



Multimodal dataset of real-time 2D and static 3D MRI of healthy French speakers

Karyna Isaieva, Yves Laprie, Justine Leclère, Ioannis K Douros, Jacques Felblinger, Pierre-André Vuissoz

► To cite this version:

Karyna Isaieva, Yves Laprie, Justine Leclère, Ioannis K Douros, Jacques Felblinger, et al.. Multimodal dataset of real-time 2D and static 3D MRI of healthy French speakers. *Scientific Data*, 2021, 8 (1), pp.258. 10.1038/s41597-021-01041-3 . hal-03507532

HAL Id: hal-03507532

<https://hal.science/hal-03507532>

Submitted on 3 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



OPEN

DATA DESCRIPTOR

Multimodal dataset of real-time 2D and static 3D MRI of healthy French speakers

Karyna Isaieva¹✉, Yves Laprie², Justine Leclère^{1,3}, Ioannis K. Douros^{1,2}, Jacques Felblinger^{1,4} & Pierre-André Vuissoz¹

The study of articulatory gestures has a wide spectrum of applications, notably in speech production and recognition. Sets of phonemes, as well as their articulation, are language-specific; however, existing MRI databases mostly include English speakers. In our present work, we introduce a dataset acquired with MRI from 10 healthy native French speakers. A corpus consisting of synthetic sentences was used to ensure a good coverage of the French phonetic context. A real-time MRI technology with temporal resolution of 20 ms was used to acquire vocal tract images of the participants speaking. The sound was recorded simultaneously with MRI, denoised and temporally aligned with the images. The speech was transcribed to obtain phoneme-wise segmentation of sound. We also acquired static 3D MR images for a wide list of French phonemes. In addition, we include annotations of spontaneous swallowing.

Background & Summary

The investigation of the movement of speech articulators has a number of applications including study of speech production¹, speech recognition², as well as some medical applications: diagnosis and rehabilitation of abnormal speech and swallowing, study of orto-facial structures implicated in sleep apnoea syndrome³. Information on motion can be obtained using different methods including electromagnetic articulography (EMA)⁴, X-ray⁵ and ultrasound imaging⁶. Nowadays, magnetic resonance imaging (MRI) holds one of the leading positions as a data acquisition method in speech sciences^{7–10} due to its non-invasiveness and absence of long-term health hazards. Contrarily to other techniques such as ultrasound, which fails to visualise the articulators separated from the sensor by air, or EMA which provides only the sensors' trajectories glued on the upper vocal tract articulators, MRI succeeds to visualise the whole vocal tract.

However, MR imaging of a speaking person is a challenging problem due to the fast motion of articulators. One of the techniques allowing a reasonable spatio-temporal resolution of recorded speech, is cine-MRI^{11,12}. However, this method requires several identical repetitions of the same target utterance, which leads to artifacts in case of non-periodicity, and increases acquisition time. Real-time MRI allows high spatio-temporal resolution without repeating and is usually based on spoiled gradient echo sequences^{8,13,14}. Acquisition can be sped-up by usage of non-cartesian (generally undersampled) schemes which ensure good coverage of the k-space centre. This approach has been employed by several research groups to study speech. A spiral encoding scheme was applied in^{15,16} and was thereafter combined with sparse-SENSE constrained reconstruction methods^{8,17}. In^{18,19}, a radial encoding scheme was used together with a compressed SENSE reconstruction. The technique²⁰ makes use of radial sampling and the regularized nonlinear inversion reconstruction. Another approach, which does not necessarily assume a non-cartesian encoding, was used for dynamic 3D imaging of the vocal tract^{14,21}.

Since the technologies listed above are not easily available, data sharing could greatly accelerate research in the field. In this context, multiple databases exist for English speakers. Real-time MRI datasets, where 460 sentences were pronounced by 4 and 10 speakers, are presented in²² and¹⁶, respectively. The databases²³ and²⁴ which were acquired from 17 and 8 speakers respectively, include both real-time and 3D static MRI. An emotional speech dataset recorded from 10 speakers was published in¹⁵. Recently, an extremely rich dataset counting 75 English speakers was presented²⁵. However, MRI datasets representing other languages are very limited. 2D dynamic

¹Université de Lorraine, INSERM, IADI, Nancy, F-54000, France. ²Université de Lorraine, CNRS, Inria, LORIA, Nancy, F-54000, France. ³Oral Medicine Department, University Hospital of Reims, 45 rue Cognacq-Jay, 51092, Reims, Cedex, France. ⁴CIC-IT, INSERM, CHRU de Nancy, Nancy, F-54000, France. ✉e-mail: Karyna.isaieva@univ-lorraine.fr

Patient Code	Gender	Age	Height, cm	Weight, kg
P1	Male	25	175	70
P2	Male	22	180	73
P3	Male	41	178	68
P4	Female	45	174	78
P5	Male	20	175	70
P6	Male	27	172	73
P7	Female	24	178	59
P8	Female	25	168	68
P9	Female	33	162	52
P10	Female	30	170	53

Table 1. Relevant speakers' information.

MRI with temporal resolution of 7 frames per second of one female Portuguese speaker was published in²⁶. Static 3D MR images of five Japanese vowels pronounced by one male speaker are presented in²⁷. A dataset of 3D vocal tract shapes was published recently²⁸ for two German native speakers. A 2D dynamic with 3D static MRI database including 2 male French speakers was also acquired earlier²⁹. Nevertheless, the available data does not allow exhaustive investigation of these languages. Moreover, all the existing publicly available databases offering high spatio-temporal resolution dynamic MRI, exploit similar acquisition technologies due to the fact that they are acquired by the same research team. Availability of datasets of different qualities could serve to get better precision in some aspects.

