Effect of head posture on phonation of French vowels
Ioannis K Douros, Pierre-André Vuissoz, Yves Laprie

To cite this version:
Ioannis K Douros, Pierre-André Vuissoz, Yves Laprie. Effect of head posture on phonation of French vowels. ICPhS 2019 - Proceedings of International Congress of Phonetic Sciences, Aug 2019, Melbourne, Australia. hal-02180486
EFFECT OF HEAD POSTURE ON PHONATION OF FRENCH VOWELS

Ioannis K. Douros¹,², Pierre-André Vuissoz², Yves Laprie¹

¹Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France,
²Université de Lorraine, INSERM U1254, IADI, F-54000 Nancy, France,
ioannis.douros@loria.fr, pa.vuissoz@chru-nancy.fr, yves.laprie@loria.fr

ABSTRACT

In this work we study how the head position in relation to that of the cerebral vertebrae affects phonation of five vowels of French. Our aim is to examine how this affects acoustic parameters, draw conclusions to adapt acoustic simulations and vocal tract geometric models for articulatory synthesis, verify whether the effect of posture is observable and evaluate the changes in the formant frequencies. We use MRI data to capture the shape of the vocal tract in three different positions per vowel, with simultaneous speech signal recording. We use acoustic simulations with no underlying plane wave hypothesis to see how the acoustic wave propagates. We simulate three head postures (up, middle/natural and down) for every vowel, compare the results with the original speech signal and validate the existence of difference both in the acoustic domain in terms of formant frequencies and in the articulatory domain, especially in the pharyngeal cavity.

Keywords: Head position, Speech phonation, Acoustic simulation, French vowels, MRI

1. INTRODUCTION

In recent years, articulatory synthesis has aroused considerable scientific interest [20, 13, 3]. This is due in particular to recent advances in the field of acoustic simulations (the possibility of taking into account the complexity of the vocal tract [7], or better coordinating glottis opening and supraglottal cavities for consonants [8] for instance) that open up new possibilities in terms of the quality of the speech produced. These advances also rely on articulatory data that are more precise and accurate thanks to the emergence of Magnetic Resonance Imaging (MRI).

The geometric shape of the vocal tract can be derived from images of a film or be generated at each time point by an articulatory model [14, 1, 2, 12]. Apart from models based on geometric primitives, the models are generally constructed using factor analysis [1] applied to a corpus of two or three-dimensional MRI images of the vocal tract. One of the issues raised by articulatory models derived from medical images of one subject is the validity of the model for other speakers.

Maeda [15] developed a procedure that consists of separately adapting the sizes of the mouth and pharynx. This distinction between the two parts of the vocal tract is based on the observation that the size of the pharyngeal and mouth cavities depends on both gender and, predictably, age. A slightly more elaborate approach adapted to a more complex articulatory model has been developed in [19] and tested with a model developed on one speaker which was used to fit mid-sagittal vocal tract shapes of another speaker.

In the first works dedicated to articulatory modeling carried out with X-ray images, speakers were sitting and adopting a fairly natural position to produce speech. More recent articulatory data are acquired with MRI in a supine position, and the head posture is largely dictated by the position of the MRI antenna and foam, which is used to prevent it from moving during acquisitions. Consequently, the position of the head is not natural, and above all it can vary significantly between two acquisitions, and a fortiori between two machines. The articulatory models that can be derived from those data implicitly incorporate the head posture. Experiments carried out to fit dynamic MRI data of one speaker with an articulatory model built for a reference speaker have shown that the adaptation procedure fails to approximate the whole vocal tract. More precisely, it turned out that the tongue can be fitted fairly well, which is not the case for the pharyngeal cavity whose width deviates from what is predicted by the model. In addition, this deviation is likely to change the acoustic properties of speech, and formant frequencies in particular.

For this reason we are interested in assessing the geometrical and acoustic consequences of head posture in speech production from MRI data by using direct formant estimation and acoustic simulations.
2. MATERIALS AND METHODS

The pipeline of this study is 1) MRI data and speech signal acquisition, 2) data processing, and 3) acoustic simulations.

2.1. Data acquisition

For the purpose of this study, we used the MRI data part of a study approved by an ethics committee and the subject gave written informed consent (ClinicalTrials.gov identifier: NCT02887053). The subject retained for the data acquisition is a healthy male French native speaker at the age of 57, without any reported speaking or hearing problems.

