Automatic generation of the complete vocal tract shape from the sequence of phonemes to be articulated

Vinicius Ribeiro, Karyna Isaieva, Justine Leclere, Pierre-André Vuissoz, Yves Laprie

Automatic Generation of the Complete Vocal Tract Shape from the Sequence of Phonemes to be Articulated

Vinicius Ribeiro[a], Karyna Isaieva[b], Justine Leclere[b,c], Pierre-André Vuissoz[b], Yves Laprie[a]

[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*

Abstract

Articulatory speech synthesis requires generating realistic vocal tract shapes from the sequence of phonemes to be articulated. This work proposes the first model trained from rt-MRI films to automatically predict all of the vocal tract articulators' contours. The data are the contours tracked in the rt-MRI database recorded for one speaker. Those contours were exploited to train an encoder-decoder network to map the sequence of phonemes and their durations to the exact gestures performed by the speaker. Different from other works, all the individual articulator contours are predicted separately, allowing the investigation of their interactions. We measure four tract variables closely coupled with critical articulators and observe their variations over time. The test demonstrates that our model can produce high-quality shapes of the complete vocal tract with a good correlation between the predicted and the target variables observed in rt-MRI films, even though the tract variables are not included in the optimization procedure.

Keywords:
Phonetic-to-articulatory, Speech production, Vocal tract shape

1. Introduction

Together with the numerical simulations of the vocal tract aero-acoustics, the generation of the vocal tract shape is a crucial step of articulatory speech synthesis. The complete vocal tract – from the glottis up to the lips and including the velopharyngeal opening – must be covered. However, it is not the only input to the articulatory synthesis. First, information about the vibration of vocal folds is required; Elie and Laprie [1] used an improved version of the two-mass model of Ishizaka and Flanagan [2]. However, it is possible to choose reasonably good values easily. Second, information about the glottal opening plays an essential role in the production of fricatives [3] and stops. The latter requires excellent coordination between the closure at the place of articulation and the opening at the glottis during the closure and the rapid opening, which gives rise to the burst. The interest in generating the vocal tract shape goes far beyond the articulatory synthesis. It also concerns, among other topics, the numerical modeling of coarticulation, the study of articulatory intra-speaker variability, and compensation phenomena. One of the first models of the complete vocal tract shape was the articulatory model of Coker [4], which relies on geometric primitives. Then Maeda [5] and others [6, 7] used principal components methods to develop models from X-ray images recorded for continuous speech with the double difficulty of processing low-quality images and having at disposal a tiny number of images (less than 1000) for a minimal amount of speech, i.e., approximately 30 seconds. More elaborated 3D geometric articulatory models were developed and then adjusted to a given speaker by exploiting a set of 3D static MRI images in parallel with models derived from images [8]. Altogether, these models allowed a crucial scientific breakthrough in understanding the phenomena of compensation and use for articulatory synthesis. However, the exploitation of these models faces a significant difficulty: controlling the set of parameters over time. Whether a geometrical or a statistical model, the vocal tract shape is defined by a vector of parameters, which requires interpolation over time to obtain the profile at each time step.

Several strategies have been proposed to predict the vocal tract shape during continuous speech. Öhman [9] presented one of the firsts, which consists of superimposing the effect of fast constriction consonant gestures on a sequence of continuous vocalic gestures. A weight describing the degree of resistance of a consonant enables the place and degree of articulation to be reached. Since the articulatory-to-acoustic transformation is non-linear, a simple linear interpolation is insufficient, and several methods use second-order filter techniques. Articulatory phonology [10] provides a theoretical framework for representing speech using constricting events (gestures), which target one particular articulator, e.g., lips, tongue tip, tongue body, velum, and glottis. The activation of these articulators corresponds to gesture scores, and their determination is the keystone of implementing task dynamics within articulatory phonology. Nam et al. [11] provides an answer by exploiting the University of Wisconsin X-ray Microbeam Database (XRMB) [12]. XRMB contains data of many speakers but with a fairly small number of sentences per person due to the hazard of the acquisition technique. The limited database size explains why only the gestures' timing was obtained by warping their onset and offset to minimize the acoustic distance between natural and synthetic speech. However, adjustments to several other parameters would have been necessary. Birkholz et al. [13] proposed tenth-order linear systems to model the dynamics of articulators from a sequence of discrete consonant and vowel targets, with the expected advantage of better-fitting bell-shaped velocity profiles observed in natural movements. The parameters were optimized with an EMA database collected for a corpus of CVCVCVCV occurrences. Despite the excellent fitting on CV sequences, this work raised the issue of choosing an appropriate degree of freedom and approximating more complex phonetic contexts.

The development of sizeable electromagnetic articulography (EMA) databases [14] enabled exploration of acoustic-to-articulatory relationships, the direct and inverse problems, but more often inversion to recover the vocal tract shape from the acoustic signal. Richmond [15] offered the first results in the field by using a Mixture Density Network (MDN), which allows estimating the probability density of the articulators' positions conditioned to the acoustic features. More recently, Biasutto-Lervat and Ouni [16] explored EMA to perform phoneme-to-articulatory mapping with the aid of a Recurrent Neural Network (RNN). The RNN is a class of artificial neural networks connected such that the previous outputs are the inputs of the following iterations keeping hidden states in between. Thus, it is capable of modeling sequential data straightforwardly. The principal issue related to the most primitive RNN architectures is the inability to model long-term dependencies causing the vanishing gradient problem. Alternatively, the Long Short-Term Memory (LSTM) [17] introduces the forget, input, and output gates, which allow controlling the information that flows from previous time-steps to the subsequent ones. The advantage of the LSTM network is the ability to model long-term dependencies at the expense of having a more significant number of parameters. The Gated Recurrent Unit (GRU), proposed by Cho et al. [18], is a type of RNNs that meets a compromise between long-term dependencies and complexity. It contains only two gates, i.e., update and reset, to control the information flow. RNNs usually follow the forward direction in the input sequence, but the backward direction is often beneficial for sequence modeling. Thus, bidirectional versions of GRUs and LSTMs are trendy nowadays. The bidirectionality is achieved by parallel using two GRU (or LSTM) layers. The first model the forward direction, and the second model it backward. In the end, the outputs of the two are concatenated. RNNs became especially popular for acoustic-to-articulatory inversion and coarticulation models. Biasutto et al. [19] model labial coarticulation from the data collected using 44 sensors attached to an actress's face. This last work is particularly interesting for our study since it inspires our network architecture.

In this work, our challenge is beyond the capacity of EMA. Our goal is to produce a complete coarticulatory model of the vocal tract, which requires observing the entire structure, which is impossible with EMA, limited to a few sensors attached on lips, and articulators inside the oral cavity (teeth, tongue tip, and tongue body). In this context, magnetic resonance imaging provides an alternative for complete visualization of the vocal tract without harming the patient's health. On the one hand, current technologies allow the acquisition of MRI images in real-time (rt-MRI) with a decent spatial resolution and appropriate frame rate [20]. On the other hand, raw MRI films do not suffice for modeling articulatory gestures. It requires complex processing to retrieve the contour of each articulator, which demands extensive manual geometric annotation of each articulator. This challenge has been explored through different approaches, with or without deep learning [21, 22]. A second challenge is audio processing since each phone's onset and offset times are required to locate the corresponding images. Forced alignment tools are available to align the speech signal with the audio transcription and extract this information. However, forced alignment tools are usually built to perform automatic speech recognition (ASR), which requires less temporal accuracy. Thus, the automatic phonetic segmentation contains many errors, essentially temporal shifts, while our task demands exact temporal boundaries. We need to have a quasi-perfectly synchronized and phonetically segmented file once we associate the phoneme with the precise corresponding gestures performed by the speaker. Hence, the phonetic segmentation requires manual corrections, which is laborious.

In our most recent work, we proposed the first attempt to predict the vocal tract shape from the phonemes to be articulated [23]. We proposed a deep neural network to predict the positions of five articulators, i.e., the tongue, the upper and lower lips, the soft palate, and the pharyngeal wall. These articulators were tracked frame-by-frame from rt-MRI films. The phonetic segmentations were obtained from the audio recordings using forced alignment and then manual corrections.

Even though the study brings excellent novelty, it presents several limitations. First, it misses many essential articulators, and the whole vocal tract from the glottis to the lips is not available. Second, the small size of the dataset prevents relevant training from being achieved. The corpus contains 107 sentences with repetitions. Even though we show the generalization capacity of the models with few data points and the generated curves are realistic, they do not always fulfill the critical articulators. We expect the critical constraints to be satisfied with a larger dataset with more examples for each phoneme. In this work, we focus on these limitations. We explore the generation of the complete vocal tract, with the addition of articulators and organs in the bottom of the vocal tract, i.e., vocal fold position, thyroid cartilage, arytenoid cartilages, and epiglottis, and upper and lower incisors, which describe the mandible movement and jaw opening.

2. Materials

2.1. Corpus

Along with this work, we extend the research to a larger dataset than that used in [23]. The dataset [24] is composed of one male French native speaker. The MRI sequences were recorded at Max Plank Institute, Göttingen, Germany. The recordings have a frame rate of 55 fps, pixel spacing of 1.412 mm, and an image resolution of 136 × 136 pixels. The audio recordings sampling frequency is 16 000 Hz.

