INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Generating high-resolution 3D real-time MRI of the vocal tract

Martin Strauch[1], *Antoine Serrurier*[2]

[1]Institute of Imaging & Computer Vision, RWTH Aachen University, Aachen, Germany
[2]Clinic for Phoniatrics, Pedaudiology and Communication Disorders,
University Hospital and Medical Faculty of RWTH Aachen University, Aachen, Germany

martin.strauch@lfb.rwth-aachen.de, aserrurier@ukaachen.de

Abstract

MRI recordings of the vocal tract allow researchers to obtain anatomical cross-sections in a non-invasive way, providing an important tool for speech production research. Acquiring MRI at equally high temporal and spatial resolution remains, however, challenging. We propose an image processing method for synthesising a real-time high spatial resolution 3D movie given real-time 2D MRI and static high spatial resolution 3D MRI data from the same speaker. We evaluate our method on a public dataset with 17 speakers, showing that a real-time 2D movie of the vocal tract during a speech task can be encoded by combinations of a small number of its frames. These combinations can be transferred to the domain of the high spatial resolution 3D data with static vocal tract articulations matched to frames of the 2D movie, synthesising a 3D movie of the speech task. Our algorithmic method provides a generic approach that can complement technical improvements of the acquisition process.

Index Terms: 3D real-time MRI, speech production, vocal tract

1. Introduction

In speech, the sound is produced by the vibration of the vocal folds and the configuration of the vocal tract. Analysing the shapes taken by the vocal tract is therefore crucial for understanding speech production mechanisms. The emergence of the non-invasive and non-ionising Magnetic Resonance Imaging (MRI) technology for speech production in the 1990s was a breakthrough for articulatory studies [1]. It allowed the safe acquisition of anatomical cross-sections of the vocal tract with visualisation of the soft and hard tissues. However, the long acquisition time forced the speaker to sustain an articulation for several seconds without movement, far from real speech experience.

A second breakthrough occurred in the 2000s and became very popular for speech production studies: In real-time MRI [2], frames are acquired at a high temporal resolution, typically several tens of images per second [3], enough to record dynamic speech. Although this was a major improvement, it came at the cost of lower image quality, resolution and anatomical accuracy, as well as the loss of 3D information.

Improving these parameters for real-time MRI of the vocal tract is the object of ongoing research. Almost all studies focus on the MRI acquisition process [4]. While this may be the most intuitive approach, it depends on the progress in MRI technology and appears technically challenging. Here, we propose an alternative approach that relies on image processing: Combining high temporal but low spatial resolution 2D MRI data with high spatial resolution 3D MRI data to obtain 3D MRI data with both high temporal and high spatial resolution. An initial, semi-automatic attempt has already been proposed [5], but our study introduces a more generic approach, fully automatic and applicable also to long speech recordings as opposed to the short vowel/consonant sequences in [5].

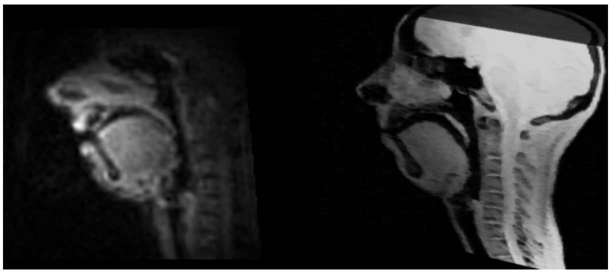

Figure 1: *Left: Frame from a 2D low-resolution real-time movie (USC data [6]). Right: The same frame from the synthesised 3D high-resolution real-time movie (midsaggital view).*

Figure 1 illustrates our method: We replace a frame of the real-time MRI movie on the left by a synthesised 3D MRI image on the right. As a result, we generate a high-resolution 3D real-time movie from the low-resolution 2D real-time MRI movie. The approach requires for each speaker a (low-resolution) real-time MRI movie and a limited number of high-resolution (possibly 3D) MRI images recorded for articulations representative of the speaker's articulatory repertoire.

Technically, we synthesise a high-resolution 3D movie based on a generative model that 1) assigns the small number of static 3D volumes to their best matching frames from the low-resolution movie and that 2) encodes the low-resolution movie through linear combinations of these frames. The same combinations can then be applied to the matching static 3D volumes, thereby creating the synthesised movie.