The MRI data was acquired on a Siemens Prisma 3T scanner (Siemens, Erlangen, Germany) with gradient of 80 mT/m amplitude and 200 mT/m/ms slew rate. We used the 3-dimensional cartesian gradient echo RF spoiled vibe sequence (TR = 3.8 ms, TE = 1.41, FOV = 235 × 260 mm, flip angle = 7.5 degrees) for the acquisition. The pixel bandwidth is 445 Hz/pixel with an image resolution of 320 × 290. Scan slice thickness is 1.2 mm and the number of slices is 36. The pixel spacing is 0.8125 and the acceleration factor is 3 iPAT.

The subject’s vocal tract is imaged while he lay supine in the MRI scanner with his head in three positions: up, middle/normal and down. Additional phantoms were used to stabilize the head in each position and help the subject reach and maintain the two extreme. Between the different head positions, the phantom position was re-initialized to ensure the maximum possible lengthening and shortening of the vocal tract. However, there were limitations to how far the subject could tilt his head due to the coil.

The recording time for the subject, including initializations, calibrations and pauses between head positions and phonemes, was 2 hours. Audio is recorded at a sampling frequency of 16 kHz inside the MRI scanner using FOMRI III (Optoacoustics, Or Yehuda, Israel) fiber optic microphone. The subject pronounces each vowel to be recorded twice before the MRI acquisition starts and once as the MRI machine is on. The latter repetition takes around 7.4 s of sustained phonation.

The subject wears ear plugs for protection from the scanner noise, but is still able to communicate orally with the experimenters via an in-scanner intercom system.

Since the sound is recorded at the same session of the MRI acquisition, there is additional noise in the audio signal. Details on how we treated this issue are described in the experiment section.

2.2. Data processing

For the processing of the MRI images, we used ITK-SNAP software [23]. This software was specifically designed for medical image segmentation. It employs some popular algorithms like geodesic active contours [5] and region competition [24] for automatic 2D/3D segmentation. It also offers some options for manual segmentation like polygon based tools and various types of paint brushes.

2.3. Acoustic simulations

For the purposes of our experiments, we used k-wave toolbox for MATLAB [22] to simulate how the acoustic wave propagates through the vocal tract until it reaches the lips. The applications of k-wave toolbox range from acoustic [21] and ultrasound wave propagation to photoacoustic tomography.

Although there are various popular methods for acoustic wave propagation, such as finite differences, finite elements and boundary element methods, in general they require a lot of time since they require a small time step and a lot of grid points per wavelength. k-wave solves these issues by interpolating a Fourier series through all of the grid points to get an estimation of the gradient. This way, the computations require fewer grid points since it employs Fast Fourier Transform (FFT), which enables them to make the calculations faster. An issue that arises is that as the wave reaches the grid boundaries, it keeps propagating by entering from the opposite site. To prevent this, k-wave implements an absorbing boundary condition called Perfect Match Layer (PML).

3. EXPERIMENTS

The experiment can be divided into three parts: 1) image information extraction, 2) formant estimation, and 3) acoustic simulations.

3.1. Image information extraction

The data that we used for the experiment is 3D MRI data of the vocal tract of five vowels of French language /a/, /œ/, /y/, /o/, /y/, in three different head positions: up, natural and down. Using tools provided by ITK-SNAP, we manually segmented the vocal tract of the mid-sagittal slice. We then used meshlab [6] to apply Laplacian smoothing filtering with a step of 3 to all the images.

In order to measure the head position, we used the measurement proposed in [16]. The main idea is to use an angle defined by two lines to define the head
position. The first line is the one that connects the interior edge of the C2-C3 cervical vertebrae. The second line is the one that connects the posterior tip of the spinous process of C1 and the tuberculum sella. As shown in Figure 1, number 1 corresponds to the first line, number 2 corresponds to the second line, while number 3 corresponds to the calculated angle. We used imageJ software [18] to manually specify the lines and make the angle computations. The average angle was $144.8 \pm 0.6^\circ$, $124.7 \pm 0.8^\circ$, $101.3 \pm 1.1^\circ$ for the up, normal and natural position respectively. Since there is at least $20^\circ$ of difference between the three positions, we were expecting to notice some difference in phonation [10].

![Figure 1: 2D segmentation of /o/ (top row) and /i/ (bottom row) vowels at up, normal and down position from left to right. Lines 1 and 2 are used to define the angle 3 of the head position](image)

3.2. Formant estimation

In the case of a standard speech signal (recorded in a fairly quiet room) formants can be extracted by applying standard algorithms, e.g. linear prediction coding (LPC) which is used in Praat [4]. For speech recorded in an MRI machine, the situation is quite different. First, the amplitude of the signal becomes much higher when the machine starts acquiring images. Since the signal must not be clipped even when the noise machine is intense, the recording level is low, and consequently the signal is poorly defined. Second, the transfer function of the optical microphone (for instance in the case of our FOMRI III (Optoacoustics, Or Yehuda, Israel)) attenuates the energy at low frequencies, and consequently the energy of the first formant is always lower than expected. This is important because LPC cannot be used anymore since F1 is often too weak to be detected. We therefore resorted to an algorithm derived from the standard linear cepstral smoothing called true envelope" [9, 17]. The advantage of cepstral smoothing is that it does not impose the implicit assumption of an all-pole model. Compared to linear cepstral smoothing, true envelope algorithm provides the additional advantage of approximating harmonics instead of smoothing the spectrum.