The corpus contains 38 acquisitions, with a median acquisition time of 81.8 seconds, a minimum of 36.3 seconds, and a maximum of 90.1 seconds. The acquisitions have a median of 18 sentences each, a minimum of 12 sentences, and a maximum of 30 sentences. Some of the sentences are repeated during the acquisition. The sentences were selected to provide a phonetically balanced coverage of French. The corpus includes 707 spoken sentences, with 38 unique phonemes plus 6 non-phonetic tokens to represent pauses, hesitations, and noise after /i/, /u/, /y/, and /o/. This corpus represents a total of 125 411 rt-MRI images. In both cases, we used Astali [25] to "force" align the speech with the transcriptions and obtain the phonetic segmentation. An expert carefully manually corrected the phonetic annotations.

2.2. Image and contour characteristics

The images correspond to 8 mm thick slices in the mid-sagittal plane, each image requiring an acquisition time of 18 ms. These characteristics give rise to two types of inaccuracies. The first is related to the fact that the slice volume is only partially occupied by tissues near articulator contours. This effect called "partial volume effect" [26] is particularly marked in the case of the tongue, which sometimes presents a marked groove, for example, for the tongue body during the articulation of the vowel /i/. In this case, the contour is blurred, and we retain the internal shape that best corresponds to the mid-sagittal plane. The second inaccuracy is related to the movement of articulators during the acquisition of an image. This effect mainly concerns the tongue tip during the articulation of the dentals and laterals like /l/ in French. We often observe a "ghost" contour effect, which gives two contours in the vicinity of the tongue tip. We chose the more marked contour and not an intermediate one.

The lower part of the pharyngeal wall is rather challenging to delineate for two main reasons. First, the image possibly does not correspond entirely to the mid-sagittal plane in this part of the vocal tract. Second, the larynx is narrower during phonation due to the rapprochement of the arytenoid cartilages and the edges of the quadrangular membrane. For the front of the larynx, we used the edge of the epiglottis and the thyroid cartilage, which remain visible and can be followed. There is no visible contour for the back of the larynx, and we used the position of the vocal folds and that of the pharyngeal wall, which we arbitrarily connected to represent the arytenoid cartilages. We consider that this approximation corresponds to a minimal error even if it is impossible to evaluate precisely.

2.3. Choice of the articulators

One of the essential points is to represent each articulator independently of the others because each articulator corresponds to a different organ, and their movements and deformations are not necessarily synchronized. This condition is essential in particular to study articulatory compensation phenomena. Another reason is that the nature, essentially the shape of an articulator (elongated or not), or that it does not appear directly on the image, e.g., bone structures, requires specific tracking techniques. We, therefore, consider the tongue from the root to the sublingual cavity, the lower and upper lips, the trace of the lower central incisor root, the velum, the epiglottis, and the vocal folds. We added the hard palate with the upper central incisor, the pharyngeal wall, and finally, the thyroid cartilage and the approximation of the arytenoid muscles to define the bottom of the vocal tract. These contours allow the air column to be fully defined from the glottis to the lips. The root of the upper incisor, the hard palate, and the root of the lower incisor are rigid structures and are therefore tracked using the correlation between a reference image and the images to be processed.

2.4. Tracking articulators

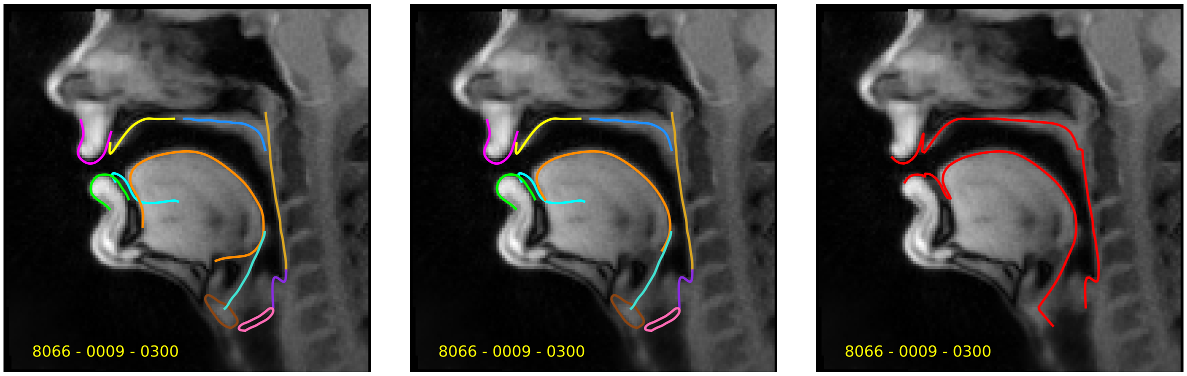

Figure 1: The figure on the left presents a sample of the tracked contours for one MRI frame. The articulators are the upper lip (magenta), the lower lip (lime green), tongue (orange), soft palate mid-line (blue), pharyngeal wall (dark golden), upper incisor (yellow), lower incisor (cyan), vocal folds (pink), thyroid cartilage (brown), epiglottis mid-line (turquoise), and arytenoid muscle (violet). The figure in the center presents the same sample after tail-clipping. The figure on the right represents the vocal tract air column.

This paper does not intend to be exhaustive on the tracking task. This section only describes the procedure followed to acquire the vocal tract shape during speech. We will publish the complete protocol in a future article with the appropriate literature review and evaluation methods.

Many works have been devoted to tracking articulator contours, from the first attempts on X-ray images to deep learning-based techniques. The first approaches mainly used methods based on active contours [27] and the optimization of a criterion corresponding to the articulator contours or homogeneous regions [28, 29]. Machine learning techniques were then developed with the advantage of learning from past examples, notably the "Active Shape Models" used by Labrunie et al. [22]. Other machine learning techniques have been used [30], but in the last few years, deep learning, particularly Convolutional Neural Networks (CNNs), have brought significant progress in terms of performance. Initially proposed for tracking the tongue contour in ultrasound images with autoencoders [31], these techniques have been used for MRI images with the help of U-Net [32] as we did [21]. Other methods like Deep Temporal Regression Network (DTRN), which take into account a series of images, present the advantage of integrating the movement for tracking several articulators [33].

For our purposes, we extended our approach using segmentation networks [21, 34] to all articulators of the vocal tract. The networks used in [23] have been re-trained with a limited number of new contours. In total, we estimate that we delineated a total number of approximately 1 000 outlines in the current dataset (in addition to the shapes of the dataset used in [23], that has 1 000 annotated images), to be compared with the total number of 125 411 images (this dataset) × 8 articulators tracked in this work. At the end of the tracking procedure, the articulators were post-processed with a strategy we call "tail-clipping", which consists of dropping points in the tails of the articulator relative to a specific point. Figure 1 presents one sample of the tracked contours together with its version after tail clipping and the estimated vocal tract air column.

The performance of the tracking is, in general, excellent, and even when there is a mean deviation of approximately 2 mm (calculated with the P2CP distance mentioned later in the text – Equation 2) between the ground truth and the prediction, it is tough to decide which one is more relevant. Figure 2 exhibits 4 cases in which the discrepancy for the tongue is 2 mm or more. It should be noted that the expert was not part of the team that annotated the images used to train the first version of the tracking system (which includes only the tongue). It can be seen that the annotator generally chose a contour somewhat inside the tongue, unlike the experts who participated on [21]. In addition, there are differences at the extremities of the tongue contour either because it is longer at the root or because the shape of the sublingual cavity is different, but this does not affect the relevance of the contours concerning the objective of articulatory synthesis. As far as we are concerned, the deviations that have the most impact on the acoustic modeling, even if they are negligible in absolute value, are those that concern the tongue tip (bottom left of Figure 2) because an error on the existence of contact between the tongue and palate or teeth can change the articulation mode during the training of the vocal tract shape prediction. Thus, we have paid particular attention to this point and corrected a few contours.

3. Methods for generating the vocal tract shape

3.1. Baseline method

Because of the absence of an existing system to compare our approach for these rt-MRI data we propose the phoneme-wise weighted mean contour as a baseline method. The phoneme-wise weighted mean contour consists of two phases. The training phase builds a lookup table with the articulators' contours corresponding to the phoneme and relative position of the frame with respect to the temporal boundaries of the sound produced.

