# CAN DEEP LEARNING HELP TO UNDERSTAND SPEECH PRODUCTION MECHANISMS?

*Antoine Serrurier*[1]

[1]*Clinic for Phoniatrics, Pedaudiology & Communication Disorders, University Hospital and Medical Faculty of the RWTH Aachen University, Germany*
*aserrurier@ukaachen.de*

**Abstract:** Deep Learning has become inescapable in all fields of research. It leads to unprecedented levels of prediction but is often associated with a loss in understanding the considered phenomenon. This study aims on the contrary at taking advantage of the performance of Deep Learning to increase knowledge in speech production. The study explores more specifically the potential of Deep Learning as an original method to help at determining the cross-speaker vowel-specific articulatory invariants, *i.e.* the stable articulatory features in the production of vowels. 228 midsagittal MRI data of 41 speakers articulating 6 vowels have been considered, for which manually traced vocal tract contours are available and aligned in a common reference coordinate system. Convolutional Neural Networks have been trained to classify the images in terms of vowel for five increasingly challenging classification scenarios, from two to six classes, in a leave-one-speaker-out scheme, with accuracies above 99%. The Grad-CAM algorithm has been applied to all test images, resulting in heatmaps identifying the determinant vocal tract regions for a robust classification of the image. The edges of this region for each image have been aligned in the reference coordinate and averaged over all instances of a vowel for a scenario. The preliminary results show that a vowel can be robustly identified from the anterior part of the vocal tract, even if the constriction, crucial for the acoustics, is located in the posterior part. Our approach demonstrates the potential of Deep Learning as a tool to increase knowledge in speech production.

## 1 Introduction

In the last years, Deep Learning (DL) has become inescapable in all field of research and speech sciences is directly concerned [1]. It has led to unprecedented prediction performance and led to exceptional breakthroughs. It is however often associated with a loss in interpretability of the considered phenomenon. This study aims on the contrary at taking advantage of the performance of DL to increase knowledge in speech production.

In speech production, a core aspect is the setting of the position and shape of the vocal tract articulators to achieve the desired acoustic-articulatory targets. Vowels are usually achieved by specific articulatory settings, driven by acoustic targets such as the formants [2]. Complementary, a large inter-speaker articulatory variability is observed [3]. Determining the cross-speaker vowel-specific articulatory invariants, *i.e.* the stable articulatory features characteristics for the production of the vowel, remains therefore challenging. Standard methods rely on normalisation procedures (see [3] for such a method and a short literature review). Determining the articulatory invariants has been taken as the case study to explore the potential of DL to contribute at solving this issue.

The chosen approach consists in training a DL network to classify input Magnetic Resonance Imaging (MRI) data of the vocal tract for different vowel classes. The network learns to identify on new image the characteristic features to perform the correct classification. These features represent the determinant vocal tract regions for the robust identification of the vow-

els. An analysis of these determinant regions is then proposed. A similar method has already been attempted [4], but in the framework of articulatory-to-acoustic mapping. The authors showed that the trained network coefficients relate to the articulatory vowel space. In addition, they also exhibited the determinant regions for their classification task. The current study aims at deepening this approach by introducing newer and more performant networks and a state-of-the-art network analysis algorithm, and by focusing on the vocal tract regions as potential markers of the articulatory invariants.

## 2 Material and methods

### 2.1 Data

The data consist of midsagittal MRI images of the vocal tract recorded from 41 speakers sustaining for a few seconds vowels in their native language. The data constitute a subset of the data presented with more details in [5]. In this study, only a subset of the oral vowels was retained, leading altogether to 228 images from 41 speakers. A summary of the data is presented in Table 1. Because of the deep learning approach, the largest possible amount of data is necessary for the training, and in particular the maximum number of speakers for each vowel. For that purpose, the same vowels of different languages are grouped into the same classes, leading to the six cross-language vowel classes /i, a, u, o, ø, ɛ/. For the dataset #5, the vowel /æ/ has been associated with the class /a/ and the vowel /ɑ:/ with the class /o/; for this dataset, there is no image for the vowel class /ø/.