In this work we report on a multi-modal MRI database consisting of 2D real-time and 3D static MR images of the vocal tract of 10 French speakers. The protocol used for the real-time MRI acquisitions for our dataset was successfully used by multiple groups in the context of the study of the articulators' motion^{30–33}. While performing investigations on the vocal tract organs, it is crucial to consider the diversity of their movements during speech production. Standard French language includes 35 phonemes (18 consonants, 14 vowels, 3 semi-vowels) which form 1290 diphones³⁴ and many complex consonant clusters. To cover this variability as much as possible, a corpus was previously developed²⁹. The corpus allows to explore numerous phenomena specific to the French language such as nasal vowels, uvular /ʁ/³⁵, French /y/, short /ʉ/, and strong anticipation of labial features³⁶. The dataset includes annotations of the speech and of spontaneous swallowing and will thus provide researchers with data having a good coverage of the French phonetics to further explore French speech production and physiological processes taking place in the vocal tract vicinity.

Methods

Participants and speech task. The participants were 5 male and 5 female native French speakers (aged 29 ± 8 years) without any speech or hearing problems. Presence of any metal in the vocal tract vicinity, which may generate susceptibility artifacts, was also an exclusion criterion. Relevant patient characteristics are listed in Table 1. A set of mid-sagittal images demonstrating speakers' anatomy is presented in Fig. 1.

All participants provided written informed consent, including written permission to publish the materials of this experiment. The data was recorded under the approved ethical protocol "METHODO" (ClinicalTrials.gov Identifier: NCT02887053). The study was approved by the institutional ethics review board (CPP EST-III, 08.10.01).

The previously designed corpus²⁹ was presented in form of a pdf-file (see Supplemental Materials) which was projected on a screen in the MRI room so that a speaker could read it during the experiment without difficulty. The corpus included two parts. The first one served for the acquisition of dynamic 2D data and included 77 sentences which were constructed to provide an almost-exhaustive coverage of the French phonetic contexts of vowels /i, a, u, y/ and some nasal vowels selected from / $\tilde{\alpha}$, \tilde{o} , $\tilde{\epsilon}$ /. Several levels of criteria were used to guide the manual construction of those sentences. After the insertion of a new sentence the first level of criteria evaluated was the number of VV for all the vowels, the number of CV for C in /p, t, k, f, s, \int , l, β , m, n/ and V in /i, a, u/ plus /y/, the number of VC with C as a coda and C in /l, β , n, m/ and V in /i, a, u, y, e, ϵ , o, ɔ /, the consonant clusters C1C2V with C in /p, t, k, b, d, g, f/, C2 in / β , l/ and V in /a, i, u, y/ (the other CCV following the same pattern with /s, \int , v/ are rare in French), and VC, of C in a coda, 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 French sentences and therefore words not corresponding to the target 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 considering groupings of close vowels. There are 6 groups of vowels (/i, e/, ϵ , a/, u, o, ɔ /, /y, ø /, / œ , ə / and nasal vowels / $\tilde{\alpha}$, \tilde{o} , $\tilde{\epsilon}$ /. This provides a second level of evaluation which helps required words to build well-constructed sentences. Other words required to form well-constructed sentences provide additional contexts with the remaining vowels. All the words (except "cartoons" and "squaw") are of French origin to avoid any ambiguity of pronunciation. The sentences were divided on groups of 3–5 to ensure comfortable duration of a session. The speech task for 2D real-time acquisition is presented in Online-only Table 1.

The second part comprised 3D static data acquisition and included 5 silent acquisitions with different positions of the tongue (against the upper teeth, against the lower teeth between the incisors, retroflex and deep

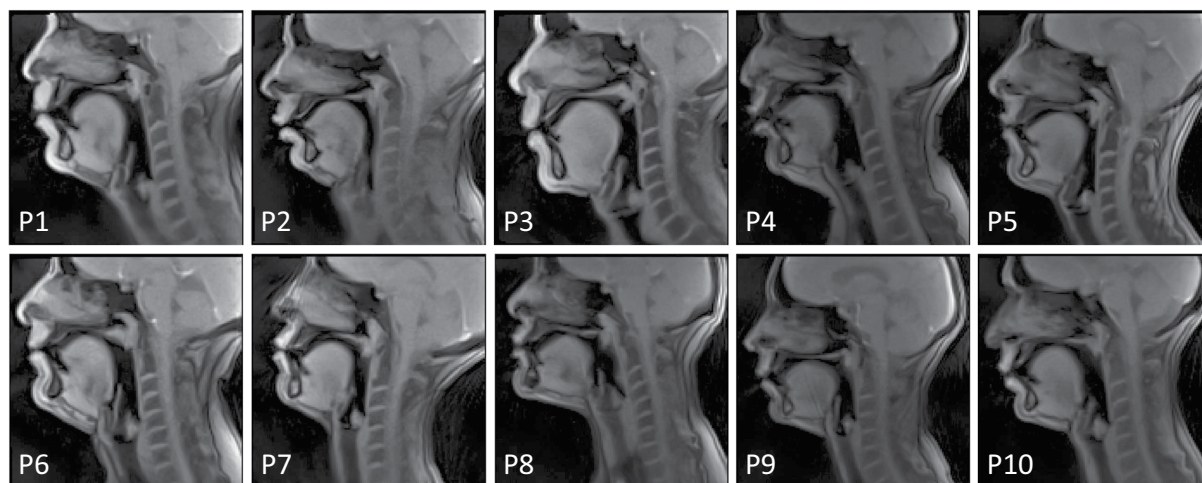


Fig. 1 Examples of real-time images of all ten speakers pronouncing /u/ (“filou” from the first sentence).

