COMPARISON BETWEEN 2D AND 3D MODELS FOR SPEECH PRODUCTION: A STUDY OF FRENCH VOWELS

Ioannis K. Douros[1,2], Pierre-André Vuissoz[2], Yves Laprie[1]

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

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

ABSTRACT

In the present work, we study the production of five vowels of French. The idea is to compare 2-dimensional models with 3-dimensional models, and examine whether the 2-dimensional articulatory models can adequately describe the acoustics of the vocal tract. For the purposes of our experiments, we used 3-dimensional MRI data to acquire the shape of the vocal tract. We then use it to simulate how the pressure wave produced from the vocal folds propagates with the K-wave Matlab toolkit. We carried out acoustic simulations for both the 3-dimensional and the 2-dimensional (mid-sagittal plane) shapes of the vocal tract and compared formant frequencies with those calculated from the denoised speech signal recorded in the MRI machine. We then compared the results of the 2-dimensional and 3-dimensional acoustic simulation with those provided by the traditional simulation used in articulatory synthesis, which relies on the plane wave propagation assumption.

Keywords: acoustic simulations, electrical simulations, French vowels, MRI of the vocal tract

1. INTRODUCTION

Speech synthesis has reached a high level of naturalness through concatenative approaches [2, 10] and more recently deep learning approaches [28]. Both are based on the use of a large corpus of prerecorded speech.

The speech corpus has to cover all phenomena that are to be treated. Hence, the more phonetic contexts, speech styles, expressions, speaker postures, etc the corpus covers, the more natural the synthesized speech. All this contributes to increasing the size of the corpus.

On the other hand, the weakness of those techniques is tightly connected to their dependence on the corpus. This means that changing the speaker characteristics, adding new expressions, or taking speech production disorders into account is almost impossible. These techniques do not contribute to the understanding of speech production and are unable to link an acoustic cue, for example the evolution of frequencies in the vicinity of a consonant, to their articulatory origin.

Unlike these approaches, which only model the result of speech production, i.e. the acoustic speech signal, articulatory synthesis [3, 14, 27] explicitly models the link between the vocal tract, vocal folds and aero-acoustic phenomena. This is achieved by solving the equations of aerodynamics and acoustics in the vocal tract and by using its geometry as an input.

The geometry of the vocal tract results from the position of the speech articulators, and thus from the activity of corresponding muscles. A first solution to model these phenomena is to use biomechanical modeling to compute the shape of all the deformable speech articulators [9, 20]. This involves modeling the behaviour of muscles and the properties of muscle tissues in a realistic way before solving the mechanical equations. Despite constant progress, this approach is often limited to predicting the position of the jaw and the tongue shape and it is still unimaginable to use a biomechanical approach to calculate the whole shape of the vocal tract.

For these reasons we prefer to use an articulatory model [1, 13] to compute the vocal tract geometry. Of course, the articulatory model must provide a geometric description close to reality in order to guarantee a good quality of synthesis. Similarly to the biomechanical approach, one of the challenges consists of collecting and processing data to construct the model. As the delineation of the articulators in the MRI images is a task that requires a certain amount of interpretation of the geometry of the vocal tract and has a very long processing time, it is often preferable to construct a two-dimensional model in the mid-sagittal plane and then calculate the transverse area at each point of the vocal tract from the glottis to the lips [7].

There have been several studies of the vocal tract using different types of articulatory data like Electromagnetic Articulography (EMA) and X-ray films in order to synthesise speech or track the movement of vocal tract parts [15], coupling electromagnetic sensors and ultrasound. Another data acquisition technique that has been widely used over the last years is Magnetic Resonance Imaging (MRI). In [23], they estimate the area function of the vocal tract from MRI images and then compare the results with the sound acquired at a different session. Even though they tried to make the audio recording condition as close to the original MRI as possible [11], there could still be significant difference between the recorded audio signal and the signal pronounced during MRI acquisition due to the difference in the auditory feedback for the speaker during the noisy MRI recordings and otherwise.

In this work, our purpose is to examine to what extent the 2-dimensional data can describe articulatory information, compared to 3-dimensional data, since it is much more efficient to use with a 2-dimensional model [7]. We also examine how the acoustics are affected in each case by comparing the sound signal with the results from 2 and 3-dimensional acoustic and 2-dimensional electrical simulations.