For each image, manually outlined contours of the vocal tract and of parts of the face are also available. All articulation contours for a single speaker are aligned on a common reference coordinate system in centimetres based on the upper teeth and the hard palate. The articulation contours of all speakers are aligned together on common bony reference points of the cranium. Further details of this process are provided in [6]. An example of an image and its corresponding contours is visible in Figure 1.

### 2.2 Method

The approach consists in classifying the images by means of DL and in analysing the trained network to determine the strategy used by the network to take the classification decision. As a classification task can be performed very differently according to the number of classes, five increasingly challenging classification scenarios ranging from two to six classes have been considered (see Table 2). It is aimed to simulate the different strategies that could be implemented to classify images from languages with two to six vowels.

**Table 1** – Overview of the datasets

| | Native language | Male/Female | Corpus | Reference |
|---|---|---|---|---|
| #1 | French | 6/5 | [i a u o ø ɛ] | [6] |
| #2 | French | 1/0 | [i a u o ø ɛ] | [7] |
| #3 | German | 7/3 | [i: a: u: o: ø: ɛ:] | [5] |
| #4 | German | 1/1 | [i: a: u: o: ø: ɛ:] | [8] |
| #5 | English | 8/9 | [i: æ u: ɑ: ɛ] | [9] |

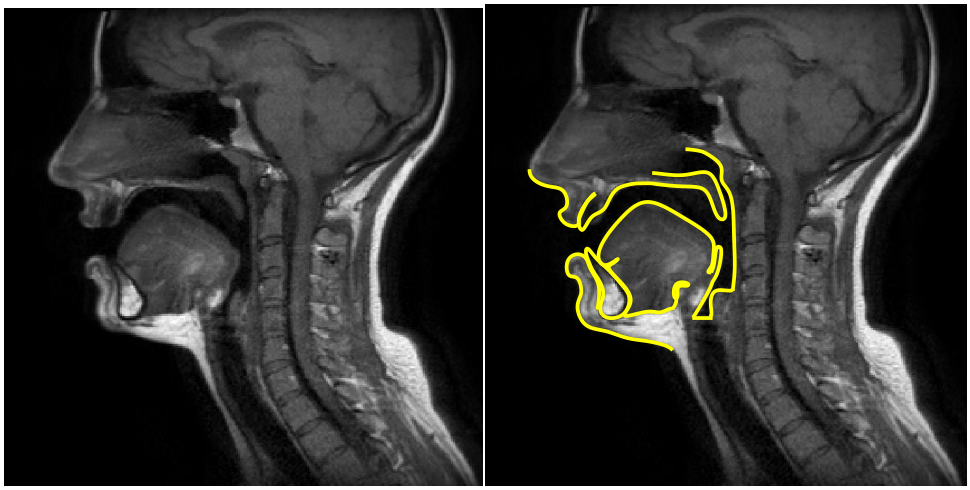

**Figure 1** - MRI image taken from the data (left) and same image superimposed with the manually segmented contours (right)

**Table 2** – Classification scenarios

| Scenario | Classes | Number of images per class |
|---|---|---|
| #1 | /ɛ ɔ/ | 41 – 40 |
| #2 | /a i u/ | 41 – 41 – 41 |
| #3 | /a i u ø/ | 41 – 41 – 41 – 24 |
| #4 | /a i u ɛ ɔ/ | 41 – 41 – 41 – 41 – 40 |
| #5 | /a i u ø ɛ ɔ/ | 41 – 41 – 41 – 41 – 40 – 24 |

For each scenario, the classification has been performed and evaluated in a leave-one-speaker-out scheme. In this scheme, all the images of one speaker, *i.e.* between two and six images depending on the scenario, have been left out to serve as test images. The remaining images have been randomly split in 80% training images and 20% validation images. The training dataset has been augmented by creating for each vowel 100 artificial MRI images as follows: (1) 100 artificial articulation contours have been randomly generated as random linear combinations of the existing articulation contours of the training data, (2) for each artificial contour, the closet existing articulation contour of the training data has been identified and (3) the corresponding image has been warped using the existing articulation contour points as source landmarks and the artificial contour points as target landmarks. An example is visible in Figure 2.