retroflex), 14 vowels and 12 consonants (each in context of several vowels). The full set of the phonemes and positions included into the speech task for 3D static acquisitions, is listed in Online-only Tables 2 and 3. The first three positions were designed to help the registration of teeth within MRI data. The participants were asked to keep the same position before and up to the end of the MRI noise in case of silent positions or vowels. For a consonant C in context of a vowel V they were instructed to phonate V until the end of the countdown of an MRI operator who then started a sequence, then to keep the articulation of C until the end of the MRI noise and then phonate V again (the consonant production in this case is called blocked articulation). This helped speakers reach the expected articulatory position for each of the consonants articulated within a given vowel context. Duration of each sequence for the static 3D data acquisition was chosen to be 7 seconds, as a compromise between the volunteers’ comfort and the image quality. While sometimes it was not very natural to keep the same position for such time, especially in the case of plosive consonants, the task appeared to be absolutely feasible.

The presentation was sent to the volunteers before the experiment so that they had at least one day to get familiar with the sentences which sometimes sound strange even if they are all well-formed French sentences. Additionally, the speakers were carefully instructed directly before the acquisition to guarantee correct understanding of the task.

Data acquisition and alignment. The MRI data was recorded at Nancy Central Regional University Hospital on a Siemens Prisma 3 T scanner (Siemens, Erlangen, Germany). The speakers were in supine position and the Siemens Head/Neck 64 coil was used.

For the 2D real-time we used radial RF-spoiled FLASH sequence¹³ with TR = 2.22 ms, TE = 1.47 ms, FOV = 22.0 × 22.0 cm, flip angle = 5°, and slice thickness was 8 mm. Pixel bandwidth was 1670 Hz/pixel. Image size was 136 × 136, and in-plane resolution was 1.6 mm. Images were recorded at a frame rate of 50 frames per second and reconstructed with a nonlinear inverse technique presented in¹³. This method represents a formulation of a nonlinear optimisation problem with respect to both image and coil sensitivity maps which is solved iteratively with regularized Gauss–Newton method. The protocol used for dynamic data acquisition differs from the protocols of the published publicly available databases, and thus the quality is also somewhat different. Radial encoding trajectories are shorter than spiral ones, so that the repetition time is lower in our case (2.22 ms comparing to 6.004 ms in the latest dataset²⁵) which probably decreases the manifestation of the off-resonance effects. Also, some residual aliasing artifacts can be remarked in case of the datasets¹⁶ and¹⁵. However, due to the larger slice thickness (8 mm comparing to 6 mm), the partial volume effects are more pronounced in our case. Protocol¹⁷ proposes higher temporal resolution (83 frames per second^{24,25}), while the protocol used in our case offers higher in-plane spatial resolution.

For 3D static data, we used 3D VIBE with TR = 3.8 ms, TE = 1.55 ms, FOV = 22.0 × 20.0 cm², flip angle = 9°, slice thickness was 1.2 mm, and in-plane resolution was 0.69 × 0.76 mm². Pixel bandwidth was 445 Hz/pixel. Image size was 320 × 290 with 36 slices. Acceleration factor was iPAT = 3. Each sequence had duration of 7 seconds. Examples of the resulting 3D images are given in Fig. 2.

Audio was recorded at a sampling frequency of 16 kHz inside the MRI scanner by using a FOMRI III optoacoustics fibre-optic microphone (FOMRI III, Optoacoustics Ltd., Mazor, Israel) placed in the scanner. The volunteers wore earplugs to be protected from the scanner noise, but were still able to communicate orally with the experimenters via an in-scanner intercom system. Since the sound was recorded at the same time with the MRI acquisition, some additional noise is present in the audio signal. In order to suppress the noise, we used the algorithm proposed in³⁷. The algorithm relies on the hypothesis that a noisy sound represents a gaussian mixture of two components (voice and MRI noise in this case), and the decomposition is based on an expectation-maximisation algorithm. The components are characterized by some source and resonator spectral features. A clear speech recording, which was required by the algorithm, was done separately for each speaker just before the dynamic MRI acquisition starts, with the same patient and microphone positions.

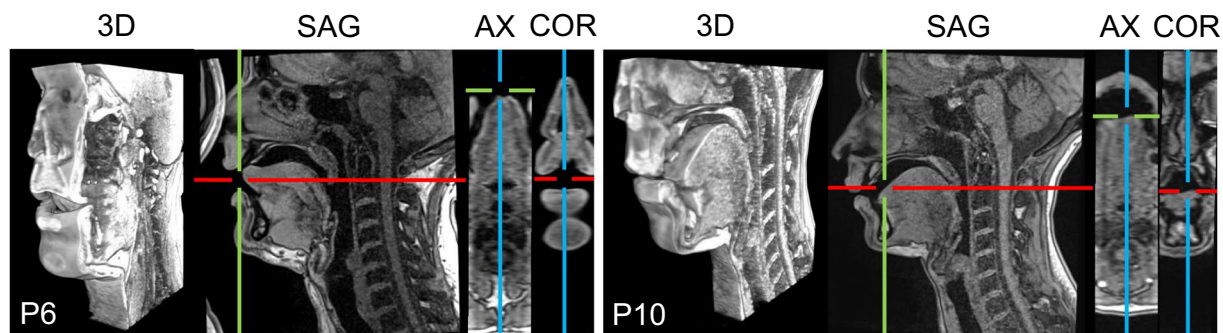


Fig. 2 Examples of 3D static images: P6 pronouncing /l/ in context of /a/, and P10 pronouncing /ã/. The 3D volumes were cropped for better visibility.