2. Method

2.1. Notation and problem statement

Given 2D and 3D MRI data from the same speaker:

- The $(m_l \times n_l)$ matrix L: 2D MRI movie (real-time; low spatial resolution) with m_l pixels and n_l frames. Free speech. The columns of L contain 2D frames flattened into vectors.
- The $(m_h \times n_h)$ matrix H: 3D MRI volumes (high spatial resolution) with m_h voxels and n_h volumes. The volumes, thereafter *basis volumes*, are isolated articulations representative of the articulatory repertoire of the speaker (volumes flattened into vectors).

Synthesise a high-resolution 3D MRI movie matrix M of dimensionality $(m_h \times n_l)$ by combining the information from L and H.

10.21437/Interspeech.2023-804

2.2. Method overview

The movie L shows one articulation or state of the vocal tract per frame. L contains redundancy in the sense that many articulations are transitions between target articulations. It is known that a well-chosen limited set of articulations is sufficient to reconstruct any articulation of a speaker [7]. Hence, we assume (1) that L can be encoded through linear combinations of a limited set of its columns (frames) and (2) that H also contains such a set of representative articulations.

Formally, we assume that L is low-rank and can be approximated by linear combinations of a small number of its columns identified by the index set $\kappa = [\kappa_1, \ldots, \kappa_c]$ of length $c \ll n_l$. Hence, if we select a sufficiently large index set κ, the norm error $||L - L(\,:, \kappa)X||$ is very small, such that we obtain a good approximation: $L \approx L(\,:, \kappa)X$. Here, $L(\,:, \kappa)$ denotes the subset of frames selected from L, and X is a $(c \times n_l)$ coefficient matrix that describes the contribution of the c selected frames to each of the $1, \ldots, n_l$ frames of L.

The coefficients in X model each frame of L as a combination of the selected frames. If we assume that $X \in \mathbb{R}^{\geq 0}$, only additive mixtures of the frames in $L(\,:, \kappa)$ are allowed. This enforces more realistic combinations: A frame can then not be modelled through subtracting a frame, but only positively as, for example, $0.7\,L(\,:, \kappa_1) + 0\,L(\,:, \kappa_2),\ \ldots,\ + 0.3\,L(\,:, \kappa_c)$, indicating that it contains mostly the articulation in frame κ_1, but with a contribution of κ_c, modelling a transition.

If we further assume that, for a sufficiently long speech task, there is at least one frame from L that corresponds to one of the basis volumes from H, we can encode L through linear combinations of the frames matched to the basis volumes in H. The same coefficients that determine the encoding of a frame $L(:, i)$ can then be applied to combine the columns of H, synthesising the volume frame $M(:, i)$ of the high-resolution movie.

In the following, we describe the steps of our method: First, we first assign the basis volumes (in practice, we use midsagittal images to facilitate the matching) from H to their best-matching frames from L, determining the index set κ (Assignment: Section 2.3). In a second step, we then find the coefficient matrix X to reconstruct $L \approx L(\,:, \kappa)X$ from the selected frames (Encoding: Section 2.4). Finally, the coefficients in X are employed to combine the articulations in H, synthesising the high-resolution movie M (Synthesis: Section 2.5).

2.3. Assigning the basis volumes from H to frames of L

2.3.1. Preprocessing

The 2D frames and 3D volumes are from the same speaker, but have been recorded with different acquisition parameters. Thus, they are not in the same coordinate space and have a different image appearance. Preprocessing is required to correct for this:

- Create 2D midsagittal projections of the 3D volumes in H that match the midsagittal view in the 2D frames in L (done in [8]). We denote the matrix containing the 2D basis images by H^{2D}.
- Correct for shift and rotation differences: (1) Register the mean frame of L to the mean basis image of H^{2D} by affine registration (performed using the Matlab function imregtform, metric: MattesMutualInformation, optimizer: OnePlusEvolutionary); (2) Apply the registration parameters to all frames of L.
- Create a binary vocal tract mask $\mathbf{v}$ based on the pixels that exhibit changes over time: These are the parts of the image where the articulators are located. In particular, compute the median absolute deviation (MAD) along the time dimension of L, followed by spatial low-pass filtering of the MAD image (convolution with a large Gaussian kernel with $\sigma = 15$) and thresholding to obtain a binary mask.
- Filter all images from L and H^{2D} by a spatial low-pass filter (convolution with a Gaussian kernel with $\sigma = 2$) to reduce noise, and then multiply with the vocal tract mask $\mathbf{v}$.
- The mean frame of the movie accounts for the largest part of the appearance difference between the low- and the high-resolution versions. Subtract the mean frame, i.e. $L - \overline{L}$ and $H^{\text{2D}} - \overline{H^{\text{2D}}}$, to reduce the appearance difference and to focus on the moving parts that deviate from the mean.

