Modeling the temporal evolution of the vocal tract shape with deep learning

Yves Laprie, Vinicius Ribeiro, Karina Isaeva, Justine Leclere, Jacques Felblinger, Pierre-André Vuissoz

MODELING THE TEMPORAL EVOLUTION OF THE VOCAL TRACT SHAPE WITH DEEP LEARNING

Yves Laprie[a], Vinicius Ribeiro[a], Karina Isaeva[b], Justine Leclere[b,c], Jacques Felblinger[b,d], Pierre-André Vuissoz[b]

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

[b] Université de Lorraine, INSERM, U1254, IADI, Nancy, F-54000, France
[c] Service de Médecine Bucco-dentaire, Hôpital Maison Blanche, Reims, F-51100, France

[d] CIC-IT 1433 INSERM CHRU Nancy, F-54000, France
Yves.Laprie@loria.fr

ABSTRACT

This paper overviews our work on the links between coarticulation modeling, approached from the point of view of predicting the vocal tract shape from the phonetic sequence, and the available real-time MRI corpora. Real-time MRI has revolutionized the acquisition of articulatory data through the image quality, the possibility of acquiring and denoising the speech signal, and the possibility of recording corpora containing several thousand sentences. Coarticulation modeling is only possible with the ability to reliably track articulator contours in many images. Tracking techniques using neural networks have provided efficient solutions comparable in reliability to humans. Finally, we show that even if recurrent neural networks trained on these corpora can successfully predict the shape of the vocal tract, it is still necessary to use constraints directly from phonetics to ensure consistency in the prediction.

Keywords: vocal tract shape, articulatory modeling, rt-MRI, RNN, deep learning

1. INTRODUCTION

Modeling coarticulation is a central issue in phonetics which thus raised many works intended to approximate the position of articulators or, equivalently, to predict the constriction locations. One of the great difficulties has always been supporting theories on data in sufficient quantity and quality. For this reason, research has been structured in two directions. The first proposed theories focused mainly on the scope of anticipation of articulatory gestures and then tried to construct small corpora to validate them. Globally, the debate centered on the scope of anticipation, depending on whether it has a fixed duration or, on the contrary, depends on the constraints imposed by the achievement of constrictions in the vocal tract. These approaches have suffered from the handicap of relying on a limited amount of data without the possibility of moving from theory to a numerical model. Farnetani [1] offers a complete overview of approaches from phonetics and phonology.

The second method consisted in proposing strongly constrained numerical models and fitting their parameters from small corpora, which is possible because the introduction of constraints reduces the number of parameters to be fitted. In practice, the Öhman model [2, 3] is probably the most widely used model. Cohen and Massaro's model [4] based on dominance functions was only used for labial modeling but could have easily been extended to other articulators. This model requires five parameters for each articulatory factor and each phoneme, with a total of about 4 000 parameters. Nam et al. [5] determined the parameters of the gestural scores associated to the articulatory phonology proposed by Browman and Goldstein [6] and the TADA task dynamics model [7]. The gestures are assumed to be known and the learning process optimizes two parameters: the onset and offset times of each gesture from a Microbeam X-ray database [8] recorded for 47 speakers who uttered at most 56 sentences. Thus, two parameters for each occurrence of the gesture are optimized to copy the speech signal best. This approach has not been used to generate the shape of the vocal tract but only to carry out automatic speech recognition. Our main idea was to impose a minimum of constraints on the learning process and the global model was therefore learned directly from the data of a speaker who pronounced 700 sentences in French. We now present a larger corpus of 2100 French sentences for one female and one male speaker that will allow

us to extend this approach. The data consisted of the contours of all articulators in the vocal tract, as shown in Figure 1.