In order to align the sound with MRI data, we used Signal Analyzer and Event Control system (SAEC) which had previously been designed³⁸. We applied it to record timestamps of the reconstruction start and the sequence stop events and send them to a channel of the opto-acoustic system which allows MRI transistor-transistor logic (TTL) commands recording. This enabled automatic synchronisation of the images and the sound. However, during manual examination it was found that the images are somewhat shifted with respect to the sound after the automatic alignment. A shift of 2 images (40 ms) was explained by the application of the temporal median filter of width 5 (100 ms) during reconstruction. Nevertheless, due to some temporal variations of the TTL signals and/or sound reception, probably caused by USB jitter, the temporal shift between the images and the sound slightly varied from series to series. Thereby, the shifts were required to be manually defined for each series by comparison of sound amplitude and/or spectrograms with MR images. From a practical point of view, for this purpose we selected the events which are clearly visible on both the acoustic signal and the images: onsets of the sounds /p/ and /b/. This event corresponds to the time moment where the lips contact occurs and results in an abrupt and considerable acoustic signal weakening. In most of the recorded sequences there were /p/ and /b/ near the recording extremities. If this was not the case, the fallback solution was to use the contact between the tongue tip and teeth alveoli for /t/ and /d/ which also corresponds to a strong weakening of the signal. Thus, the full pipeline consisted of four steps. 1. Data acquisition. 2. Automatic alignment using the TTL timestamps. 3. Sound cropping, so that it fits the MRI acquisition time interval, which was necessary for the denoising, and the denoising itself. 4. Manual shift determination and re-alignment. The illustration of the different steps of the alignment routine is given in Fig. 3.

Transcription of the continuous speech corpus. Speech transcription is the temporal sentence-wise, word-wise or phoneme-wise segmentation of an audio recording. Together with the synchronisation, it provides correspondence between images and pronounced phonemes which is helpful for many practical applications. Even though the speech task was pre-defined, some speakers made mistakes or repeated some syllables or words. To take these deviations with respect to the expected pronunciation into account, each recording was inspected by an investigator and all the hesitations or repetitions were included to the text used for the alignment. The transcription was done in two steps. First, the sentences start and end were manually annotated from the sound, which was denoised and cut to fit the MRI acquisition time interval (see Fig. 3). This was done using Transcriber 1.5.2 (<http://trans.sourceforge.net/en/presentation.php>). This software generates .trs files which include the timestamps and the corresponding text. Text annotations and audio signal were synchronized by a forced alignment automatic speech recognition system Astali (<http://ortolang108.inist.fr/astali/>) trained on French. It used the .trs text annotations together with the denoised sound to perform the temporal segmentation (both word-wise and phoneme-wise). The phonemes are stored inside a file using SAMPA phonetic annotation system³⁹.

Swallowing detection. Swallowing is defined as a series of mechanisms allowing transportation of food, drinks or saliva to the stomach. This mechanism occurs in four steps: (1) A pre-emptive phase, with lip closure; (2) an oral phase, which corresponds to the bolus transportation from the front to the posterior area of the oral cavity, in order to reach the pharynx; (3) a pharyngeal phase, where the bolus continues its way through the pharynx, and (4) an oesophageal phase allowing the bolus to penetrate into the stomach.

The oral phase is particularly interesting, owing to the complexity of the muscle contractions and anatomical movements achieved during this step. During this phase, which lasts about one second, the mandible is stabilized when the teeth are in contact and in maximal intercuspation occlusion (MIO), after lip closure. At the same time, the tongue initiates a propulsion movement which begins in the anterior hard palate, and then performs a second contraction to propel the bolus at the rear up to the pharyngeal areas. Following this contraction, the oropharynx is closed to prevent bolus penetration into the upper airways (UA). The bolus can then continue its way to the pharynx.

Swallowing pathologies are numerous, and imaging devices remain limited to study in real-time physiological movements of the anatomical structures of the upper airways. In order to facilitate this observation and identify the oral phase of swallowing, we propose a protocol to determine the start and end positions of swallowing on our MRI images in real time. (1) The image counting starts when the apex of the tongue touches the hard palate. However, for some images recorded during speech, the tongue might already be in this position while swallowing