The presented work can be divided into two main parts: 1) the data acquisition and processing, and 2) the simulations and the comparison of the results.

2. MATERIALS AND METHODS

2.1. Data acquisition

For the purpose of this study, we used an MRI data part of a study approved by an ethics committee and the subject gave written informed consent (ClinicalTrials.gov identifier: NCT02887053). The subject used for the data acquisition is a healthy male French native speaker at the age of 32, without any reported speaking or hearing problems.

The MRI data was acquired on a Siemens Prisma 3T scanner (Siemens, Erlangen, Germany) with a gradient of $80mT/m$ amplitude and $200mT/m/ms$ slew rate. We used the 3-dimensional cartesian vibe sequence ($TR = 3.57\ ms$, $TE = 1.43$, $FOV = 22 \times 20\ mm$, $flip\ angle = 9\ degrees$) for the acquisition. The pixel bandwidth is $445Hz/pixel$ with an image resolution of 256×174. Scan slice thickness is 1.2 mm and the number of slices is 120. The pixel spacing is 0.8597 and the acceleration factor is 3 $iPAT$. The acquisition time was $7.4\,s$ which allows the subject to maintain phonation easily.

The subject's vocal tract was imaged while he lay supine in the MRI scanner. The recording time for the subject, including calibration and pauses between phonemes, was 2 hours.

Audio was recorded at a sampling frequency of 16 kHz inside the MRI scanner using FOMRI III (Optoacoustics, Or Yehuda, Israel) fiber optic microphone. The subject starts producing the phoneme just before the MRI recording starts and sustains phonation until the end of the acquisition. 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. In order to de-noise it, we used the de-noising algorithm proposed in [19].

We apply this algorithm to our data using the FASST toolbox [21].

2.2. Segmentation

For the purposes of our experiments, we used the ITK-SNAP software [30] to segment the volume of the vocal tract. ITK-SNAP provides a great variety of tools for segmenting images, both automatically and manually.

As far as automatic segmentation is concerned, ITK-SNAP implements two active contour segmentation algorithms, region competition and geodesic active contours [4], [31].

For manual segmentation, ITK-SNAP offers two types of tools, the most interesting among them is the adaptive brush that adjusts itself to follow the image boundaries. The brush tool can be used for both to $2D$ and $3D$ image segmentation.

2.3. Acoustic simulation

For the acoustic simulations, we employ the k-wave Matlab toolbox [26]. This toolbox has a wide range of applications like photoacoustic tomography ultrasound wave propagation [29], and acoustic propagation. [25].

Several numerical methods have been developed to solve the partial differential equations of acoustics, like finite differences, finite elements, and boundary element methods [24]. These methods offer significant advantages as they can calculate acoustic characteristics accurately and implement frequency dependent losses at boundaries. However, in many cases these methods are significantly slow. This happens due to the fact that they require a small time step to achieve adequate accuracy and a lot of grid points per wave length. In the method used by k-wave these problems are solved by interpolating a Fourier series through all of the grid points in order

to get the estimation of the gradient. This approach solves the problems of the previously referred methods as it a) requires fewer grid points (only two) per wave length since the base function of the Fourier series is the sinusoid and b) it can be fast since it employs Fast Fourier Transform (FFT) to calculate the amplitudes of the simulated signals. A problem that arises is that when a wave approaches the computational grid boundaries, it keeps propagating to the medium by entering from the opposite site of the computational grid. This happens because of the usage of the FFT algorithm for the computation. To tackle this issue, k-wave adds a specific type of layer to the boundaries of the computational grid by implementing an absorbing boundary condition, called Perfect Match Layer (PML), which prevents this phenomenon.

Finally, k-wave toolbox has a great number of parameters that can be customised for a simulation, most of them concerning the grid and time sparsity, the properties of the mediums, the sensors, the sources, the number of dimensions ($1D/2D/3D$), the number of PML, etc.