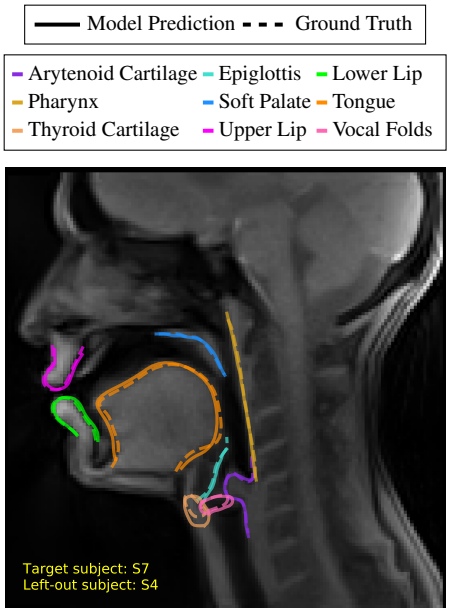

Figure 1: Articulator contours. Authorized reproduction from [9].

2. ARTICULATORY DATA AND ACQUISITION TECHNIQUES

Access to data has always been a major obstacle to the study of coarticulation. Cineradiography was the first source of data, but the hazards to which the subjects were exposed limited the duration of the films to about 30 seconds per year at most and this technique was finally abandoned in the late 1980s for ethical reasons. Then, electromagnetic articulography (EMA) allowed the data acquisition in larger volumes with limited risks.

The strength of EMA is to offer a high sampling frequency of up to 800 HZ but with the limitation of monitoring the position of only a few sensors, what is more with the difficulty of having to glue these sensors for a duration reduced to 45 minutes in the best case. The substantial increase in the recording duration is essential to enable machine learning since it heavily relies upon data availability. For this reason, EMA databases, particularly those acquired by Richmond [10], have been widely used to perform articulatory acoustic inversion experiments. Nevertheless, the limitation of the number of sensors prohibits the determination of tongue root and larynx movements, which play a crucial role in the length of the vocal tract and thus on the acoustic properties of speech.

2.1. MRI specificities w.r.t. speech production

Magnetic resonance imaging and its real-time version significantly evolved the observation of articulatory gestures. It is now possible to acquire excellent quality data at a frequency of 50 Hz, and even recently 80 Hz, with a sufficient spatial resolution (1.4 mm pixels) for our application [11] with the system developed at the Max Planck Institute for Multidisciplinary Sciences by Uecker et al. [12]. The speech signal can be recorded with an optical microphone and be almost entirely denoised using recent techniques [13]. The speech recorded in this way suffers only from the Lombard effect due to machine noise and the supine position.

Despite these very substantial advances, there are still limitations imposed by the MRI acquisition technique itself. The first, called "partial volume effect" [14], is related to the thickness of the mid-sagittal slice. This thickness is of the order of 8 mm to obtain a sufficient signal-to-noise ratio (SNR), hence, a good contrast. Therefore, the volume giving rise to a pixel is $1.4 \times 1.4 \times 8$ mm in our data. In some cases, this volume is only partially occupied by the imaged organs, especially near the edges of the tongue at the groove and the apex level, leading to pixels that are sometimes difficult to classify as part of the tongue or not. The second limitation is related to the MRI acquisition technique. The acquisition is performed in the image's Fourier domain, and a complete image would require as many elementary acquisitions as the number of lines in the image. The real-time acquisition technique is based on acquiring a small number of data – in this case, nine variable rays in Fourier space for the algorithm of Uecker et al. [12] – which reconstructs the complete image with regularization techniques. Each elementary acquisition (one ray in Fourier space) requires a little more than 2 ms, and the fast organs, mainly the apex and the lips, can therefore move while obtaining an image. This effect is particularly noticeable in the case of the apex, which comes into contact with the alveolar or post-alveolar region during articulation of the /l/, /d/, /n/ and /t/ sounds. For the apex, and to a lesser extent for the lips, the two effects are cumulative, so it is always challenging to determine the exact position of the apex and the lips. For this reason, the images cannot be considered as representing the ground truth.

2.2. Acquisition strategy

One challenge is acquiring large corpora of several thousand sentences to have excellent coverage of phonetic contexts for a speaker. For ethical and practical reasons, it is not possible to exceed one hour and 15 minutes in the MRI machine; thus, it is necessary to organize several recording sessions with the same speaker posture to guarantee the homogeneity of the articulatory data concerning the relative positions of the oral and pharyngeal cavities. Therefore, we made a blocking foam perfectly adapted to the MRI antenna and the speaker's head. First, a cast of the back of the head was made in alginate, which was used to create a plaster positive. Next, with a 3D-printed copy of the antenna, we placed the plaster cast in the same position as the speaker's head during the acquisition. We filled the space between the two with expanding foam (FlexFoam-iT III). The foam-blocking helmet (Figure 2d) can be used during the acquisition to ensure the head's position is the same during all sessions. For the last corpus constructed that will be used for future coarticulation models, six recording sessions were required for approximately 2 100 French sentences.