Figure 2 shows the narrow band spectrum (curve with harmonics) and the true envelope spectrum (smooth curve), obtained with Winsnoori [11].

Once the MRI acquisition starts the MRI noise takes over speech. Denoising this speech offers a much better speech quality perception, but at the same time degrades the intrinsic acoustic properties of speech. In particular, the detection of spectral peaks corresponding to formants becomes chaotic, and we had great difficulty in determining the formants in the denoised speech. We thus visually checked that harmonics of speech still present were compatible with the formants of the vowel just before acquisition (although they were strongly dominated by the MRI machine noise).

3.3. Acoustic simulations

We used k-wave toolbox for MATLAB in order to make acoustic simulations. At this point, the data are in the form of a 3D shaped surface with sagittal width of 1.2 mm which is the slice thickness of our 3D MRI. Each 3D surface mesh is then transformed into a voxel-based grid from which we take its projection to a 2D plane parallel to the sagittal slice.

In order to run the simulation, one should specify the simulation parameters mainly based on three aspects: 1) having a good approximation of the realistic conditions, 2) stability of the simulation and 3) acquiring sufficient duration of simulated signal.

In our case, the computational grid has a size of $128 \times 128$ (coronal $\times$ axial) with grid spacing $d_x = d_y = 1 \text{mm}$. k-wave uses Fast Fourier Transform (FFT) for computations; therefore, it is suggested to use numbers with low prime factors (ideally powers of 2) as grid dimensions. A PML layer of 10 grid points was added at the boundaries of ev-
ery side to solve the issue of circular wave propagation that happens when simulating with k-wave, the problem that was mentioned in the method section. We manually select the source position of every vowel to correspond to the position of the vocal folds. As a source, we used a disc of 5 grid point radius that emits a delta pulse of pressure (spreading equally to all directions) and 1 Pa amplitude. The medium properties inside the vocal tract were $c_{in} = 350m/s$, $d_{in} = 1kg/m^3$ and the properties outside were $c_{out} = 1000m/s$, $d_{out} = 1000kg/m^3$, where $c_{in}$, $c_{out}$ are the speed, and $d_{in}$, $d_{out}$ are the densities inside and outside the vocal tract respectively. We simulated the signal for 30ms, with a time step of $3 \times 10^{-8}$, resulting in 1000001 signal samples. The pressure signal was recorded with a sensor placed at the end of the vocal tract and the maximum allowed frequency of the grid was 175KHz. Every simulation takes 3 hours and 20 minutes on average.

Finally, we computed the transfer function of every vocal tract and computed the peaks that appear in the frequency domain to compare them with the formants of the original audio signal as shown in Table 1.

### 4. DISCUSSION

The visual inspection of images shows that the up position (bigger angle between the pharyngeal and mouth cavity) increases the volume of the back cavity corresponding to the pharynx but reduces its length. Conversely, the down position essentially results in a change in the angle between the two cavities, and in a smaller size of the front cavity but does not significantly change the volume of the pharyngeal cavity. The acoustic impact of these modifications are visible in Table 1. It should be noted that the variations in formant frequencies between the neutral position and the other two positions are confirmed by acoustic simulations in terms of direction, even if their magnitude is sometimes different. This last point can be explained by the fact that numerical simulations are bidimensional and that formants were measured with some difficulties in the noisy speech.

In terms of formant frequencies the effect is quite negligible for the first formant of vowels that have a large pharyngeal cavity because the volume increase of the Helmholz resonator is proportionally quite small. On the other hand, the effect is more pronounced for $F2$, either because it corresponds to the half wavelength for the pharyngeal cavity for /i/ which results in a lower value for the neutral and down positions (longer pharyngeal cavity), or because it corresponds to the Helmholz resonator of the mouth in the case of /u/ and /o/ for the neutral and down positions, which results in a higher value (smaller volume). Future work will focus on techniques for adapting articulatory models from these data and observations.

### 5. ACKNOWLEDGEMENT

This work was financed by Lorraine Université d’Excellence (LUE) grant and ANR ArtSpeech.

### 6. REFERENCES