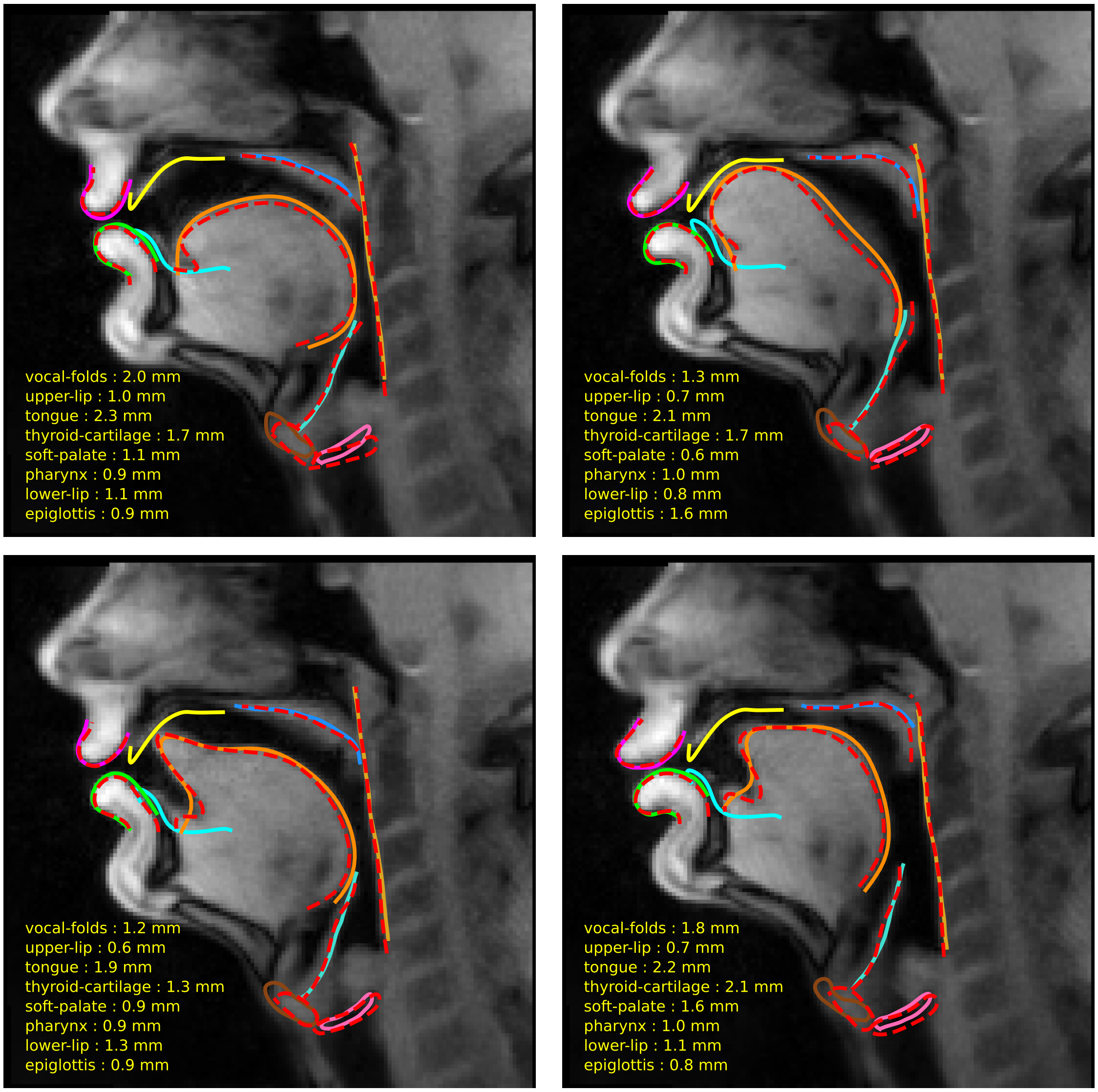

Figure 2: Tracking examples. The red dashed curves are delineated by hand, and the solid ones are provided by the tracking algorithm. The deviations between the two curves are given for each contour.

The inference phase searches for the current phoneme in the lookup table. Then, it calculates the weighted mean contour using the absolute difference between the current phoneme and the lookup phonemes relative positions as the weight. A softmin guarantees that the sum of the weights is one.

3.2. Proposed method

We utilize the same network described in [23], which is inspired by the work of Biasutto et al. [19]. Figure 3 depicts the overall schematic of the architecture, which is an encoder-decoder. The encoder contains two layers of bidirectional GRU followed by a reshaping layer, while the decoder is composed of eleven Articulator Predictor heads. Each head is responsible for estimating the shape of one articulator. The Articulator Predictor contains a sequence of layer normalization and linear layers with `ReLU` activation. In the end, the sigmoid activation guarantees that the outputs are between 0 and 1. Even though the network does not implement any mechanism that ensures the continuity between predicted curves, the GRU learns the temporal dependencies, guaranteeing that our model predicts temporally-consistent curves.

The network takes as input the sequence of phonemes to be articulated. Phoneme duration is encoded by repeating the same phoneme to match the MRI frame duration (18 ms). Therefore, a phoneme of t ms would be repeated $\frac{t}{18}$ times in the input sequence. Each network head outputs N_{samples} (x, y) pairs, each pair corresponding to one sample in the articulator shape. The network output combines the eleven heads outputs for each input token, i.e., the network output will have the shape $(\text{SequenceLength}, N_{\text{articulators}}, 2, N_{\text{samples}})$.

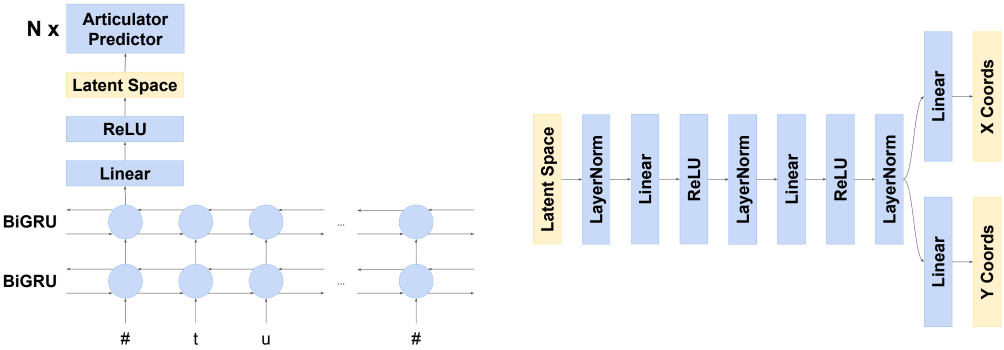

Figure 3: The proposed network architecture. The recurrent encoding path is presented on the left and the Articulator Predictor block on the right. Each RNN cell output is reshaped and fed into the Articulator Predictor. Only the first cell is drawn for simplicity. From Ribeiro et al. [23].

From the 38 acquisitions, we kept eight for testing. We split the remaining 30 acquisitions into training and validation in the 5-fold cross-validation framework, with 24 acquisitions for training and six acquisitions for validation. It is important to stress that the same cross-validation folds were used for the baseline and proposed solutions. Models were trained for 300 epochs with 20 epochs of patience for early stopping. The Adam optimizer [35] was used, with a weight decay of 10^{-6} and a learning rate of 10^{-4}, which was reduced by a factor of ten after ten epochs without improvements in the validation loss. We implemented the code with `PyTorch` [36].

During the acquisition process, the speaker changed his position. Thus, there is a shift in the location of the speaker's head between acquisitions. To make the training invariant to the head positioning, we re-centralize all of the target contours, using the upper incisor as the center of the coordinate system. We fix the upper incisor's root position in the 0-1 grid and adjust the other articulators' coordinates accordingly.

3.3. Evaluation strategy

We evaluate the performance of our models with two distinct strategies. The first is based directly on the generated curves. Models were trained by minimizing the Euclidean distance between the predicted spatial coordinates and the target positions. Note that even though each neural network head makes predictions independently, they are trained jointly, and they share the same latent space. The loss function is given by

$$\mathcal{L}(p, \hat{p}) = \frac{1}{N_{\text{articulators}} \times N_{\text{samples}}} \sum_{i=1}^{N_{\text{articulators}}} \sum_{j=1}^{N_{\text{samples}}} d(p_{ij}, \hat{p_{ij}}) \tag{1}$$

where p and $\hat{p}$ are the ground truth and the predicted points, respectively, $N_{\text{articulators}}$ is the number of articulators, N_{samples} is the number of predicted spatial contour samples, and d is the Euclidean distance. In our case, $N_{\text{articulators}} = 11$ and $N_{\text{samples}} = 50$.

During the test phase, we measured the point-to-closest-point [22] given between the model's outputs and the target curve and the Pearson's correlation between the predicted and the target trajectories of the articulators' curves over time.

The point-to-closest-point distance is given by

$$p2cp_{\text{mean}}(U, V) = \frac{1}{2}(\frac{1}{n}\sum_{i=1}^{n} \min_{j \in \{1,2,...,n\}} d(v_i, u_j) + \frac{1}{m}\sum_{i=1}^{m} \min_{j \in \{1,2,...,m\}} d(u_i, v_j)) \tag{2}$$

where $U = \{u_1, u_2, ..., u_n\}$ and $V = \{v_1, v_2, ..., v_m\}$ are the target and predicted contours with 2D coordinates with n and m points, respectively, and $d(u_i, v_j)$ is the Euclidean distance between points u_i and v_j.

Equation 2 is slightly modified from the proposed by Labrunie et al. [22]. We generalize them to arrays of arbitrary sizes to avoid resampling in specific cases where the predicted curve and the ground truth sizes could not match.

We should stress that the contours used as the ground truth are automatically delineated by the algorithm described in subsection 2.4. Despite its outstanding robustness, some contours might be erroneous. We should also stress that the shape of the arytenoid muscle and the lower and upper incisors are arbitrarily drawn and do not correspond to the actual shape.

The second evaluation strategy is based on the vocal tract variables (TV). The vocal tract variables are measurements made in specific points of the vocal tract, representing the constriction between the articulators. Figure 4 presents a visual representation of the TVs, and Table 1 presents their names and the associated constrictors. Speech gestures are intended to reach articulatory goals are defined in terms of TV [37].