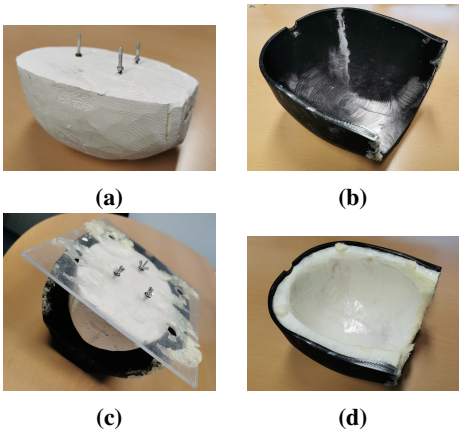

Figure 2: (a) Plaster cast of the head, (b) copy of the antenna, (c) adjustment of the cast in the antenna, (d) blocking foam.

The MRI acquisition parameters are identical to those in [11]. Each acquisition lasts 80 or 90 seconds, allowing the images and the speech signal to be easily processed. Coarticulation modeling requires precise phonetic segmentation because it is essential that the images, and thus the position of the articulators, correspond perfectly to the articulated phoneme. Forced-alignment tools developed for automatic speech recognition (ASR) were used to align the speech signal with the text transcription. The resulting phonetic segmentations were then manually corrected with great care. Coarse errors were corrected, and we separated the closure from the burst stops so that the vocal tract shape prediction could model the relevant closures. This correction work requires about an hour and a half for every 177 acquisitions in the new corpus.

In addition to phonetic segmentation, it is necessary to obtain the position of each articulator. The emergence of deep convolutional networks for image segmentation [15, 16] allowed considerable progress. The system initiated in [17] was extended to cover all articulators from the glottis to the lips in [9], and the mandible and hard-palate positions were estimated using image correlation. We manually delineated the contours in about 2 000 images, which is reasonable compared to the current corpus size of more than 500 000 images. The performance is very close to that of a human with the advantage of providing a remarkable homogeneity in the tracking behavior, unlike a group of human experts which always presents disparities. Due to lack of space, this system is not fully described, but we refer to Ribeiro et al. [9] for a detailed description. Figure 1 gives an illustration of the tracking results.

3. PREDICTING THE VOCAL TRACT SHAPE AND COARTICULATION EFFECTS

One of our primary goals is to model coarticulation and vocal tract shape prediction as independently as possible from constrained numerical models. Modeling can be done at two levels: the geometrical representation of the vocal tract shape and then the prediction of the vocal tract shape. Many articulatory models have been proposed to approximate the shape of the vocal tract. The models can be based either on the use of geometric primitives (lines, circles, and others, in two dimensions, or planes, disks, and others, in three dimensions) [18], or on the analysis of a set of articulatory contours extracted from a corpus of midsagittal images of the vocal tract [19, 20, 21]. For both models, the vocal tract shape is represented by a vector of about ten parameters in the best case, seven for the Maeda model [22] but at the cost of relatively simplified modeling of the vocal tract shape. In the case of models obtained by applying a data analysis technique, a marked geometrical compensation can be generally observed, complicating the choice of the best articulatory vector to represent one shape. Thus, we decided to forego intermediate modeling and use articulatory contours as obtained by the tracking.