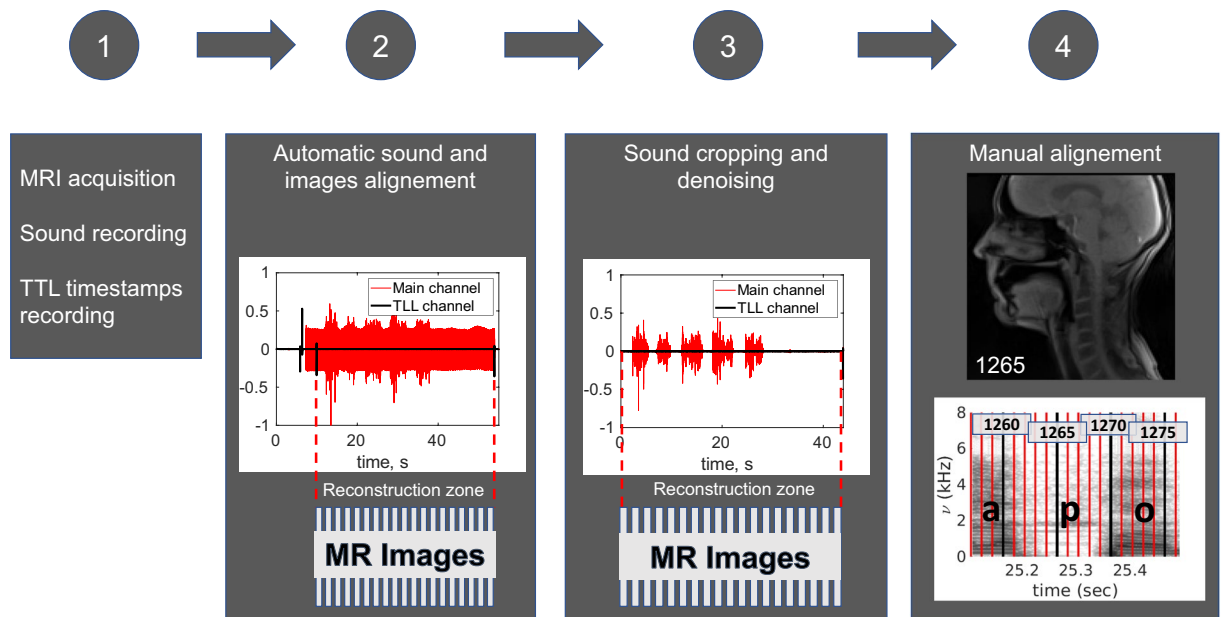


Fig. 3 Data processing routine illustrated with the data of P9. The figure should be read from left to right, and each item corresponds to a processing stage. The first item lists the acquired data, second and third items illustrate sound alignment and denoising results, and the last item aims to explain the manual alignment principles. The articulators' position on the MR image 1265 corresponds to the onset of /p/ (the lips have just come in contact). The plot below the MR image shows a spectrogram of the sound with vertical lines denoting centres of MRI acquisition intervals (each 5-th line is black for better readability). From the spectrogram, the image 1265 is supposed to correspond to the middle of /p/, which is not the case. The sound is haste by approximately 3 images = 60 ms and thus the sound and the images should be re-aligned.

begins. In these cases, we took as a landmark the most anterior contact of the dorsal part of the tongue with the hard palate. (2) At the beginning of the oral phase of swallowing, the elevation of the hyoid bone is observed. The end of the oral phase has been described as the moment when the space between the tongue and the velum and/or the soft palate reappears. At the same time, the hyoid bone returns to its initial position, and the oropharynx relaxes to allow the ventilatory flow to recover.

Data Records

The data is available on figshare⁴⁰. Each of the 10 folders contains the data of a speaker. To summarize, a folder (except the second one) contains 16 dynamic and 76 static series. Each of the dynamic series counts 1800 to 2200 frames and has duration about 1 minute, which results to overall 34800 images (approximately 15 minutes) per subject. Each static series consists of 36 slices.

The dataset is organized as follows: the root directory contains ten speakers' folders with names "PXX" (XX here is a patient code). The speaker data is divided into three folders: DCM_2D, DCM_3D and OTHER. Inside DCM_2D and OTHER folders there are 16 subfolders with names SYY (YY is a series number) which correspond to the different dynamic acquisitions. Files stored inside the subfolders of DCM_2D are DICOM files with the MRI 2D dynamic data, and files contained inside subfolders of the OTHER folder are the corresponding denoised cropped sound, TEXT_ALIGNMENT_PXX_SYY.trs file (sentence alignment), TEXT_ALIGNMENT_PXX_SYY.textgrid file (alignment of words and phonemes), SWALLOWING_PXX_SYY.trs which has the same format as "sentence alignment" and contains swallowing timestamps, and an example video VIDEO_PXX_SYY.avi generated with the provided code (after compression).

The 3D static data can be found inside the subfolders of the DCM_3D folder. The name of those subfolders corresponds to the target phoneme and possibly its vowel context, for instance "tu" for /t/ in the vocalic context of /u/. There are also three static positions to help determine the position of teeth which are not visible on MRI scans. They are denoted as UP (for the tongue touching the upper incisors), DOWN (for the tongue touching lower incisors) and CONTACT (for incisors in contact). The latter was used to check consistency of the teeth positions (as determined by UP and DOWN).

The correspondence between the folder names and the speech task performed in frames of an MRI series, is given in the Online-only Tables 1–3. The dataset structure is also illustrated in Fig. 4.

The text alignment data are the Transcriber sentence annotation files (.trs files) which include pronounced text and the timestamps of start and end of each sentence on the one hand, and word and phoneme segmentation files (.textgrid Pratt files)⁴¹ on the other hand. The code which allows reading these files is provided.

For P2, 3D static images of consonants phonation were not acquired because of technical reasons, thus only dynamic 2D data and 3D static images of vowels and silent positions were included. All other speakers performed the full list of acquisitions.