2.3.2. Matching process

For the ith basis image $H^{\text{2D}}(\,:, i)$, we find the frame $L(\,:, j)$ that minimises a dissimilarity function $d(H(\,:, i),\, L(\,:, j))$. It is defined as 1 - the Jaccard (also known as Tanimoto) similarity between the preprocessed images $\mathbf{h} \in H^{\text{2D}}$ and $\mathbf{l} \in L$:

$$d(\mathbf{h}, \mathbf{l}) = 1 - \frac{\mathbf{h} \cdot \mathbf{l}}{||\mathbf{h}||^2 + ||\mathbf{l}||^2 - \mathbf{h} \cdot \mathbf{l}}$$

We thereby obtain an assignment of the $i = 1, \ldots, n_h$ basis images in H^{2D} to n_h distinct frames of L identified by the index set κ, and referred to as *basis frames*. Technically, we employ the dissimilarity function $d()$ to compute the entries of the $(n_h \times n_l)$ cost matrix D and then solve an optimal assignment problem with the Hungarian algorithm [9, 10].

2.4. Encoding the low-resolution movie L

We compute the coefficient matrix $X \in \mathbb{R}^{\geq 0}$ by solving a non-negative least squares problem (NNLSQ). With a NNLSQ solver (Matlab function lsqnonneg; Lawson-Hanson algorithm [11]), we can find the $X \in \mathbb{R}^{\geq 0}$ that minimises $||L - L(\,:, \kappa)X||$ given the movie L and the basis frames $L(\,:, \kappa)$.

If many basis frames contribute with non-zero coefficients to the reconstruction of each frame, we may obtain a better approximation to L, but potentially at the cost of blurriness resulting from averaging many images. The non-negative least squares approach (NNLSQ) computes a matrix X that is both good at reconstructing L and sparse (containing few non-zero coefficients), such that each frame is modelled based on only few most relevant basis images.

If we assume an ideal, hypothetical case in the absence of noise where $L(\,:, \kappa)$ contains in fact the exact generators of L, i.e. $L = L(\,:, \kappa)X$, $X \in \mathbb{R}^{\geq 0}$, meaning that each frame is an additive mixture of one or more of the generators, then NNLSQ yields the optimal X, and L is recovered exactly from only c of its frames.

2.5. Synthesising the high-resolution movie M

The $(n_h \times n_l)$ coefficient matrix X models each frame of L as a combination of the n_h basis frames in L, and these basis frames are matched to their basis images counterparts in H^{2D} $(m_h \times n_h)$. Hence, we can use X to compute the $(m_h \times n_l)$ high-resolution 3D movie matrix M by matrix multiplication, $M := HX$ (or the 2D variant $M^{\text{2D}} := H^{\text{2D}}X$), modelling each frame of the synthesised movie as a linear combination of the n_h basis volumes from H with coefficients optimised on the movie L.

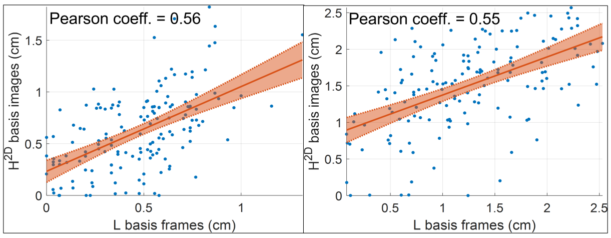

Figure 2: *Assignment: Values for the matched H^{2D} vs. L frames for the lip aperture (left) and the velum-tongue distance (right) for the 5 manually annotated speakers (dots); linear regression line (solid line) and 95% confidence interval (filled area).*