The first experiment presented in [23] relies on an deep encoder-decoder neural network built using a bidirectional Gated Recurrent Units (GRU) [24] and a linear decoder with `ReLU` nonlinearities to reconstruct the articulators' shapes. Model training is performed from the contours of 30 90-second acquisitions and the evaluation focuses on the geometric accuracy and realization of four classical articulatory variables: lip aperture (LA), tongue tip constriction degree (TTCD), tongue body constriction degree (TBCD) and velum opening (VEL), which have a significant acoustic impact. The experiments produced very good results despite the limited amount of data available. As we have seen above, the physical origin of the real-time MRI images does not allow to know precisely the position of the apex and the lips, which play an essential role in the realization of complete or partial constrictions of the vocal tract. Also, one of the important points to consider in the vocal tract shape prediction is the model's ability to compensate for a perturbation imposed on an articulator. For these reasons, it would have been preferable to have constraints that better control the degree of a particular constriction. Since each contour is represented by a vector of independent points forcing critical articulators constrictions were not direct since it can lead to very unrealistic shapes [25].

On a second approach, we trained an autoencoder to reduce the articulators' dimensionality, representing each contour by a set of control parameters in the autoencoder latent space [25]. The results are limited to the tongue for simplicity, but the extension to the complete vocal tract is possible. The latent space is analogous to the linear components retained by principal components analysis (PCA). Unlike PCA, an autoencoder does not guarantee the orthogonality, thus, we added a covariance minimization in the feature space to the loss function. We relied on the work done on articulatory model construction [20] to set the size of the latent space to eight. With the autoencoder acting as an implicit articulatory model, we used a similar bidirectional GRU to predict the tongue's principal components as a function of the phoneme sequence to be articulated. The final shape can be reconstructed using the autoencoder's decoder. The advantage of this method is that it allows adding phonetic constraints imposed on the articulatory variables. Taking these constraints into account has a double interest. First, it guarantees that the synthetic vocal tract shapes are compatible with the expected acoustic properties in terms of closure at the place of articulation, which is important in the perspective of articulatory synthesis. Secondly, these constraints are intended to compensate for the limitations of real-time MRI presented in Section 2.1, making it very difficult to detect contacts between the apex or tongue body and the alveolar region or palate. Figure 3 presents the results obtained for the tongue when using the encoder-decoder network from [23] (Figure 3a) and the autoencoder approach from [25] (Figure 3b).

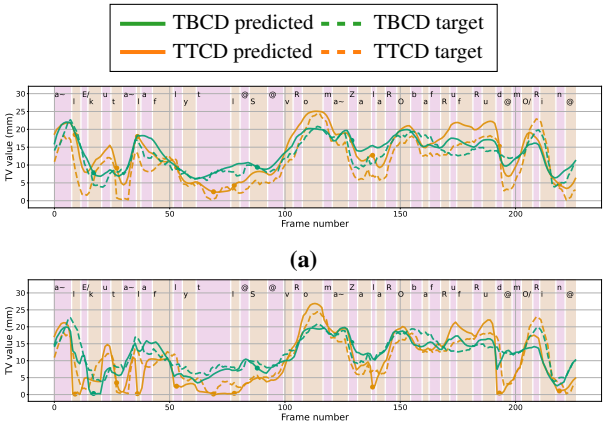

Figure 3: TBCD and TTCD trajectories for the utterance "En écoutant la flûte, le chevreau mangea la robe à froufrous de Maurine". Authorized reproduction from [25].

4. CONCLUSION

As we have shown, the development of real-time MRI and deep learning has profoundly changed the approach to coarticulation. While the number of images was for a long time limited to a few hundred and it was unimaginable to have an automatic tracking of the articulators' contour with a good accuracy, it is now possible to process very large corpora (nearly a million images and more than 2 100 sentences for a single speaker in the case of the last acquired corpus) which allows the use of automatic learning models, often inspired or derived from recurrent neural networks, which are much less constrained than those used in the past. Our past work shows how to design a deep learning-based articulatory synthesizer that allowed us to easily take into account phonetic constraints and will enable the study of articulatory compensation mechanisms.

5. ACKNOWLEDGEMENTS

This research was supported by the French ANR project Full3DTalkingHead, CPER IT2MP, Région Lorraine and FEDER.

6. REFERENCES

[1] E. Farnetani, "Labial coarticulation," in *In Coarticulation: Theory, data and techniques*, W. J. Hardcastle and N. Hewlett, Eds. Cambridge: Cambridge university press, 1999, ch. 8.

[2] S. Öhman, "Coarticulation inVCV utterances: Spectrographic measurements," *J. Acoust. Soc. Am.*, vol. 39, no. 1, pp. 151–168, 1966.