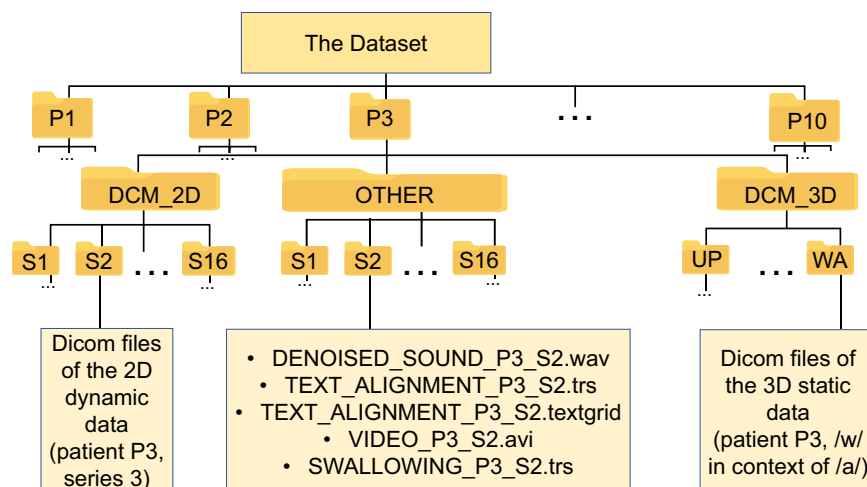


Fig. 4 The dataset structure illustrated with the example of the second dynamic series and the 3D static image of the consonant /w/ in context of /a/ of the P3 data.

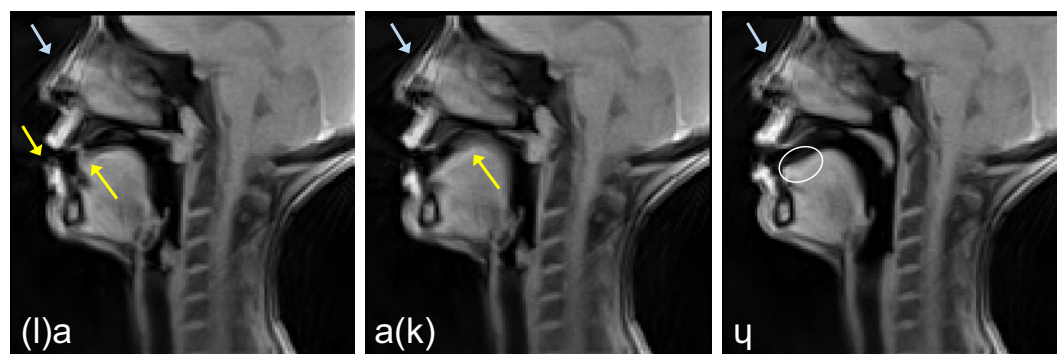


Fig. 5 Artifacts and imperfections that can occur illustrated on the images of P7. Blue arrows point to an aliasing artifact, which, however, does not affect the quality of the vocal tract imaging. Yellow arrows point to motion artifacts that take place in case of rapid change of the articulators' position (like the transition from /l/ to /a/, and from /a/ to /k/). The white ellipse points to a tongue region with slightly lower intensity which is caused by relatively large slice thickness and tongue shape variations in the left-right direction (partial volume effects).

Technical Validation

The dynamic 2D images were visually inspected by the researchers. In general, the images quality was good. Some minor artifacts can be observed as it is pointed out in Fig. 5. Some partial volume effects occurred because of relatively large slice thickness (8 mm). Despite the temporal filter, some residual radial aliasing artifacts were observed. Very fast articulatory gesture can lead to blurring, for example when the tongue tip approaches the alveolar region for /t/, /d/ or /l/.

The video-sound alignment was verified by a researcher having more than 20 years of experience in the field (author Y.L.). This was done by cross checking of the sound track and images using the videos included to the database as described in Methods section. In the case a misalignment found, a proper shift was applied to the sound.

The text alignment was checked by authors K.I., P.-A.V. and Y.L. by comparing the sound, the images and the text. Some errors caused by the fact that certain speakers made many pronunciation errors and/or hesitations were generally corrected, however some residual errors can still have place. The list of the corrected mistakes is presented in Online-only Table 4. An experienced reader can see that there are some minor phoneme-wise sound segmentation errors (order of 1 frame which corresponds to 20 ms) in the case of plosive consonants (explained by the fact that almost no sound is produced). We chose to keep the automatic annotations.

The vocal tract shape during 3D static data acquisition, which should correspond to a required phoneme, was visually inspected directly during the course of the experiment by authors K.I. and P.-A.V. In case of obviously wrong positions, the data was reacquired up to three times. The resulting 3D images were checked by author Y.L. The cases of phoneme/image inconsistency for the static 3D data are given in the Table 2. In addition to individual comments given in the Table 2, here are some general trends. Blocked articulations (freezing a position just before producing a consonant) is not a natural gesture in speech production. It is especially difficult to control the velum position since there is no acoustic feedback. This explains why the velum is in a lower position in some cases where it is expected to be in a higher position, for instance stop consonants. Some speakers who were not familiar

Subjects	Comments
P1	Good images. Very extreme retroflex shape.
P2	Only some images were recorded because of a technical problem.
P3	The subject did not understand instructions correctly (no contact for stops, lips not closed for /p/) but sustained sounds, i.e. vowels and fricatives, are correct.
P4	The overall quality of images is not very good. Oral vowel shapes are correct, and in to a lesser extent fricatives. Many blocked articulations do not exhibit expected features (contact for stops, velum position...).
P5	Very good images. /li/ was not articulated correctly.
P6	Good images. Slight move for /a/ and /mɜ/.
P7	The contact between the fixed and the mobile articulators is not reached for several stop consonants. The velum is in the upper position for /m/, under-anticipation of the tongue position for /p/, exaggeration of the tongue tip position between upper and lower teeth for /t/, no contact at the place of articulation of some /k/. The articulators' positions are, in general, not very natural.
P8	Strong forward position of the mandible. Instead of being in the upper position the velum is in the lower position for many oral articulations. Several stop articulations without contact between the fixed and mobile articulators (for /p/ and /t/).
P9	The subject did not understand instructions despite several trials. Some vowels are articulated with the mouth closed and some tongue shapes are very unusual for /k/ which has not been articulated correctly.
P10	Good images. The velum is sometimes in the lower position (/a/ and /la/ for instance) for oral sounds.