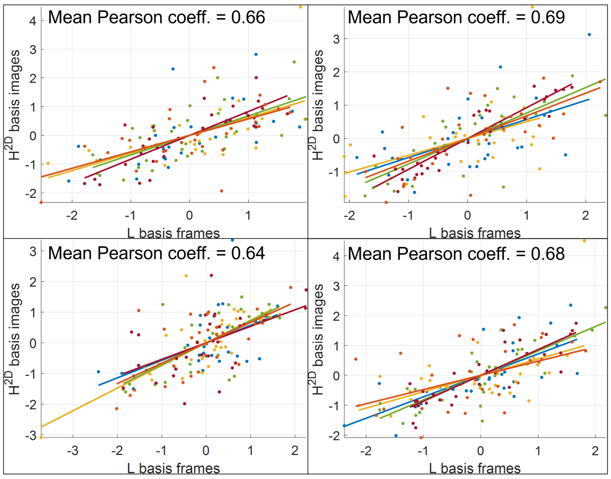

Figure 3: *Assignment: Values for the matched H^{2D} vs. L frames for the 4 tongue parameters (top-left: front-back, top-right: up-down, bottom-left: tip up-down, bottom-right: back position) and linear regression line (solid lines) per speaker for the 5 manually annotated speakers.*

3. Results

3.1. Data

For evaluation, we relied on the publicly available USC Speech and Vocal Tract Morphology MRI Database [6] with 17 American English native speakers. We used a subset consisting of: (1) 2D midsagittal real-time MRI recordings of the "north wind passage", about 1000 frames of spatial resolution 2.9 mm^2/pixel recorded at 23.18 frames/s, representing the movie L, and (2) 3D MRI recordings representative of the speakers' articulatory repertoire, $n_h \approx 30$ articulations sustained for 7 s and of spatial resolution 1.5625 mm^3/pixel, representing the basis volumes H. The manually outlined vocal tract contours of the H^{2D} sets are available from a previous study [8].

3.2. Results

There are two possible sources of errors:

1. The assignment of the basis volumes H to their best-matching frames in the low-resolution L may be wrong. Then, the movies M and M^{2D} are combined from wrong basis volumes or images.
2. The encoding of the low-resolution movie matrix L in terms of the selected frames κ may not be accurate enough, i.e. the norm error $||L - L(:,\kappa)X||$ is too high. If the coefficients in X do not reconstruct L well, they will also not be suitable for synthesising the new movie M.

3.2.1. Assignment

We evaluated the assignment between the basis images H^{2D} and the basis frames of L based on features describing shape and position of the articulators. The features were derived from the existing contours for H^{2D} and from manual annotations on 5 speakers for L.

We considered three major articulators, the lips, the velum and the tongue, measuring the aperture of the lips and the distance of the middle point of the inferior face of the velum to the tongue. For the tongue, principal components analyses were applied on four different subsets of the contours to capture the articulatory components describing (1) the frontward-backward position and shape, (2) the upward-downward position and shape, (3) the upward-downward position and shape of the tongue tip and (4) the position and shape of the back of the tongue. This approach is inspired from known shape components of the tongue [7]. The principal component scores for the basis frames and images characterise the tongue according to these components.

We plotted the feature values obtained for each basis image in H^{2D} against the values for their matching frames from L for the lip aperture and the tongue-velum distance (Figure 2), and for the tongue parameters (Figure 3). The position and shape of the articulators were consistent between the H^{2D} basis images and the L basis frames, but with a relatively high level of noise.

3.2.2. Encoding

To analyse how well the the low-resolution movie L can be reconstructed by combination of the frames κ, we computed the norm reconstruction accuracy (in percent of the norm of L):

$$\text{reconstruction accuracy} := 100 - \left(\frac{||L - L(:,\kappa)X||^2}{||L||^2} * 100\right)$$

Figure 4 shows the frame-by-frame reconstruction accuracy along the time axis of the movie L for a single speaker. The accuracy was generally high with a mean of 96.6% (dotted line). We subsequently calculated the distance between the original movie and a random time permutation of the reconstructed

speaker	1	2	3	4	5	6	7	8	9
rec. accuracy	96.6	96.42	97.41	96.93	97.39	96.49	97.82	96.34	97.97
rec. accuracy (permuted)	91.26 (0.09)	89.94 (0.09)	93.26 (0.07)	91.65 (0.08)	91.53 (0.1)	89.34 (0.1)	94.59 (0.04)	89.78 (0.08)	94.06 (0.05)
speaker	10	11	12	13	14	15	16	17	
rec. accuracy	97.14	97.09	96.85	96.94	96.87	96.62	96.6	97.21	
rec. accuracy (permuted)	93.08 (0.05)	93.23 (0.06)	91.35 (0.1)	90.87 (0.08)	90.61 (0.08)	91.53 (0.08)	89.42 (0.09)	92.42 (0.08)	