[3] ——, "Numerical model of coarticulation," *J. Acoust. Soc. Am.*, vol. 41, pp. 310–320, 1967.

[4] M. Cohen and D. Massaro, "Modeling coarticulation in synthetic visual speech," in *Computer Animation*, D. Thalmann and N. Magnenat-Thalmann, Eds. Springer-Verlag, 1993.

[5] H. N. V. Mitra, M. Tiede, E. Saltzman, L. Goldstein, C. Epsy-Wilson, and M. Hasegawa-Johnson, "A procedure for estimating gestural scores from natural speech," in *11th Annual Conference of the International Speech Communication Association - INTERSPEECH 2010*, Makuhari, Chiba, Japan, 2010.

[6] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, no. 3-4, pp. 155–180, 1992.

[7] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "TADA: An enhanced, portable task dynamics model in MATLAB," *The Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2430–2430, 2004.

[8] J. Westbury, P. Milenkovic, G. Weismer, and R. Kent, "X-ray microbeam speech production database," *The Journal of the Acoustical Society of America*, vol. 88, no. S1, pp. S56–S56, 1990.

[9] V. Ribeiro, K. Isaieva, J. Leclere, R. Karpinski, J. Felblinger, P.-A. Vuissoz, and Y. Laprie, "Automatic tracking of vocal tract articulators in real-time magnetic resonance imaging," *Available at SSRN 4192628*.

[10] K. Richmond, "Preliminary inversion mapping results with a new ema corpus," in *10th Annual Conference of the International Speech Communication Association - INTERSPEECH 2009*, Brighton, 2009, pp. 2835–2838.

[11] K. Isaieva, Y. Laprie, J. Leclère, I. K. Douros, J. Felblinger, and P.-A. Vuissoz, "Multimodal dataset of real-time 2D and static 3D MRI of healthy French speakers," *Scientific Data* , vol. 8, no. 1, p. 258, Oct. 2021.

[12] M. Uecker, S. Zhang, D. Voit, A. Karaus, K. D. Merboldt, and J. Frahm, "Real-time MRI at a resolution of 20 ms," *NMR Biomed*, vol. 23, no. 8, pp. 986–94, Oct. 2010.

[13] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. on Audio Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, 2012.

[14] E. Bellon, M. Haacke, P. Coleman, D. Sacco, D. Steiger, and R. Gangarosa, "MR artifacts: A review," *AJR. American journal of roentgenology*, vol. 147, pp. 1271–81, 12 1986.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[17] K. Isaieva, Y. Laprie, N. Turpault, A. Houssard, J. Felblinger, and P.-A. Vuissoz, "Automatic tongue delineation from mri images with a convolutional neural network approach," *Applied Artificial Intelligence*, vol. 34, no. 14, pp. 1115–1123, 2020.

[18] P. Birkholz and D. Jackel, "A three-dimensional model of the vocal tract for speech synthesis," in *15th International Congress of Phonetic Sciences - ICPhS'2003, Barcelona, Spain*, Aug 2003, pp. 2597–2600.

[19] D. Beautemps, P. Badin, and G. Bailly, "Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling," *J. Acoust. Soc. Am.*, vol. 109, no. 5, pp. 2165–2180, 2001.

[20] 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.

[21] V. Ribeiro, K. Isaieva, J. Leclere, P.-A. Vuissoz, and Y. Laprie, "Towards the Prediction of the Vocal Tract Shape from the Sequence of Phonemes to be Articulated," in *Proc. Interspeech 2021*, 2021, pp. 3325–3329.

[22] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marschal, Eds. Kluwer Academic Publishers, 1990.

[23] V. Ribeiro, K. Isaieva, J. Leclere, P.-A. Vuissoz, and Y. Laprie, "Automatic generation of the complete vocal tract shape from the sequence of phonemes to be articulated," *Speech Communication*, vol. 141, pp. 1–13, Apr. 2022.

[24] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[25] V. Ribeiro and Y. Laprie, "Autoencoder-Based Tongue Shape Estimation During Continuous Speech," in *Proc. Interspeech 2022*, 2022, pp. 86–90.