A pre-trained *EfficientNet B0* Convolutional Neural Network (CNN) [10] has been loaded and trained with the augmented training data. The trained network has then been evaluated on the test images. This procedure has been repeated until all speakers have served exactly once as test speaker. Proceeding this way ensures that the trained networks are evaluated on speakers that have not been used for the training. The drawback is that it requires to train as many networks as speakers for each scenario and that each test speaker is associated with a different network.

For each scenario, the overall classification accuracy is provided. Although the objective in this study is not to achieve the best possible accuracy, bad classification results would imply that the networks are not able to perform robust classification and would raise doubts on the results of the study.

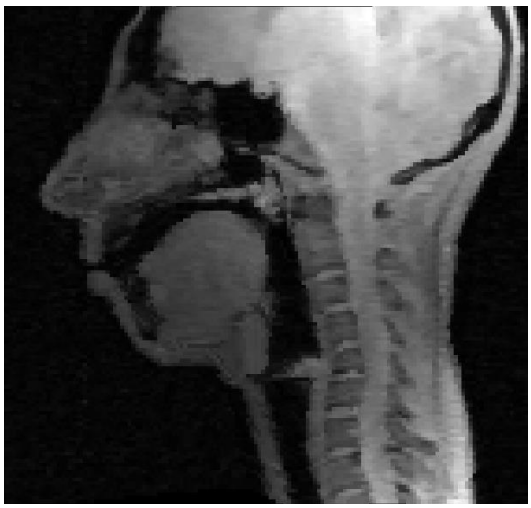

**Figure 2** – Augmented image for the vowel /u/

For all correctly classified images, the Grad-CAM algorithm [11] has been applied on all convolutional layers of their corresponding network. This algorithm calculates the gradient of the output of a network layer. It outputs a matrix of size similar to the size of the layer image and is automatically resized to the size of the input image according to the rules used by the network. The result is a heatmap image which can be overlaid on the input image to highlight the regions playing an important role in the current layer. In a CNN, the image resolution decreases with the layers of the network to concentrate on meaningful features to perform the task. For this reason, the Grad-CAM outputs calculated on the first layers tend to have higher resolution but lower significance while outputs calculated on the last layers tend to have lower resolution but higher significance. For the last layer, the Grad-CAM output highlights the image regions on which the final classification score is calculated. Taking the average of the Grad-CAM outputs of all layers of a network highlights the regions playing a recurring important role towards the final classification result. In other words, the last layer Grad-CAM output may inform about the regions on which the classification decision is taken while the average Grad-CAM output may inform about the reason why the last layer region emerged (see Figure 3 for an example).

From each last layer Grad-CAM output, a surface equivalent to 12 square centimetres encompassing the most significant region for the classification decision has been extracted. The results have then been averaged together with the associated articulation contours over all instances of each vowel of each scenario. It provides for each vowel of each scenario the most significant region for the classification decision.

## 3 Results

The classification accuracy for each scenario is visible in Table 3. An example of Grad-CAM outputs is presented in Figure 3. The results for the surface averaged over all instances of each vowel of each scenario have been presented for the two extreme scenarios, *i.e.* two (Figure 4) and six (Figure 6) classes classification, as well as for the /a i u/ scenario containing the quantal vowels (Figure 5). In order to visualise whether the most significant region can be different according to the scenario, the Figure 7 displays for the vowels /a i u/, present in four of the five scenarios, the edges of the same surfaces for the four scenarios where they are present.