Table 1: *Reconstruction accuracies (in percent of $||L||$). Top: original low-resolution movie. Bottom: Mean (standard deviation) over 100 random permutations of the time dimension.*

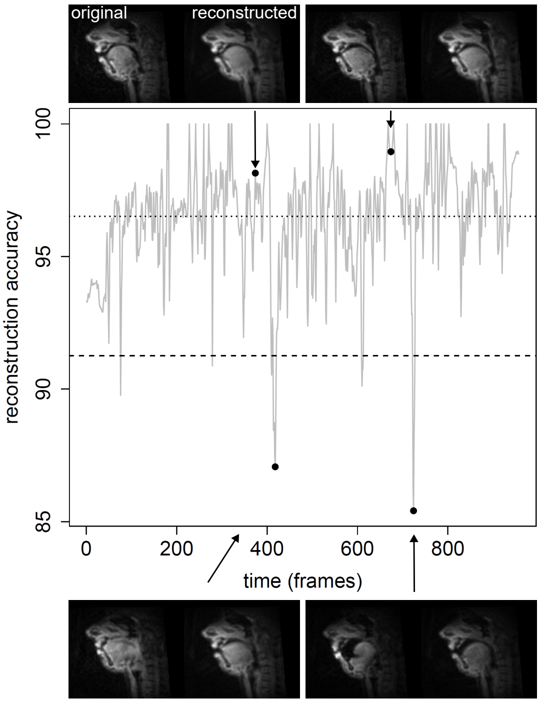

Figure 4: *Encoding: Reconstruction accuracy by frame (speaker 1). Dotted line: mean reconstruction accuracy (matching frames). Dashed line: mean reconstruction accuracy over 100 random permutations (non-matching frames). Image pairs visualise reconstruction at four time points (left: original, right: reconstructed). Top: two reconstructions with high accuracy. Bottom: two reconstructions with accuracy lower than baseline (dashed line).*

movie to serve as baseline. The mean reconstruction accuracy over 100 of such permutations (dashed line) was 91.26%. The vast majority of the reconstructed frames was clearly above this baseline (Figure 4). For all speakers, the frames κ that were selected by matching the basis images in $H^{2\mathrm{D}}$ could reconstruct the low-resolution movie L well with a mean reconstruction accuracy of 96.98%. (Table 1). The frame reconstructions were also specific as established by the permutation analysis: Across all speakers, the reconstruction accuracy was higher for the original L than for the permuted variants of L (Table 1).

3.2.3. Accuracy of the synthesised movie

The synthesised movie M was evaluated by calculating the similarity between the low-resolution movie L and the synthesised high-resolution movie $M^{2\mathrm{D}}$ and, as a baseline, the similarity between L and 100 random time permutations of $M^{2\mathrm{D}}$: Overall, the mean Jaccard similarity was higher between matching than between non-matching frames resulting from random permutations (Figure 5a). Hence, on average the synthesised frames in H were specific for their matching frames from L.

Figure 5b shows examples for the synthesised frames. The supplementary material[1] contains examples for basis images matched between $H^{2\mathrm{D}}$ and L and synthesised high-resolution 3D movies for two speakers.

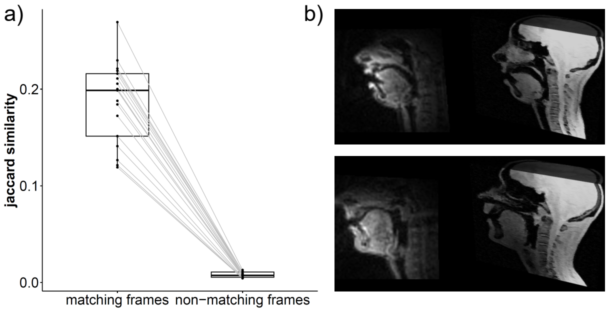

Figure 5: *Synthesised movie:* ***a)*** *Mean Jaccard similarity of matching frames (and non-matching frames as controls) from the low-resolution movie L and the synthesised high-resolution movie M^{2D} (after subtracting the mean frame and vocal tract masking). Boxplots summarise the mean Jaccard similarities for all* 17 *speakers from the USC data.* ***b)*** *Qualitative examples. Left: original 2D real-time movie, Right: synthesised 3D real-time movie (midsaggital view).*

4. Discussion

The generative model for synthesising a 3D real-time movie is interpretable in the sense that it is transparent how each frame is computed. The accuracy of both the assignment and the encoding stage can be measured, and if both are sufficiently accurate we can be confident in the accuracy of the synthesised movie.