2.4. Electrical simulation

To perform the electrical simulation we used some of the tools provided from the Xarticul software [16], [22]. Xarticul offers multiple tools, like an easy way to delineate and process articulator contours, semi-automatic articulatory measurements and construction of articulatory models [12]. Xarticul can perform acoustic simulations from the area function by using the algorithm proposed in [6].This algorithm is based on the Transmission Line Circuit Analog (TLCA) method [18]. The main idea of the algorithm is to model every tube used to describe the vocal tract as a circuit of electrical units whose parameters (electrical) correspond to the physical (acoustic). For example, the current and the voltage of the circuit in the TLCA correspond to the volume velocity and acoustic pressure respectively. Therefore, instead of describing the vocal tract as a continuous connection of tubes, one can describe it as a continuous connection of electric circuits. The main advantage of this approach is that it allows to model time-varying geometries of the vocal tract [6].

3. EXPERIMENTS

Our experiments can be divided into three main stages: a) image segmentation, b) acoustic simulations and c) electric simulations.

3.1. Image segmentation

For the purposes of our experiments, we used five of the vowels from the database described in the previous section, /a, œ, i, o, y/. First, we processed the images with 3DSlicer [8] (http://www.slicer.org) to apply lanczos interpolation in order to make corrections to the image's axis. Then, we used the tools provided by the ITK-SNAP software to automatically segment the 3D volume of the vocal tract and manually corrected the result. The area of interest begins at the glottis and extends to the lips at the point where the lips stop being simultaneously visible at the coronal plane. We used two classes and 10000 points as nearest neighbours in order to assign each point to the appropriate class for the creation of the probabilistic map, and 10 balls on average per vowel as "seeds" with various sizes based on the region of the vocal tract where they were placed. Then we applied the active contour algorithm which required between $300-500$ iterations to cover the whole vocal tract. The amount of iterations is greatly based on the vowel and the initial number, size and position of the "seeds". For the manual segmentation we used the adaptive brush tool with the default parameters to acquire the vocal tract mesh Figure 1. Finally we used meshlab [5] to smooth every mesh by applying Laplacian smoothing filter with step 3. For each vowel, about 4 hours of processing was required, with the biggest amount of time spent on the manual segmentation step.

3.2. Acoustic simulations

For the acoustic simulations, we used k-wave toolbox for Matlab [26]. For every vowel examined, simulations were carried out in both $2D$ and $3D$. First, the mesh was transformed into a volumetric representation using voxels. Then we specified the parameters for the $2D$ and $3D$ simulations. Since k-wave uses FFT, the number of grid points was set so as to have low prime factors, ideally a power of 2. For the $3D$, we used a grid size of $128 \times 128 \times 128$ grid points ($sagittal \times coronal \times axial$) with $d_x = d_y = d_z = 1mm$. We also used a PML layer of 10 grid points at the boundaries of every side of the grid, to avoid the wave penetrating the opposite side, as explained in the previous section. As a source we used a ball which emits a $delta$ pulse of pressure, spreading equally in all directions. The source has radius of 5 grid points, amplitude 1 Pa and was placed at input of the vocal tract, which was specified manually for every vowel. To record the simulated pressure we used a sensor placed at the end of the vocal tract. The medium properties inside the

vocal tract were $c_{in} = 350m/s, d_{in} = 1kg/m^3$ and the properties outside, i.e. in the tissues that delimit the vocal tract, 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. The time step is set according to the two medium characteristics (here tissues and air) and the accepted value is $3 * 10^{-8} sec$ to guarantee a good stability. The amount of time steps computed was 1000001. The maximum allowed frequency of the grid was $175KHz$. For the $2D$ case, we run the simulations on the $y-z$ plane using a disc instead of a ball on the mid-sagittal plane of the vocal tract. All the other parameters remained the same between the two simulations. The amount of time required for the simulation of each vowel is about 75 hours for $3D$, while for $2D$ is approximately 3 hours and 20 minutes. Finally, we calculated the transfer function of every vocal tract and computed their peak frequencies (Table 1), to compare them with the formants computed with the electrical simulation.

3.3. Electric simulations

For the electric simulations we used the mid-sagittal planes from the $3D$ MRI acquisition. We used Xarticul to manually delineate the articulator contours of each vowel. Afterwards, we used 40 tubes (Figure 1) to estimate the area function of the vocal tract in order to compute its formants (Table 2). Finally, we used selective LPC to compute the formants of the original audio signal (Table 3). We also made a comparison with the values given in the literature for French [17] (Table 3).