**Table 3** – Classification accuracy for each scenario

| #1 | #2 | #3 | #4 | #5 |
|---|---|---|---|---|
| 100% | 100% | 99.3% | 99.5% | 98.7% |

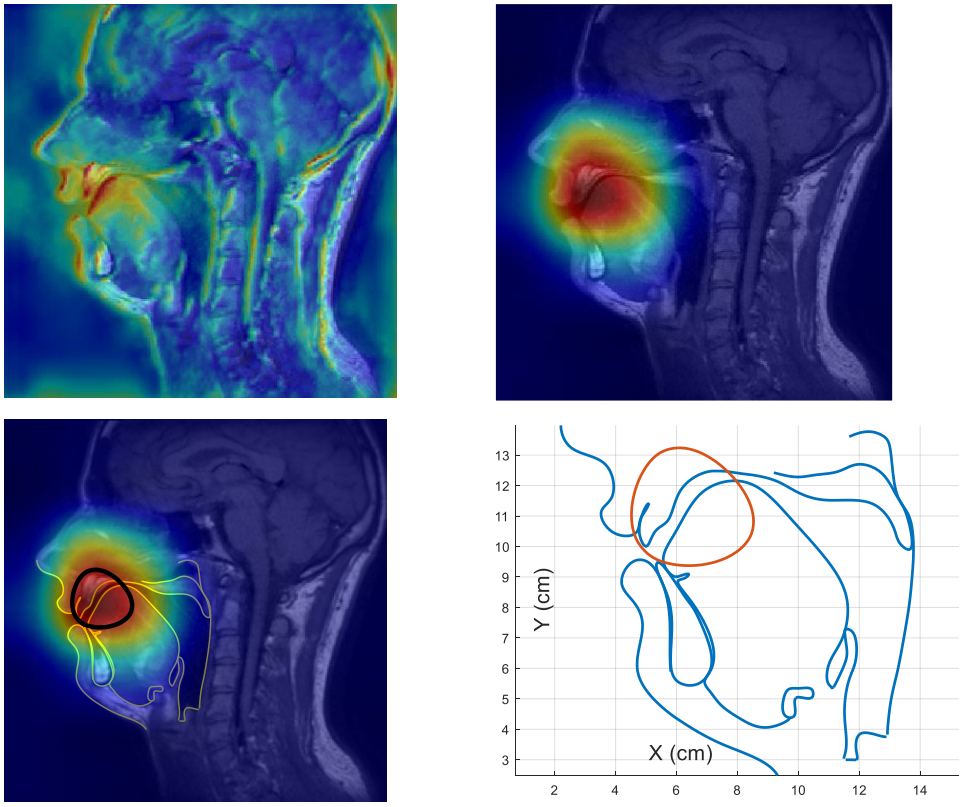


**Figure 3** – MRI image of [i] superposed with the output of the Grad-CAM algorithm for the scenario /a i u/ averaged over all network layers (top left), for the last layer only (top right) and for the last layer only superimposed with the articulation contours (yellow) and the edges of the surface of 12 cm² around the most significant region (black) (bottom left). Same articulation contours and surface edges in the common reference coordinate system in cm (bottom right).

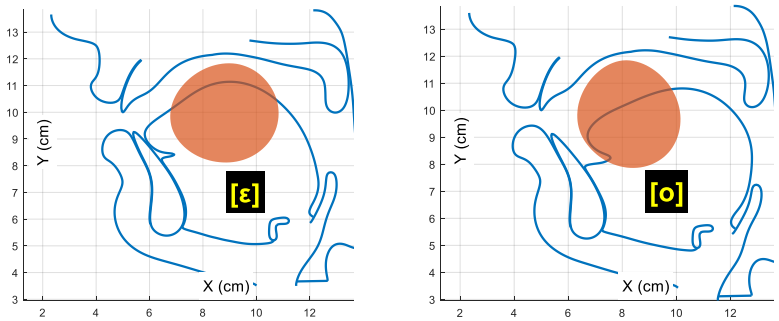


**Figure 4** - Averaged contours of the vowels superimposed with the averaged most significant regions for the classification decision for the scenario /ɛ o/.

