, "Construction and evaluation of an articulatory model of the vocal tract," in *19th European Signal Processing Conference - EUSIPCO-2011*, Barcelona, Spain, Aug. 2011.

[20] Y. Laprie, B. Elie, A. Tsukanova, and P.-A. Vuissoz, "Centerline articulatorymodels of the velum and epiglottis for articulatory synthesis of speech," in *EUSIPCO 2018 - 26th European Signal Processing Conference, Sep 2018, Rome, Italy*, 2018.

[21] I. K. Douros, P.-A. Vuissoz, and Y. Laprie, "Comparison between 2d and 3d models for speech production: a study of french vowels," in *International Congress on Phonetic Sciences, 5-9 August, Melbourne, Australia*, 2019, accepted for publication.

[22] ——, "Acoustic impact of geometric approximation at the level of velum and epiglottis on French vowels," in *International Congress on Phonetic Sciences, 5-9 August, Melbourne, Australia*, 2019, accepted for publication.

[23] A. Tsukanova, I. K. Douros, A. Shimorina, and Y. Laprie, "Can static vocal tract positions represent articulatory targets in continuous speech? Matching static MRI captures against real-time MRI for the French language," in *International Congress on Phonetic Sciences, 5-9 August, Melbourne, Australia*, 2019, accepted for publication.