Table 1: 2D / 3D formants computation from acoustic simulations in Hertz

	$F1$	$F2$	$F3$
/a/	674 / 694	1349 / 1231	3169 / 2545
/œ/	416 / 460	1444 / 1444	2499 / 2544
/i/	337 / 304	2394 / 2207	3136 / 3237
/o/	404 / 440	900 / 787	2728 / 2605
/y/	281 / 285	1798 / 1802	2192 / 2198

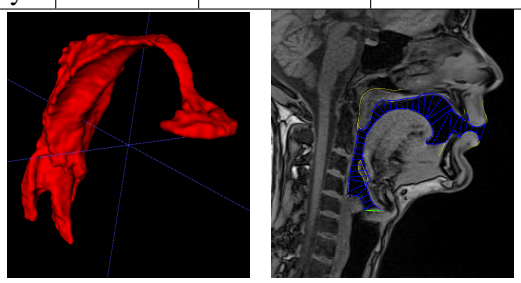

Figure 1: 3D volume of /i/ (left), separation of the vocal tract into acoustic tubes for /o/ (right)

Table 2: 2D formants computation from electrical simulations in Hertz

	$F1$	$F2$	$F3$
/a/	510	1200	2190
/œ/	408	1276	2168
/i/	280	1684	2927
/o/	491	905	2185
/y/	393	1911	2205

Table 3: Theoretical/measured values of French vowels formants in Hertz

	$F1$	$F2$	$F3$
/a/	684 / 689	1256 / 1256	2503 / 2604
/œ/	517 / 443	1391 / 1335	2379 / 2436
/i/	308 / 380	2064 / 2306	2976 / 3193
/o/	383 / 430	793 / 732	2283 / 2619
/y/	300 / 336	1750 / 1854	2120 / 2228

4. DISCUSSION

The first remark concerns the sounds produced by the speaker in the MRI machine. As shown in Table 3 which gives the average values for French speakers, the measured values are a little far from the expected values. This is especially true for the first formant of close vowels /i,o/ which is higher in frequency. This would mean that the pharyngeal cavity is smaller due to the subject posture. The visual examination of the images shows a slightly shifted articulation in some cases. Second there is a good agreement between the results of 2D/3D simulations and the formants F1 and F2 determined from the speech signal recorded. However, for F3 the 3D simulation turns out to give results closer to those of natural speech than those of the 2D simulation, probably because the 3D volume gives a geometry closer to the real one.

The third remark concerns the comparison between the acoustic and electrical simulations. It turns out that the electric simulation is not as good as the acoustic simulation to reproduce the formants. Since there is a good agreement between the 2D and 3D acoustic simulation, the most probable hypothesis is that either splitting of the vocal tract into small tubes or the estimation of the area function from the mid-sagittal shape is not completely satisfactory.

Future direction for research will focus these points so as to improve the quality of articulatory synthesis.

5. ACKNOWLEDGEMENT

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

6. REFERENCES

[1] Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., Savariaux, C. 2002. Three-dimensional linear articulatory modeling of tongue, lips and face based on mri and video images. 30(3), 533–553.

[2] Bellegarda, J. R. 2007. Lsm-based unit pruning for concatenative speech synthesis. *ICASSP 2007* volume 4 IV–521–IV–524.

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

[4] Caselles, V., Kimmel, R., Sapiro, G. 1997. Geodesic active contours. *International journal of computer vision* 22(1), 61–79.

[5] Cignoni, P., Callieri, M., Corsini, M., Dellepiane, M., Ganovelli, F., Ranzuglia, G. 2008. Meshlab: an open-source mesh processing tool. *Eurographics Italian Chapter Conference* volume 2008 129–136.

[6] Elie, B., Laprie, Y. 2016. Extension of the single-matrix formulation of the vocal tract: Consideration of bilateral channels and connection of self-oscillating models of the vocal folds with a glottal chink. *Speech Communication* 82, 85–96.

[7] Ericsdotter, C. 2007. Detail in vowel area functions. *ICPhS 2007* 513–516.

[8] Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.-C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., others, 2012. 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic resonance imaging* 30(9), 1323–1341.