Table 2. Evaluation of articulatory shapes produced.

with phonetics were unable to respect the instructions and despite several explanations, they did not understand how to do reach the expected articulatory positions, especially those corresponding to stop consonants. For the same reason, subjects reached a far better articulatory position for phonated items, (i.e. vowels and fricatives) simply because the condition is more natural. The strong MRI noise probably strengthened the Lombard effect for vowels which slightly changed the articulation. We decided to keep this inconsistent data, since it can still have some applications, i.e. as dataset augmentation in case of machine learning.

Code availability

A MATLAB code which generates a video from the dynamic data is available on GitHub <https://github.com/IADI-Nancy/ArtSpeech>. This code also allows reading alignment annotation .trs and .textgrid files, and reading audio and dicom files. It uses mPRAAT third party toolbox which is available on bbTomas GitHub <https://github.com/bbTomas/mPraat>. The code was tested on Linux distribution of MATLAB2018b and MATLAB2020a. A Python toolbox for .textgrid files parsing also exists⁴².

Received: 26 January 2021; Accepted: 25 August 2021;

Published online: 01 October 2021

References

- Elie, B. & Laprie, Y. Simulating alveolar trills using a two-mass model of the tongue tip. *J. Acoust. Soc. Am.* **142** (2017).
- Douros, I. K., Katsamanis, A. & Maragos, P. Multi-View Audio-Articulatory Features for Phonetic Recognition on RTMRI-TIMIT Database. in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 5514–5518 (2018).
- Kim, Y.-C. *et al.* Real-time 3D magnetic resonance imaging of the pharyngeal airway in sleep apnea. *Magn. Reson. Med.* **71**, 1501–1510 (2014).
- Katz, W. F., Mehta, S., Wood, M. & Wang, J. Using electromagnetic articulography with a tongue lateral sensor to discriminate manner of articulation. *J. Acoust. Soc. Am.* **141**, EL57–EL63 (2017).
- Badin, P. Fricative consonants: acoustic and X-ray measurements. *J. Phon.* **19**, 397–408 (1991).
- Fabre, D., Hueber, T., Girin, L., Alameda-Pineda, X. & Badin, P. Automatic animation of an articulatory tongue model from ultrasound images of the vocal tract. *Speech Commun.* **93**, 63–75 (2017).
- Lingala, S. G., Sutton, B. P., Miquel, M. E. & Nayak, K. S. Recommendations for real-time speech MRI. *J. Magn. Reson. Imaging* **43**, 28–44 (2016).
- Zhao, Z., Lim, Y., Byrd, D., Narayanan, S. & Nayak, K. S. Improved 3D real-time MRI of speech production. *Magn. Reson. Med.* **85**, 3182–3195 (2021).
- Gomez, A. D., Stone, M. L., Woo, J., Xing, F. & Prince, J. L. Analysis of fiber strain in the human tongue during speech. *Comput. Methods Biomech. Biomed. Engin.* **23**, 312–322 (2020).
- Carignan, C. *et al.* Analyzing speech in both time and space: Generalized additive mixed models can uncover systematic patterns of variation in vocal tract shape in real-time MRI. *Lab. Phonol. J. Assoc. Lab. Phonol.* **11** (2020).
- Masaki, S. *et al.* MRI-based speech production study using a synchronized sampling method. *J. Acoust. Soc. Japan* **20**, 375–379 (1999).
- Woo, J., Xing, F., Lee, J., Stone, M. & Prince, J. L. A spatio-temporal atlas and statistical model of the tongue during speech from cine-MRI. *Comput. Methods Biomech. Biomed. Eng. Imaging & Vis.* **6**, 520–531 (2018).
- Uecker, M. *et al.* Real-time MRI at a resolution of 20 ms. *NMR Biomed.* **23**, 986–994 (2010).
- Fu, M. *et al.* High-resolution dynamic speech imaging with joint low-rank and sparsity constraints. *Magn. Reson. Med.* **73**, 1820–1832 (2015).
- Kim, J. *et al.* USC-EMO-MRI corpus: An emotional speech production database recorded by real-time magnetic resonance imaging. in *International Seminar on Speech Production (ISSP)*, Cologne, Germany **226** (2014).
- Narayanan, S. *et al.* Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). *J. Acoust. Soc. Am.* **136**, 1307–1311 (2014).
- Lingala, S. G. *et al.* State-of-the-Art MRI Protocol for Comprehensive Assessment of Vocal Tract Structure and Function. in *Interspeech* 475–479 (2016).
- Burdumy, M. *et al.* Acceleration of MRI of the vocal tract provides additional insight into articulator modifications. *J. Magn. Reson. Imaging* **42**, 925–935 (2015).