The TVs' trajectories must reflect critical articulators. Critical articulators are those whose position is imposed to achieve the target place of articulation. They are resistant

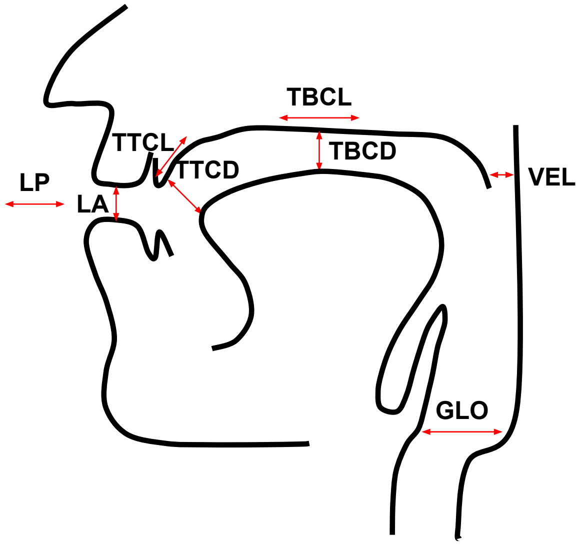

Figure 4: Visual representation of the vocal tract variables. Inspired by Saltzman and Munhall [39].

to context and have co-articulatory effects on neighboring phones [38]. For example, lip closure, i.e. a full contact between lips, is mandatory for phonemes /p/ and /b/. Similarly, the tongue tip must touch the teeth to produce the phonemes /t/, and /d/. For /k/, the tongue dorsum must approach the palate, and for nasal vowels, the soft palate must lower to let the air go through the nasal cavities. Thus, we measure the four closely connected TVs, i.e., LA, TTCD, TBCD, and VEL. We calculate the Pearson's correlation between the target and the predicted TVs' trajectories.

The *t*-test is used to account for the statistical significance of the results. To make the values more interpretable, the measures in the $[0, 1]$ grid are converted to millimeters by multiplying them by image resolution and pixel spacing.

Table 1: Vocal tract variables and their associated constrictors. Reproduced from Browman and Goldstein [37].

	Tract Variable	Constrictors
LP	Lip protrusion	Upper & lower lips, jaw
LA	Lip aperture	Upper & lower lips, jaw
TTCL	Tongue tip constrict location	Tongue tip, tongue body, jaw
TTCD	Tongue tip constrict degree	Tongue tip, tongue body, jaw
TBCL	Tongue body constrict location	Tongue body, jaw
TBCD	Tongue body constrict degree	Tongue body, jaw
VEL	Velum aperture	Velum
GLO	Glottal aperture	Glottis

4. Results

The direct evaluation of the vocal tract shapes generated by our system consists in calculating their distance with those produced by the subject and implicitly relies on the hypothesis of articulatory uniqueness. For isolated sounds or sounds considered independently of their context, this hypothesis is false for theoretical reasons (as confirmed by Atal et al. [40] and more recently by our inversion experiments [41]) and practical observations. However, as soon as anatomical constraints are taken into account and continuous speech is considered, and in the absence of constraints limiting the movement of articulators, the variability of gestures remains reduced and amounts to minor compensatory effects. In addition, the recording conditions in the MRI very strongly restrict head movements and do not favor the expressions, which both could increase articulatory variability. Thus, the distance between the automatically generated vocal tract shapes and those recorded provides the first evaluation of our approach.

Table 2 presents the mean point-to-closest-point distances and the x- and y-correlations along time per articulators. Figure 7 offers samples of the predicted curves for a selection of phonemes that have critical articulators. Figure 5 presents a comparison between the baseline and the proposed methods when predicting vocal tract shapes for vowels (left) and consonants (right). Figure 6 compares vowels and consonants predicted by the baseline (left) and the proposed methods (right).

The second evaluation exploits vocal tract variables which allow measuring how well critical articulators reach their targets for the production of consonants and some vowels. The achievement of the correct position of those articulators is crucial to guarantee the expected acoustic properties. Table 3 shows the mean and standard deviation, the minimal and the maximum correlations per TV. Figure 8 presents the trajectories of each TV and its corresponding phonemes for four test sentences. Figure 9 presents the correlation distribution for each TV. Figure 10 shows the predicted and target trajectories with the lowest correlations for each TV. Figure 11 presents the ground truth, the baseline, and the proposed methods' predictions for the same sentence.

The supplementary material contains videos of the vocal tract shape prediction (baseline and proposed) for sentences out of the train/validation/test distribution. Two sentences in French that are not part of the corpus were recorded and phonetically annotated. Then, the vocal tract shape was generated using the method proposed by this work. Note that the audio in the video is auxiliary and corresponds to the actual recording. The model does not generate audio.

5. Discussion

From the individual articulators' perspective, Table 2 shows that our model can produce the complete vocal tract shapes with tiny errors, consistently beating the baseline in terms of x and y-correlations. In the few cases (lower incisor, pharyngeal wall, and upper incisor) where the baseline drives better mean P2CP distances than the proposed method, the difference is minimal (< 0.2 mm) and without statistical significance. It is important to stress that once the upper incisor's position is kept as the reference, zero correlations and an error very close to zero are expected.

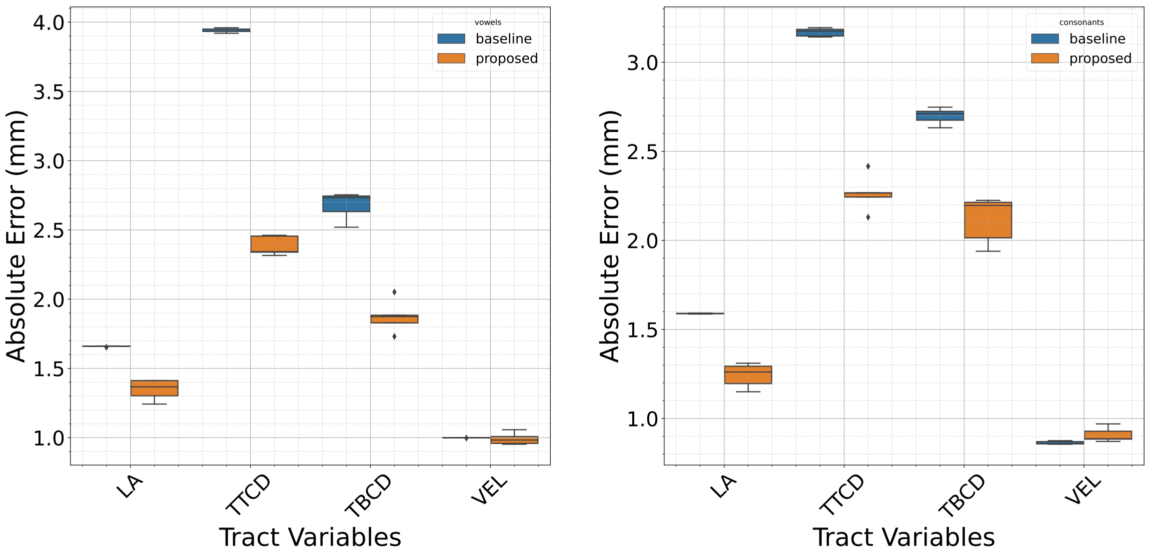

Figure 5: Comparison between the TVs absolute errors for vowels (left) and consonants (right) predicted by the baseline and the proposed methods.

In Figure 7, we notice that the targets are less well reached for short and voiced consonants, such as /l/ or /d/ for example. The sampling rate probably explains most of the deviations from the expected target. Indeed, each frame corresponds to a time interval of 18 ms, which is almost half the duration of the /l/ sound in most intervocalic /l/ cases. Similarly, the closure of /d/ is shorter than that of /t/. From a learning point of view, this means that there are both fewer examples than for the generally longer unvoiced consonants and that the forms considered are more likely to have been acquired at the boundary between the target sounds, /l/ or /d/ for example, and the neighboring sounds resulting in less characteristic articulatory positions.

From the tract variables perspective, Table 3 shows that the proposed model systematically beats the baseline by a large margin. For LA, TTCD, and TBCD, the correlation provided by our method is considered high (> 0.8), while for VEL it is considered low (< 0.5). Figure 9 shows that in all of the TVs, there is a substantial density above 0.5. Even though most TV correlations are acceptable, they have low minimal values. Hence, it is important to pay attention to Figure 10. We observe that for all TVs, the worst cases happen when that TV is free, i.e., it is not critically attached to the phonemes.

We hypothesized that for consonants with well-defined places of articulation, the weighted mean contour would be a satisfactory estimator and may even beat the neural network. However, for vowels, which have more degrees of freedom, the weighted mean contour would not capture all of the variability, being a poor estimator. Therefore, we expected the neural network to provide better predictions in these cases.