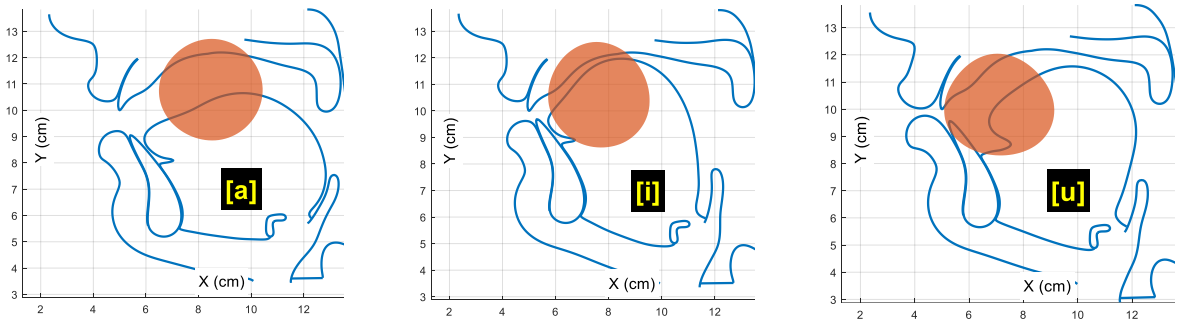


**Figure 5** - Averaged contours of the vowels superimposed with the averaged most significant regions for the classification decision for the scenario /a i u/.

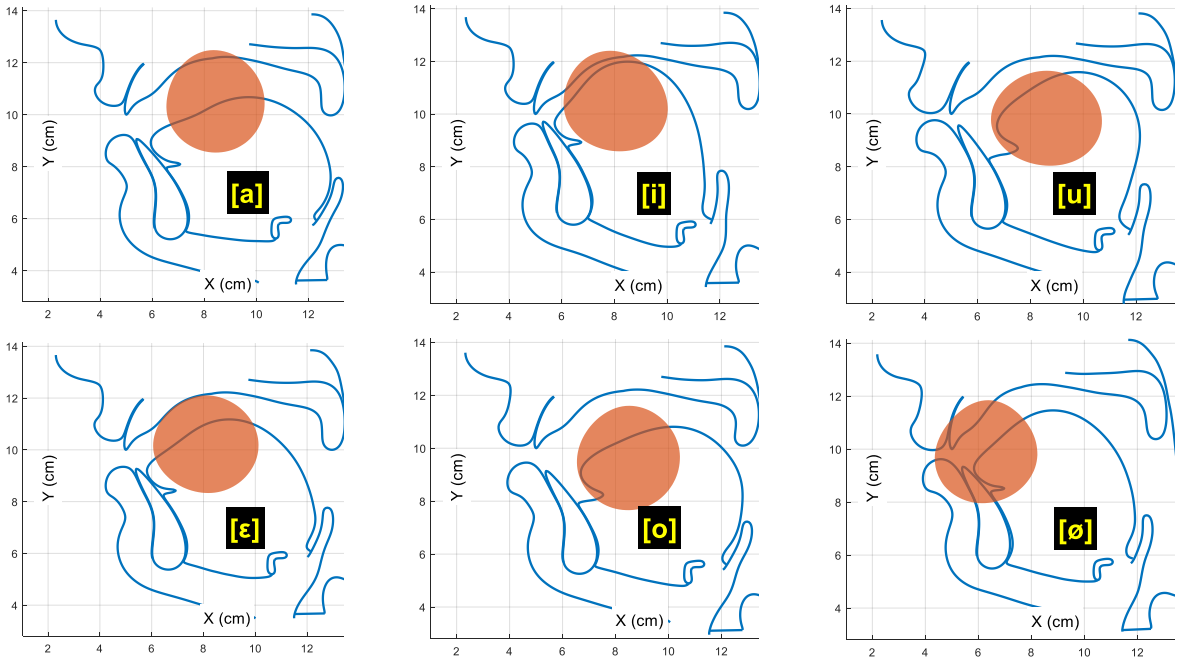


**Figure 6** - Averaged contours of the vowels superimposed with the averaged most significant regions for the classification decision for the scenario /a i u ø ɛ o/.