[9] Fels, S., Vogt, F., Van Den Doel, K., Lloyd, J., Stavness, I., Vatikiotis-Bateson, E. 2006. Artisynth: A biomechanical simulation platform for the vocal tract and upper airway. *ISSP 2006* volume 138. Citeseer.

[10] Hunt, A. J., Black, A. W. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. *ICASSP 1996* volume 1. IEEE 373–376.

[11] Kröger, B. J., Winkler, R., Mooshammer, C., Pompino-Marschall, B. 2000. Estimation of vocal tract area function from magnetic resonance imaging: Preliminary results. *ISSP 2000* 333–336.

[12] Laprie, Y., Busset, J. Construction and evaluation of an articulatory model of the vocal tract. *EUSIPCO 2011, year = 2011, pages = 466–470, organization = IEEE,*.

[13] Laprie, Y., Busset, J. 2011. A curvilinear tongue articulatory model. *ISSP 2011* Canada, Montreal.

[14] Laprie, Y., Loosvelt, M., Maeda, S., Sock, E., Hirsch, F. August 2013. Articulatory copy synthesis from cine x-ray films. *Interspeech 2013* Lyon, France.

[15] Laprie, Y., Loosvelt, M., Maeda, S., Sock, R., Hirsch, F. 2013. Articulatory copy synthesis from cine x-ray films. *InterSpeech 2013*.

[16] Laprie, Y., Sock, R., Vaxelaire, B., Elie, B. 2014. Comment faire parler les images aux rayons x du conduit vocal. *SHS Web of Conferences* volume 8. EDP Sciences 1285–1298.

[17] Lonchamp, F. 1984. Les sons du français — analyse acoustique descriptive. Cours de phonétique Institut de Phonétique, Université de Nancy II.

[18] Maeda, S. 1982. A digital simulation method of the vocal-tract system. *Speech communication* 1(3-4), 199–229.

[19] Ozerov, A., Vincent, E., Bimbot, F. 2012. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing* 20(4), 1118–1133.

[20] Perrier, P., Payan, Y., Zandipour, M., Perkell, J. 2003. Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study. *JASA* 114(3), 1582–1599.

[21] Salaün, Y., Vincent, E., Bertin, N., Souviraa-Labastie, N., Jaureguiberry, X., Tran, D. T., Bimbot, F. 2014. The flexible audio source separation toolbox version 2.0. *ICASSP*.

[22] Sock, R., Hirsch, F., Laprie, Y., Perrier, P., Vaxelaire, B., Brock, G., Bouarourou, F., Fauth, C., Ferbach-Hecker, V., Ma, L., others, 2011. An x-ray database, tools and procedures for the study of speech production. *ISSP 2011* 41–48.

[23] Story, B. H., Titze, I. R., Hoffman, E. A. 1996. Vocal tract area functions from magnetic resonance imaging. *JASA* 100(1), 537–554.

[24] Takemoto, H., Mokhtari, P., Kitamura, T. 2010. Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method. *JASA* 128(6), 3724–3738.

[25] Treeby, B. E., Cox, B. 2010. Modeling power law absorption and dispersion for acoustic propagation using the fractional laplacian. *JASA* 127(5), 2741–2748.

[26] Treeby, B. E., Cox, B. T. 2010. k-wave: Matlab toolbox for the simulation and reconstruction of photoacoustic wave fields. *Journal of biomedical optics* 15(2), 021314.

[27] Tsukanova, A., Elie, B., Laprie, Y. 2017. Articulatory speech synthesis from static context-aware articulatory targets. *ISSP 2017*.

[28] Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W., Kavukcuoglu, K. 2016. Wavenet: A generative model for raw audio. *SSW* 125.

[29] Wise, E. S., Treeby, B. E. 2013. Full-wave nonlinear ultrasound simulation in an axisymmetric coordinate system using the discrete sine and cosine transforms. *IUS 2013*. IEEE 1374–1377.

[30] Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., Gerig, G. 2006. User-guided 3d active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage* 31(3), 1116–1128.

[31] Zhu, S. C., Yuille, A. 1996. Region competition: Unifying snakes, region growing, and bayes/mdl for multiband image segmentation. *IEEE transactions on pattern analysis and machine intelligence*.