Figure 5 shows that the baseline model is not mainly a better estimator than the proposed model for either vowels or consonants. The neural network provides a better estimation for three of the four tract variables. The difference between the methods is negligible for the velum, which has two positions, open and closed. According to our expectations, Figure 6 shows that the models generally offer a better estimate for consonants than for vowels, a

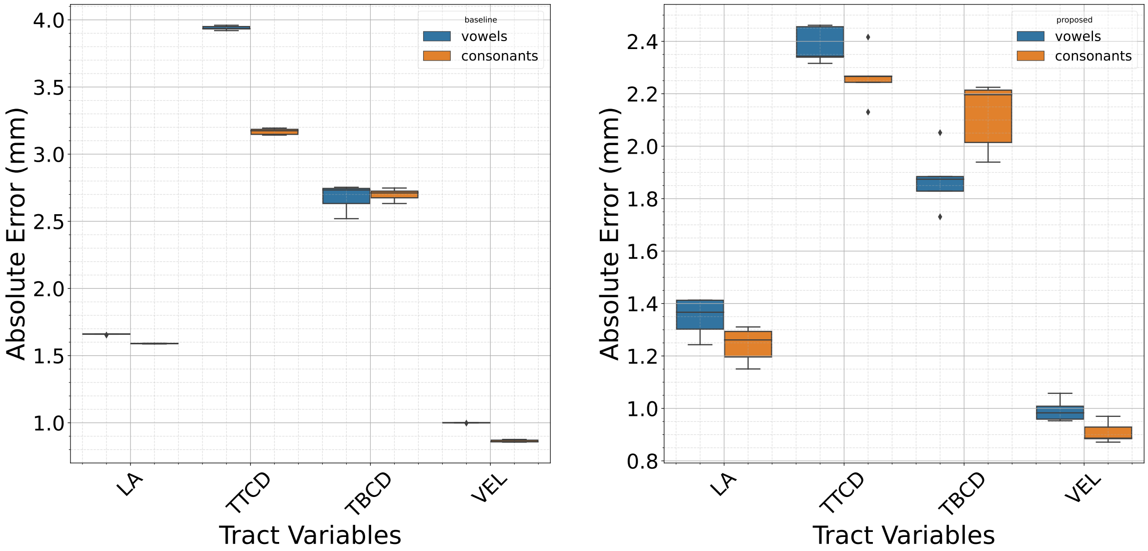

Figure 6: Comparison between the TVs absolute errors for vowels and consonant predicted by the baseline (left) and the proposed (right) methods.

behavior that is shared between the baseline and the proposed methods. However, it is noticeable that for the TBCD, the estimation for the consonants is either equal (baseline) or worse (proposed).

The trajectories depicted in Figure 8 show that most of the TVs that are attached to critical articulators follow the correct path. However, it is not completely adequate. On the one hand, we observe a lip closure (LA) for /p/, /b/, and /m/, a contact between the tongue tip and the upper incisor (TTCD) for /t/, and an increase in the gap between the velum and the pharyngeal wall (VEL) for nasal vowels. On the other hand, for short phonemes, such as /l/ and /d/, the tongue tip approaches the upper incisor, but there is no contact. Also, for /k/, the trajectories reveal the difficulty of the model to estimate TBCD for consonants. We can find cases where the gap between the tongue body and the palate is not fully closed.

6. Conclusions

To the best of our knowledge, this is the first machine learning technique approach that allows the complete vocal tract shape, including all articulators and motionless walls from the glottis to the lips, to be generated for any sequence of phonemes to be articulated. The very accurate phonetic segmentation of the speech signal, the successful tracking of contours in the rt-MRI films, and the training on a database of more than 125 000 images allow relevant vocal tract shapes to be generated in the positions of critical articulators and temporal trajectories.

We examined the results to verify their phonetic relevance to the speaker's trajectories qualitatively, and we observed minor articulatory compensation effects between the jaw opening and the lips.

However, there are some limitations. One is related to the automatic tracking of articulators and the nature of rt-MRI images, which are not captured instantaneously. The tracking

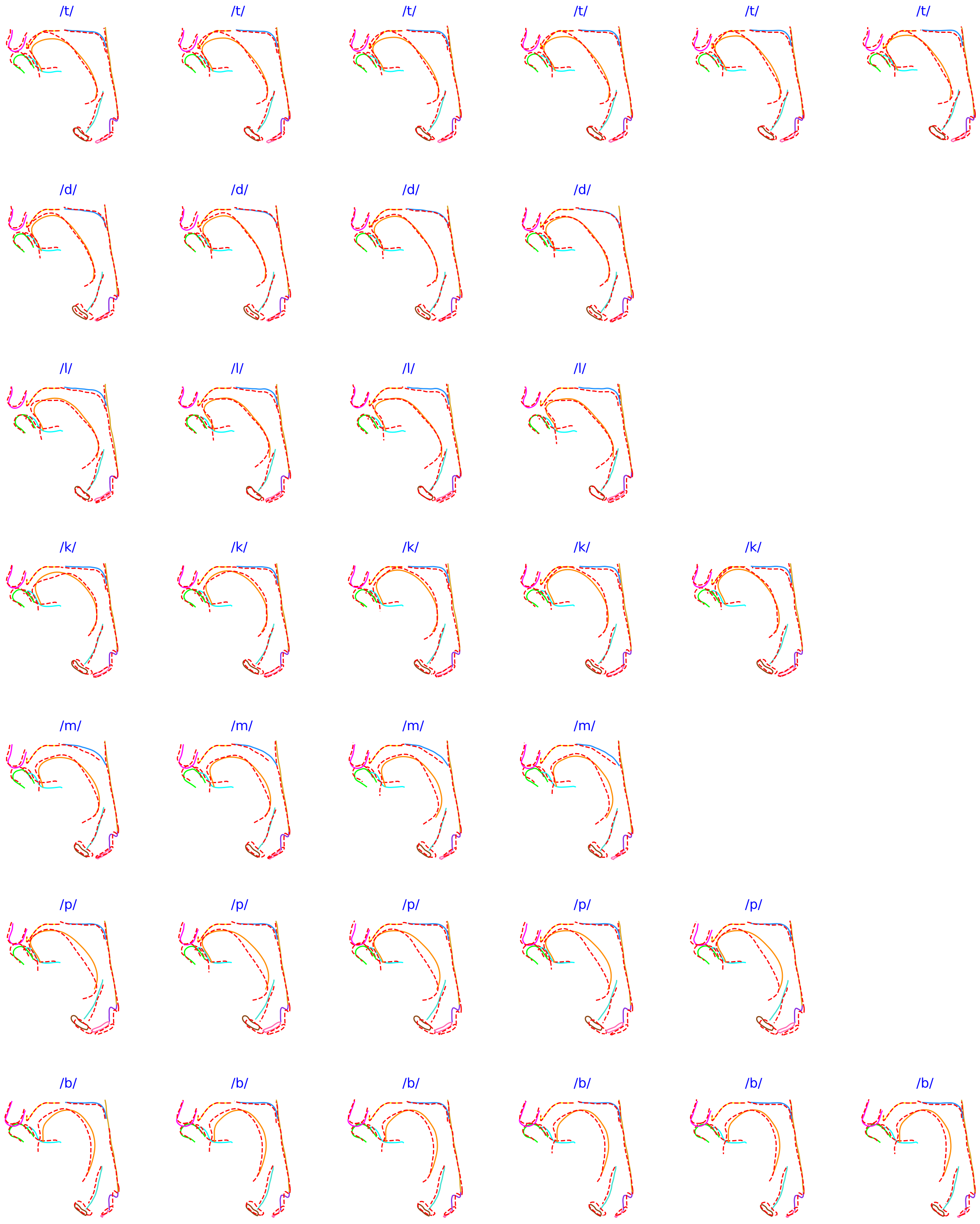

Figure 7: Samples of the proposed method predictions for the phonemes /t/, /d/, /l/, /k/, /m/, /p/, /b/, and in order of appearance from top to bottom. The solid lines represent model predictions. The red dashed line represents the ground truth. Each row represents one phoneme occurrence, and the columns represent the time steps (18 ms).

model sometimes fails to reconstruct the tongue tip for short dental sounds, e.g., /d/ and /l/. These errors harm the learning of the generative model and introduce an upper bound of the performance.

Two approaches to adapting the predictions to a new speaker can be considered. The first is the one used to adapt an articulatory model to another speaker as it was done by Maeda [42] and Laprie and Busset [43]. It consists in deforming, essentially by changing the length of the cavities of the mouth and pharynx on the one hand and the angle between these two cavities on the other hand. This transformation is fairly similar to the one adopted by Birkholz [44] to adapt his geometrical model, which also uses the hard palate shape and the temporo-mandibular joint geometrical parameters as well. This kind of transformation preserves the contact of the critical articulators for the consonants, and our method of generating vocal tract shapes should therefore work for other speakers based on some static MRI images of the vocal tract in order to perform this geometrical adaptation. In the absence of additional rt-MRI dynamic articulatory data, more subtle differences in articulatory strategy can obviously not be taken into account. Beside those adaptation procedures based on static images of the vocal tract there exist image-free adaptation techniques based on vowel's acoustic data with the drawback of being less accurate [45].

A second possible way would consist of exploiting a more extensive multi-speaker database either by developing an anatomical embedding or by constructing a generic speaker model via an anatomical atlas approach. The second approach has the advantage of requiring a limited number of subjects. Douros [46] developed a dynamic articulatory atlas using dynamic MRI, which exploits 12 male and female speakers. The approach first relies on a geometric fit that neutralizes the anatomical particularities of the speakers. More precisely, images of the same articulatory configuration of the vocal tract are used to determine for each pair of speakers a geometrical transformation from one to the other. The average transformation is used to define the anatomy of the generic speaker. It is then possible to normalize the dynamic part by aligning the dynamic data. This second part requires a very high temporal accuracy of the phonetic segmentation and thus a very long manual control step.

Articulatory synthesis has many potential applications, and its interest lies in the possibility of linking the dynamic shape of the vocal tract and other physical parameters to the acoustic signal. It allows studying the acoustic impact of perturbations affecting the vocal tract shape or motor control in medical applications, investigating the articulatory origin of expressions, or developing articulatory and acoustic feedback for speech therapy. One of the problems has been the determination of the vocal tract shape for the sequence of phonemes in a sentence. This step is often done by hand, which is an extreme limitation. This work provides an effective solution to this problem and thus should extend the application areas of articulatory synthesis.