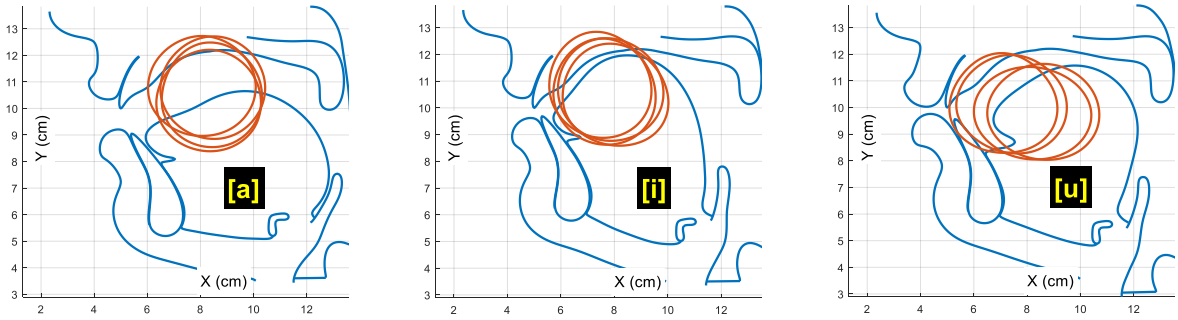


**Figure 7** - Averaged contours of the vowels superimposed with the edges of the averaged most significant regions for the classification decision for the scenarios #2 to #5.

## 4 Discussion and conclusion

We can observe the high level of performance of the networks despite the relatively limited training data for a deep network. This ensures that the features extracted from the images by the networks are robust features to take the classification decision.

Consistently for all vowels of all scenarios, despite some outliers, the classification decision is taken from the anterior part of the vocal tract, with small variations depending on the vowel and the scenario. It can also be observed that the classification scenario, more or less challenging, has only little influence on the region on which is taken the decision. We can see in Figure 7 more variability for /u/ than for /a i/, but it remains in the anterior part of the vocal tract. This suggests that the regions are in itself robust for a classification decision and depends only little on the list of other vowels it has to be discriminated from.

It is known that the vocal tract constriction is crucial for the vowels to achieve the desired acoustic targets and tends therefore to show less inter-speaker variability [3], forming a good candidate marker to identify a vowel. While the constriction seems indeed to emerge as the decisive region for the classification of /i/, it is not the case for the back vowels /a u o/. Rather, it seems that the realisation of a constriction leads to complementary open regions somewhere else in the vocal tract (buccal cavity for /a/, front cavity for /u/) from which a robust identification of the vowels appears possible. Further analyses are necessary to confirm and quantify these preliminary results.

The Figure 3 helps to understand the process of the network leading to the final classification decision: for the represented instance of the vowel /i/, it identifies the region containing the edges of the palate and tongue blade and takes the decision from this region. Further analyses are required to take deeper advantage of this representation in the interpretation of the most significant region for the classification decision.

Due to the relatively limited data for DL training, a leave-one-speaker-out rotating scheme was chosen. This led to the generation of a different network for the evaluation of each speaker of each scenario. For consistency, it would be better to have a single network per scenario. This calls for more data, although this might partly be solved in the future by data augmentation.

The initial motivation was to test whether the classification process would vary according to the number of vowels in the vowel system of a language, assuming that languages with a large number of vowels would require a decision taken on decisive regions while languages with a limited number of vowels could allow more flexibility. This led to the construction of the five scenarios of the study. However, due to the limited number of data, the vowels of languages with different number of vowels in their vowel system were mixed, limiting the possible conclusions on that matter: first the realisation of vowels of different language might slightly differ and second the production of a [a] in a three-vowel system language might allow much more articulatory flexibility than the production of a [a] in a ten-vowel system language for instance. In other words, the realisation of a vowel depends on the number of vowels in the vowel system. Further analyses involve the design of language-specific scenarios.