We found that the real-time movies L could be encoded with high reconstruction accuracies $\geq 96\%$. For this encoding to be useful for synthesising the 3D movie, the frames need to be assigned to the correct basis images in H^{2D}. Our results show that parameters describing position and shape of the articulators were indeed correlated between the matched images, although not perfectly.

The present approach is limited by the across-modality assignment of basis images in H^{2D} and L. While we employ preprocessing and an automatically generated vocal tract mask to focus on the relevant parts, the assignment is nevertheless performed on a pixel level. Future versions could implement the assignment based on automatic segmentations of the vocal tract [12], such that the articulators shape and position can be matched and evaluated explicitly. Another limitation is that a basis image in H^{2D} might not have a good match in the real-time movie L. This could be addressed by using all available movies from the same speaker as a reservoir or by attempting to synthesise a basis image in the low-resolution domain.

This paper presents a proof of the concept, showing how real-time MRI movies and static 3D MRI volumes can be combined to synthesise 3D real-time movies. Future versions will address improvements of the assignment step, utilizing for example deep learning techniques for segmentation or image synthesis, or could rely on data where a limited set of corresponding basis images covering the speaker's articulatory repertoire has been recorded in both 2D and in 3D MRI.

Integrating high temporal and high spatial resolution MRI from the same speaker is a generic way of improving data quality in vocal tract measurements, and it can be applied on top of technical advances in the MRI acquisition process.

[1] https://github.com/tonioser/supplementary-material-Interspeech2023-paper804

5. References

[1] T. Baer, J. Gore, L. Gracco, and P. Nye, “Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels,” *The Journal of the Acoustical Society of America*, vol. 90, pp. 799–828, 1991.

[2] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, “An approach to real-time magnetic resonance imaging for speech production,” *The Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1771–1776, Apr 2004. [Online]. Available: https://asa.scitation.org/doi/10.1121/1.1652588

[3] S. G. Lingala, B. P. Sutton, M. E. Miquel, and K. S. Nayak, “Recommendations for real-time speech MRI,” *Journal of Magnetic Resonance Imaging*, vol. 43, no. 1, pp. 28–44, Jan 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/jmri.24997

[4] Y. Lim, Y. Zhu, S. G. Lingala, D. Byrd, S. Narayanan, and K. S. Nayak, “3D dynamic MRI of the vocal tract during natural speech,” *Magnetic Resonance in Medicine*, vol. 81, no. 3, pp. 1511–1520, nov 2018.

[5] I. K. Douros, A. Tsukanova, K. Isaieva, P.-A. Vuissoz, and Y. Laprie, “Towards a Method of Dynamic Vocal Tract Shapes Generation by Combining Static 3D and Dynamic 2D MRI Speech Data,” in *Proc. Interspeech 2019*, 2019, pp. 879–883.

[6] T. Sorensen, Z. Skordilis, A. Toutios, Y.-C. Kim, Y. Zhu, J. Kim, A. Lammert, V. Ramanarayanan, L. Goldstein, D. Byrd, K. Nayak, and S. Narayanan, “Database of volumetric and real-time vocal tract MRI for speech science,” in *Proceedings of Interspeech 2017*, Stockholm, Sweden, 2017, pp. 645–649. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-608

[7] 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,” *The Journal of the Acoustical Society of America*, vol. 109, pp. 2165–2180, 2001.

[8] A. Serrurier and C. Neuschaefer-Rube, “Morphological and acoustic modeling of the vocal tract,” *The Journal of the Acoustical Society of America*, vol. 153, no. 3, pp. 1867–1886, mar 2023.

[9] H. W. Kuhn, “The Hungarian method for the assignment problem.” *Nav. Res. Logist.*, vol. Q2, p. 83–97, 1955.

[10] Y. Cao, “Hungarian algorithm for linear assignment problems (v2.3),” *MATLAB Central File Exchange*, 2020. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/20652-hungarian-algorithmfor-linear-assignment-problems-v2-3

[11] C. L. Lawson and R. J. Hanson, *Solving Least-Squares Problems*. Prentice Hall, 1974, ch. chapter 23, p. p. 161.

[12] 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, jun 2022.