7. Acknowledgements

This research was partly supported by the French ANR project Full3DTalkingHead.

Table 2: Mean euclidean distance, x- and y-correlations for each predicted articulator. We include the metrics for the arytenoid muscle and the lower and upper incisors for completeness. However, their target shapes are approximations. Bold values represent the best result between the baseline and the proposed methods. The † symbol indicates statistical significance ($p < 0.05$).

Articulator	Method	Mean P2CP distance (mm)	ρ_x	ρ_y
Arytenoid muscle*	Baseline	2.740 ± 0.043	0.084 ± 0.007	0.372 ± 0.002
	Proposed	**2.622 ± 0.102** †	**0.243 ± 0.035** †	**0.824 ± 0.008** †
Epiglottis	Baseline	2.403 ± 0.062	0.317 ± 0.008	0.353 ± 0.002
	Proposed	**2.182 ± 0.090** †	**0.521 ± 0.011** †	**0.801 ± 0.015** †
Lower incisor*	Baseline	**2.065 ± 0.105**	0.162 ± 0.016	0.218 ± 0.007
	Proposed	2.239 ± 0.141	**0.237 ± 0.022** †	**0.362 ± 0.010** †
Lower lip	Baseline	1.667 ± 0.081	0.273 ± 0.010	0.520 ± 0.009
	Proposed	**1.551 ± 0.080**	**0.576 ± 0.009** †	**0.725 ± 0.016** †
Pharyngeal wall	Baseline	**1.025 ± 0.027**	0.108 ± 0.014	0.246 ± 0.005
	Proposed	1.038 ± 0.108	**0.322 ± 0.033** †	**0.591 ± 0.014** †
Soft palate	Baseline	2.590 ± 0.185	0.323 ± 0.009	0.287 ± 0.013
	Proposed	**2.500 ± 0.173**	**0.594 ± 0.021** †	**0.571 ± 0.012** †
Thyroid cartilage	Baseline	2.478 ± 0.017	0.100 ± 0.005	0.333 ± 0.003
	Proposed	**2.355 ± 0.168**	**0.224 ± 0.011** †	**0.790 ± 0.005** †
Tongue	Baseline	3.662 ± 0.100	0.469 ± 0.003	0.424 ± 0.003
	Proposed	**3.437 ± 0.186** †	**0.661 ± 0.010** †	**0.678 ± 0.007** †
Upper incisor*	Baseline	**0.019 ± 0.000** †	0.000 ± 0.000	0.000 ± 0.000
	Proposed	0.118 ± 0.073	0.000 ± 0.000	0.000 ± 0.000
Upper lip	Baseline	1.109 ± 0.040	0.286 ± 0.014	0.430 ± 0.008
	Proposed	**1.077 ± 0.028**	**0.493 ± 0.022** †	**0.524 ± 0.007** †
Vocal folds	Baseline	3.227 ± 0.048	0.153 ± 0.004	0.332 ± 0.001
	Proposed	**3.127 ± 0.084** †	**0.343 ± 0.012** †	**0.763 ± 0.008** †

Table 3: Correlation between the predicted and the target trajectories for each measured tract variable. Bold values represent the best result between the baseline and the proposed methods. The † symbol indicates statistical significance ($p < 0.05$).

Tract Variable	Method	Correlation	Min Correlation	Max Correlation
LA	Baseline	0.621 ± 0.0004	-0.272	0.965
	Proposed	$\mathbf{0.790 \pm 0.016}$ †	$\mathbf{-0.152}$	**0.988** †
TTCD	Baseline	0.568 ± 0.007	-0.921	0.862
	Proposed	$\mathbf{0.784 \pm 0.014}$ †	$\mathbf{-0.480}$ †	**0.973** †
TBCD	Baseline	0.654 ± 0.003	-0.376	0.876
	Proposed	$\mathbf{0.813 \pm 0.018}$ †	$\mathbf{-0.185}$	**0.984** †
VEL	Baseline	0.325 ± 0.003	-0.582	**0.863** †
	Proposed	$\mathbf{0.398 \pm 0.023}$ †	$\mathbf{-0.415}$ †	0.839

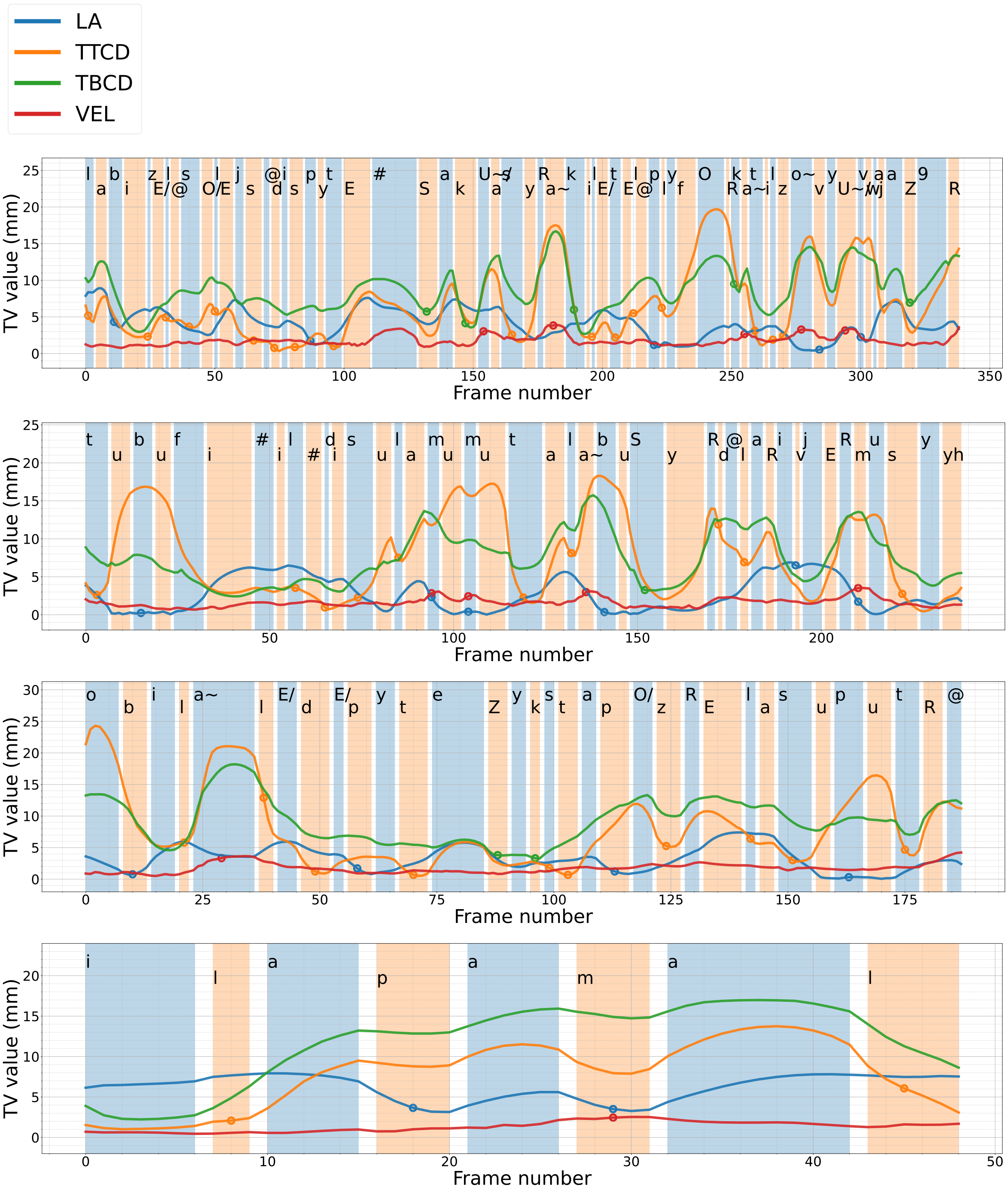

Figure 8: Trajectories predicted by the proposed method for different sentences in the test dataset. Each curve represents the one measured TV. The alternating colors and the vertical shift between the phonetic labels are used only to facilitate the visualization of phonetic intervals and have no special meaning. The circles indicate that the corresponding TV is critical for that phoneme.