The preliminary results show that in terms of articulation, a vowel can be robustly identified from the anterior part of the vocal tract, even if the constriction, crucial for the acoustics, is located in the posterior part and tends to show smaller inter-speaker variability. It further suggests that constrictions and open regions might be complementary associated, and that the realisation of a vowel articulation could also be driven by the obtention of specific open regions, having an impact on the motor planning. Finally, our methodology shows the potential of DL as a tool for further understanding speech production mechanisms. Planned further

analyses include quantitative characterisation and more realistic language-specific scenarios comprising all vowels and possibly consonants.

## Acknowledgements

The author is very grateful to all people involved in the design, acquisition or processing of the datasets: P. Badin, T. Sawallis, L. Lamalle, S. Romanzetti, B. Kröger, C. Busch, J.-A. Valdés Vargas, G. Ananthakrishnan, M. Eslami, A. Hülsdünker and F. Stepp. This research project is supported by the START-Program of the Faculty of Medicine, RWTH Aachen University. This work was also supported by the Brain Imaging Facility of the Interdisciplinary Center for Clinical Research (IZKF) Aachen within the Faculty of Medicine at RWTH Aachen University.

## References

[1] LeCun, Y., Bengio, Y., and Hinton, G.: *Deep learning.* In *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[2] Ladefoged, P., and Johnson, K.: *A Course in Phonetics*. Cengage Learning, 2011.

[3] Serrurier, A., Badin, P., Boë, L.-J., Lamalle, L., and Neuschaefer-Rube, C.: *Inter-speaker variability: speaker normalisation and quantitative estimation of articulatory invariants in speech production for French*. In *Proceedings of Interspeech 2017*, Stockholm, Sweden, 2017, pp. 2272–2276. doi: 10.21437/Interspeech.2017-1126.

[4] van Leeuwen, K. G. *et al.*: *CNN-Based Phoneme Classifier from Vocal Tract MRI Learns Embedding Consistent with Articulatory Topology*. In *Proc. Interspeech 2019*, 2019, pp. 909–913. doi: 10.21437/Interspeech.2019-1173.

[5] Serrurier, A. and Neuschaefer-Rube, C.: *Morphological and acoustic modelling of the vocal tract.* In *The Journal of the Acoustical Society of America*, In review.

[6] Serrurier, A., Badin, P., Lamalle, L., and Neuschaefer-Rube, C.: *Characterization of inter-speaker articulatory variability: a two-level multi-speaker modelling approach based on MRI data*. In *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. 2149–2170, Apr. 2019, doi: 10.1121/1.5096631.

[7] Badin, P., Sawallis, T. R., Crépel, S., and Lamalle, L.: *Comparison of articulatory strategies for a bilingual speaker: Preliminary data and models*. In *10th International Seminar on Speech Production, ISSP10*. pp. 17–20, 2014.

[8] Birkholz, P., Kürbis, S., Stone, S., Häsner, P., Blandin, R., and Fleischer, M., 'Printable 3D vocal tract shapes from MRI data and their acoustic and aerodynamic properties', *Scientific Data*, vol. 7, no. 1, Aug. 2020, doi: 10.1038/s41597-020-00597-w.

[9] Sorensen, T. *et al.*: *Database of volumetric and real-time vocal tract MRI for speech science*. In *Proceedings of Interspeech 2017*, Stockholm, Sweden, 2017, pp. 645–649. doi: 10.21437/Interspeech.2017-608.

[10] Tan, M. and Le, Q. V.: *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. 2019, doi: 10.48550/ARXIV.1905.11946.

[11] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D.: *Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization*. In *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 618–626. doi: 10.1109/iccv.2017.74.