19. Burdumy, M. *et al.* One-second MRI of a three-dimensional vocal tract to measure dynamic articulator modifications. *J. Magn. Reson. imaging* **46**, 94–101 (2017).
20. Uecker, M., Hohage, T., Block, K. T. & Frahm, J. Image reconstruction by regularized nonlinear inversion—joint estimation of coil sensitivities and image content. *Magn. Reson. Med.* **60**, 674–682 (2008).
21. Fu, M. *et al.* High-frame-rate full-vocal-tract 3D dynamic speech imaging. *Magn. Reson. Med.* **77**, 1619–1629 (2017).
22. Narayanan, S. *et al.* A multimodal real-time MRI articulatory corpus for speech research. in *Twelfth Annual Conference of the International Speech Communication Association* (2011).
23. Sorensen, T. *et al.* Database of Volumetric and Real-Time Vocal Tract MRI for Speech Science. in *INTERSPEECH* 645–649 (2017).
24. Töger, J. *et al.* Test-retest repeatability of human speech biomarkers from static and real-time dynamic magnetic resonance imaging. *J. Acoust. Soc. Am.* **141**, 3323–3336 (2017).
25. Lim, Y. *et al.* A multispeaker dataset of raw and reconstructed speech production real-time MRI video and 3D volumetric images. *arXiv Prepr. arXiv2102.07896* (2021).
26. Teixeira, A. *et al.* Real-time mri for portuguese. in *International Conference on Computational Processing of the Portuguese Language* 306–317 (2012).
27. Kitamura, T., Takemoto, H., Adachi, S. & Honda, K. Transfer functions of solid vocal-tract models constructed from ATR MRI database of Japanese vowel production. *Acoust. Sci. Technol.* **30**, 288–296 (2009).
28. Birkholz, P. *et al.* Printable 3D vocal tract shapes from MRI data and their acoustic and aerodynamic properties. *Sci. Data* **7**, 1–16 (2020).
29. Douros, I. *et al.* A Multimodal Real-Time MRI Articulatory Corpus of French for Speech Research. in *Interspeech* (2019).
30. Labrunie, M. *et al.* Automatic segmentation of speech articulators from real-time midsagittal MRI based on supervised learning. *Speech Commun.* **99**, 27–46 (2018).
31. Iltis, P. W. *et al.* Simultaneous dual-plane, real-time magnetic resonance imaging of oral cavity movements in advanced trombone players. *Quant. Imaging Med. Surg.* **9**, 976 (2019).
32. Krohn, S. *et al.* Multi-slice real-time MRI of temporomandibular joint dynamics. *Dentomaxillofacial Radiol.* **48**, 20180162 (2019).
33. Isaieva, K. *et al.* Measurement of Tongue Tip Velocity from Real-Time MRI and Phase-Contrast Cine-MRI in Consonant Production. *J. Imaging* **6**, 31 (2020).
34. Boeffard, O., Cherbonnel, B., Emerard, F. & White, S. Automatic segmentation and quality evaluation of speech unit inventories for concatenation-based, multilingual PSOLA text-to-speech systems. in *Third European Conference on Speech Communication and Technology* (1993).
35. Hannahs, S. J. French phonology and L2 acquisition. *French Appl. Linguist.* **16**, 50–74 (2007).
36. Ladefoged, P. & Johnson, K. *A course in phonetics*. (Cengage learning, 2014).
37. Ozerov, A., Vincent, E. & Bimbot, F. A general flexible framework for the handling of prior information in audio source separation. *IEEE Trans. Audio. Speech. Lang. Processing* **20**, 1118–1133 (2012).
38. Odille, F. *et al.* Noise Cancellation Signal Processing Method and Computer System for Improved Real-Time Electrocardiogram Artifact Correction During MRI Data Acquisition. *IEEE Transactions on Biomedical Engineering* **54**, 630–640 (2007).
39. Fourcin, A. J. Progress Overview for the SAM Project. in *First European Conference on Speech Communication and Technology* (1989).
40. Isaieva, K. *et al.* Multimodal dataset of real-time 2D and static 3D MRI of healthy French speakers. *figshare* <https://doi.org/10.6084/m9.figshare.c.5270387> (2021).
41. Boersma, P. & Weenink, D. Praat: Doing phonetics by computer (Version 5.3.82)[Computer software]. *Amsterdam Inst. Phonetic Sci.* (2012).
42. Buschmeier, H. & Wlodarczak, M. TextGridTools: A TextGrid processing and analysis toolkit for Python. in *Tagungsband der 24. Konferenz zur elektronischen sprachsignalverarbeitung (ESSV 2013)* (2013).

Acknowledgements

This research was funded by the projects ArtSpeech and Full3DTalkingHead of ANR (Agence Nationale de la Recherche), France; CPER “IT2MP”, “LCHN” and FEDER. We thank Claire Dessale and the CIC-IT team for their help with the legal issues and the team of MRI technicians of the Nancy Brabois Central Regional University Hospital.

Author contributions

All authors listed have made a substantial contribution to the work and approved it for publication. Y.L. designed the corpus, Y.L., P.-A.V. and K.I. designed the experiment, P.-A.V., K.I. and I.D. performed the experiment, K.I., P.-A.V. and J.L. processed the data, Y.L., P.-A.V. and K.I. performed the technical validation, J.F. took part in design of the ethical protocol, J.F., P.-A.V. and Y.L. acquired the funding, K.I. prepared the original draft of the paper, P.-A.V., Y.L., J.L., I.D. and J.F. reviewed and edited the text.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-021-01041-3>.

Correspondence and requests for materials should be addressed to K.I.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

© The Author(s) 2021