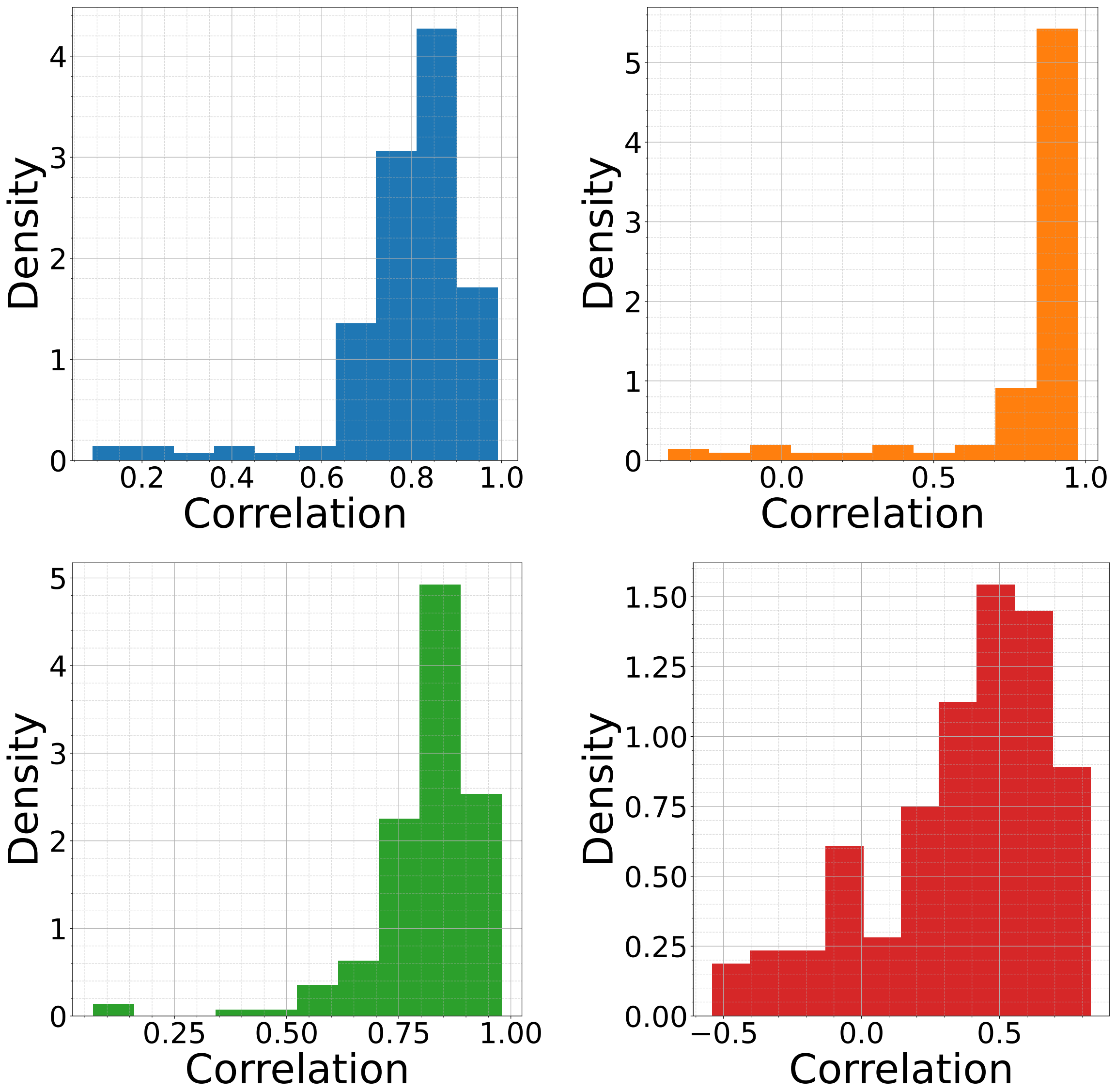

Figure 9: Distributions of correlations between predicted and target trajectories of the TVs in the test dataset for the first fold. From top to bottom and left to right, the TVs are LA, TTCD, TBCD, and VEL. The distributions corresponds to the **proposed method**.

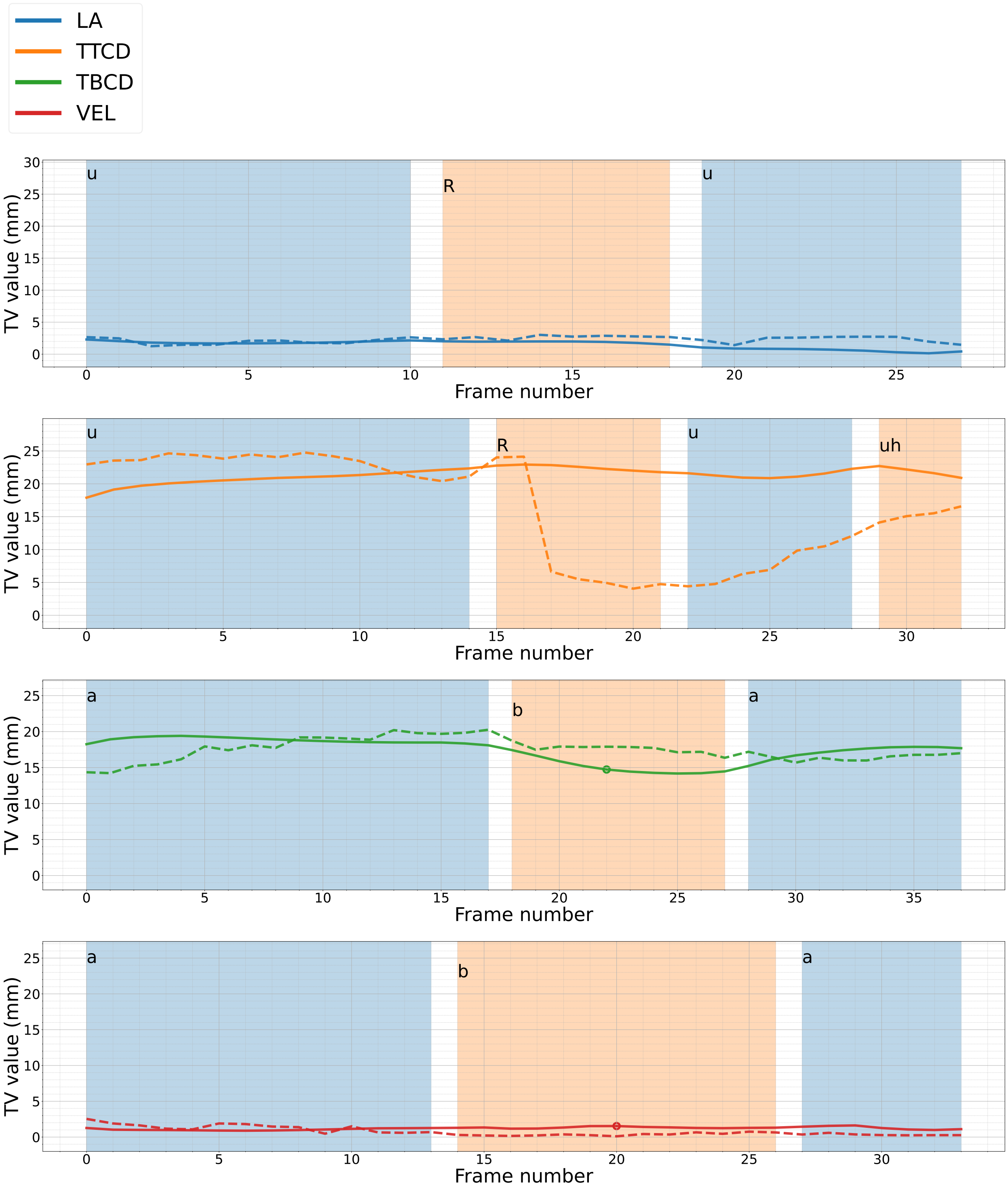

Figure 10: Trajectories with lowest correlation for each measured vocal tract variable for the first fold. From top to bottom, the TVs are LA, TTCD, TBCD, and VEL. The solid lines corresponds to the predictions and the dashed lines corresponds to the target trajectories. The alternating colors and the vertical shift between the phonetic labels are used only to facilitate the visualization of phonetic intervals and have no special meaning. The circles indicate that the corresponding TV is critical for that phoneme. The predicted trajectories corresponds to the proposed method.

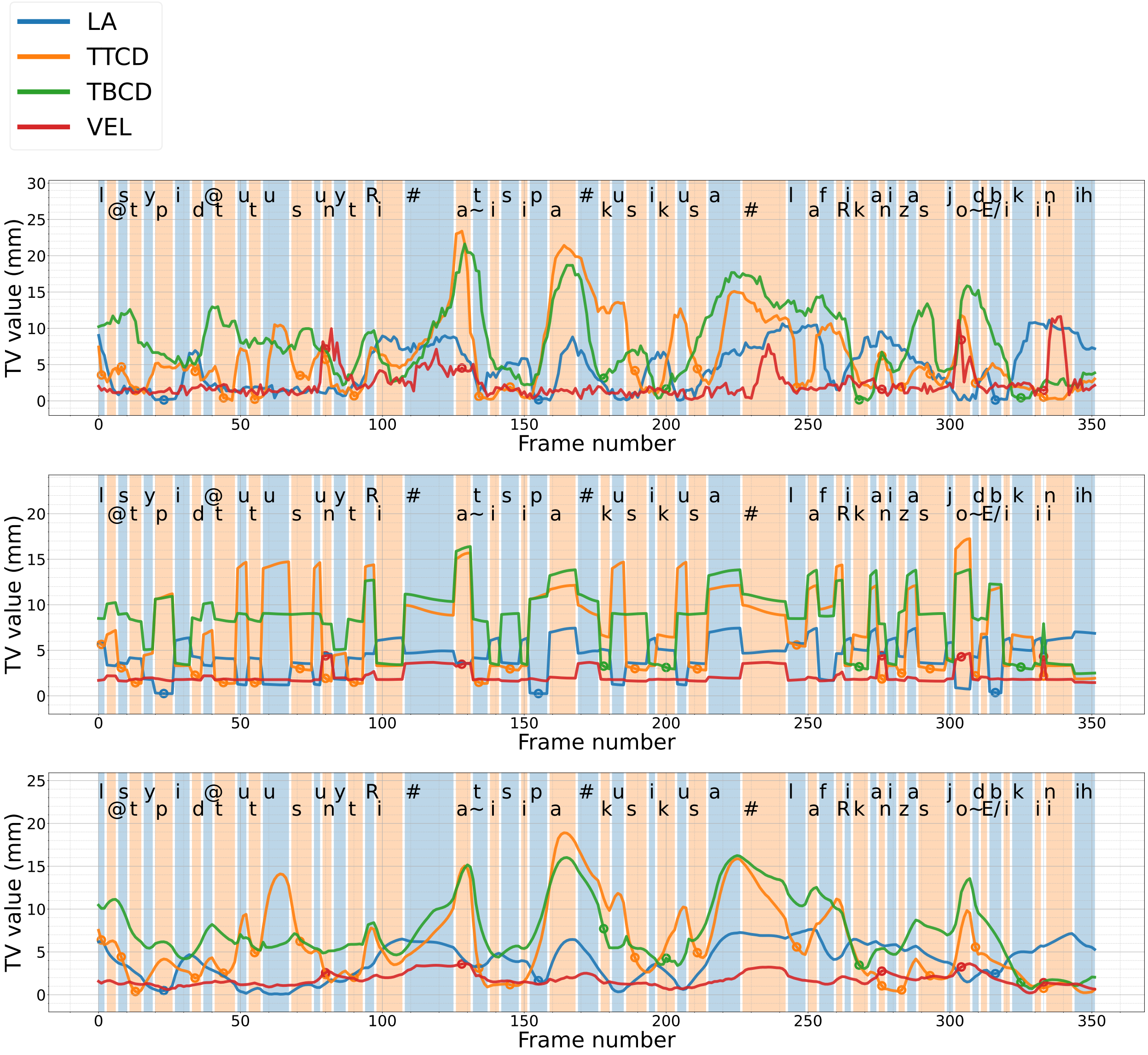

Figure 11: Ground truth, baseline method prediction and proposed method predictions, respectively, for the sentence *“Le stupide toutou sous-nutri anticipa couci-couça l’africanisation des bikinis.”*

References

[1] B. Elie and Y. Laprie. Copy synthesis of running speech based on vocal tract imaging and audio recording. In *22nd International Congress on Acoustics (ICA)*, Buenos Aires, Argentina, September 2016. URL `https://hal.archives-ouvertes.fr/hal-01372310`.

[2] K. Ishizaka and J. L. Flanagan. Acoustic properties of a two-mass model of the vocal cords. *Bell Syst. Technol. J.*, 51:1233–1268, 1972.

[3] B. Elie and Y. Laprie. Acoustic impact of the gradual glottal abduction on the production of fricatives: A numerical study. *Journal of the Acoustical Society of America*, 142(3):1303–1317, September 2017. doi: 10.1121/1.5000232. URL `https://hal.archives-ouvertes.fr/hal-01423206`.

[4] C. H. Coker. A model of articulatory dynamics and control. *Proceedings of the IEEE*, 64(4):452–460, 1976.

[5] S. Maeda. Un modèle articulatoire de la langue avec des composantes linéaires. In *Actes 10èmes Journées d'Etude sur la Parole*, pages 152–162, Grenoble, May 1979.

[6] 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. *Journal of the Acoustical Society of America*, 109(5):2165–2180, 2001.

[7] B. Potard, Y. Laprie, and S. Ouni. Incorporation of phonetic constraints in acoustic-to-articulatory inversion. *The Journal of the Acoustical Society of America*, 123(4): 2310–2323, 2008.

[8] 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*, pages 2597–2600, Aug 2003.

[9] S. EG Öhman. Coarticulation in vcv utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America*, 39(1):151–168, 1966.

[10] C. P. Browman and L. Goldstein. Articulatory gestures as phonological units. *Phonology*, 6:201–251, 1989.

[11] H. Nam, V. Mitra, M. Tiede, E. Saltzman, L. Goldstein, C. Espy-Wilson, and M. Hasegawa-Johnson. A procedure for estimating gestural scores from natural speech. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[12] J. Westbury, P. Milenkovic, G. Weismer, and R. Kent. X-ray microbeam speech production database. *The Journal of the Acoustical Society of America*, 88(S1):S56–S56, 1990.

[13] P. Birkholz, B. J. Kroger, and C. Neuschaefer-Rube. Model-based reproduction of articulatory trajectories for consonant–vowel sequences. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1422–1433, 2010.

[14] K. Richmond, P. Hoole, and S. King. Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[15] K. Richmond. Preliminary inversion mapping results with a new ema corpus. 2009.

[16] T. Biasutto-Lervat and S. Ouni. Phoneme-to-articulatory mapping using bidirectional gated rnn. In *Interspeech 2018-19th Annual Conference of the International Speech Communication Association*, 2018.

[17] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[18] 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.

[19] T. Biasutto, S. Dahmani, S. Ouni, et al. Modeling labial coarticulation with bidirectional gated recurrent networks and transfer learning. In *INTERSPEECH 2019-20th Annual Conference of the International Speech Communication Association*, 2019.

[20] M. Uecker, S. Zhang, D. Voit, A. Karaus, KD Merboldt, and J. Frahm. Real-time mri at a resolution of 20 ms. *NMR in Biomedicine*, 23(8):986–994, 2010.

[21] K. Isaieva, Y. Laprie, N. Turpault, A. Houssard, J. Felblinger, and PA Vuissoz. Automatic tongue delineation from mri images with a convolutional neural network approach. *Applied Artificial Intelligence*, 34(14):1115–1123, 2020.

[22] M. Labrunie, P. Badin, D. Voit, A. A. Joseph, J. Frahm, L. Lamalle, C. Vilain, and LJ Boë. Automatic segmentation of speech articulators from real-time midsagittal mri based on supervised learning. *Speech Communication*, 99:27–46, 2018.

[23] V. Ribeiro, K. Isaieva, J. Leclere, PA Vuissoz, and Y. Laprie. Towards the Prediction of the Vocal Tract Shape from the Sequence of Phonemes to be Articulated. In *Proc. Interspeech 2021*, pages 3325–3329, 2021. doi: 10.21437/Interspeech.2021-184.

[24] I. Douros, J. Felblinger, J. Frahm, K. Isaieva, A. Joseph, Y. Laprie, F. Odille, A. Tsukanova, D. Voit, and PA Vuissoz. A multimodal real-time mri articulatory corpus of french for speech research. In *INTERSPEECH 2019-20th Annual Conference of the International Speech Communication Association*, 2019.

[25] Astali. URL `http://ortolang108.inist.fr/astali/fr/`.

[26] E. Bellon, M. Haacke, P. Coleman, D. Sacco, DA Steiger, and R. Gangarosa. Mr artifacts: A review. *AJR. American journal of roentgenology*, 147:1271–81, 12 1986. doi: 10.2214/ajr.147.6.1271.

[27] Y. Laprie and B. Marie-Odile. Extraction of tongue contours in x-ray images with minimal user interaction. pages 268 – 271 vol.1, 11 1996. ISBN 0-7803-3555-4. doi: 10.1109/ICSLP.1996.607097.

[28] J. Kim, N. Kumar, S. Lee, and S. Narayanan. Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data. In *Proceedings of 10th Int. Seminar Speech Prod., Köln, Germany*, pages 222–225, 2014.

[29] E. Bresch and S. Narayanan. Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. *IEEE Transactions on Medical Imaging*, 28(3):323–338, 2009. doi: 10.1109/TMI.2008.928920.

[30] H. Takemoto, T. Goto, Y. Hagihara, S. Hamanaka, T. Kitamura, Y. Nota, and K. Maekawa. Speech organ contour extraction using real-time mri and machine learning method. In *INTERSPEECH*, 2019.

[31] A. Jaumard-Hakoun, K. Xu, P. Roussel-Ragot, G. Dreyfus, M. Stone, and B. Denby. Tongue contour extraction from ultrasound images based on deep neural network. In *The International Congress of Phonetic Sciences*, Glasgow, United Kingdom, August 2015. URL `https://hal.archives-ouvertes.fr/hal-01366237`.

[32] 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*, pages 234–241. Springer, 2015.

[33] S. Asadiabadi and E. Erzin. Vocal tract contour tracking in rtmri using deep temporal regression network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3053–3064, 2020. doi: 10.1109/TASLP.2020.3036182.

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

[35] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[36] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, A. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[37] C. P. Browman and L. Goldstein. Articulatory phonology: An overview. *Phonetica*, 49 (3-4):155–180, 1992.

[38] S. Silva and A. J. S. Teixeira. Critical articulators identification from rt-mri of the vocal tract. In *INTERSPEECH*, pages 626–630, 2017.

[39] E. L. Saltzman and K. G. Munhall. A dynamical approach to gestural patterning in speech production. *Ecological psychology*, 1(4):333–382, 1989.

[40] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *JASA*, 63 (5):1535–1555, May 1978.

[41] S. Ouni and Y. Laprie. Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. *JASA*, 118(1):444–460, 2005.

[42] S. Maeda. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In W. J. Hardcastle and A. Marchal, editors, *Speech production and speech modelling*, pages 131–149. Kluwer Academic Publisher, Amsterdam, 1990.

[43] 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, August 2011.

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

[45] B. Potard and Y. Laprie. Automatic adaptation of a vocal tract model. In *Proceedings EUSIPCO 2010 18th European Signal Processing Conference, Aalborg, Denmark*, 2010.

[46] I. Douros. *Towards a 3 dimensional dynamic generic speaker model to study geometry simplifications of the vocal tract using magnetic resonance imaging data*. PhD thesis, Université de Lorraine, 2020. URL `https://hal.univ-lorraine.fr/tel-03008224`.