INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

CNN-based phoneme classifier from vocal tract MRI learns embedding consistent with articulatory topology

K.G. van Leeuwen[1,2,3], *P. Bos*[1,2], *S. Trebeschi*[2], *M.J.A. van Alphen*[1], *L. Voskuilen*[1], *L.E. Smeele*[1], *F. van der Heijden*[1,3], *R.J.J.H. van Son*[1,4]

[1]Department of Head and Neck Oncology and Surgery, Netherlands Cancer Institute, Amsterdam
[2]Department of Radiology, Netherlands Cancer Institute, Amsterdam
[3]TechMed Center, University of Twente, Enschede
[4]Amsterdam Center for Language and Communication, University of Amsterdam, Amsterdam
The Netherlands

kickyvanleeuwen@gmail.com, r.v.son@nki.nl

Abstract

Recent advances in real-time magnetic resonance imaging (rtMRI) of the vocal tract provides opportunities for studying human speech. This modality together with acquired speech may enable the mapping of articulatory configurations to acoustic features. In this study, we take the first step by training a deep learning model to classify 27 different phonemes from midsagittal MR images of the vocal tract.

An American English database was used to train a convolutional neural network for classifying vowels (13 classes), consonants (14 classes) and all phonemes (27 classes) of 17 subjects. Classification top-1 accuracy of the test set for all phonemes was 57%. Error analysis showed voiced and unvoiced sounds often being confused. Moreover, we performed principal component analysis on the network's embedding and observed topological similarities between the network learned representation and the vowel diagram. Saliency maps gave insight into the anatomical regions most important for classification and show congruence with known regions of articulatory importance.

We demonstrate the feasibility for deep learning to distinguish between phonemes from MRI. Network analysis can be used to improve understanding of normal articulation and speech and, in the future, impaired speech. This study brings us a step closer to the articulatory-to-acoustic mapping from rtMRI.

Index Terms: speech analysis, magnetic resonance imaging (MRI), convolutional neural networks (CNN), articulatory-to-acoustic mapping, deep learning, vocal tract

1. Introduction

Within speech research, it has been a long-standing challenge to be able to estimate the acoustic features corresponding to a specific vocal tract configuration, also called articulatory-to-acoustic mapping. This is not a trivial problem since there is much variability between subjects. Also during speech production, the fast transitions of the articulators are difficult to capture with current measurement methods.

X-ray is one of the methods to extract articulatory information, however it has the disadvantages of bad soft tissue contrast and potentially hazardous radiation. Electropalatography can only measure when and where the tongue touches the palate and requires a customized electropalate. In electromagnetic articulography (EMA), sensor coils are placed on the tongue which can cause heterogeneity among speakers and interference with natural articulation. Advances in real-time MRI (rtMRI) make it possible to image the whole vocal tract and soft tissue articulators at a sufficient frame rate needed for speech analysis. The advantages of this technique are that no potentially hazardous radiation is needed, and nothing is placed in the mouth that could interfere with the articulators' movement. This technique is also suitable for patients with vocal tract pathology, e.g. patients who have had a (partial) tongue resection or experience pain in these areas. The advantages go at the expense of the increased complexity to acquire the data and patient-level disadvantages such as the high noise level inside an MRI scanner [1]–[4].

The University of Southern California gathered a dataset with rtMRI and MRI from sustained sounds (USC Speech and Vocal Tract Morphology MRI Database)[5]. In this study, we use deep learning and single frame MR images of the sustained phonemes to model the relation between the vocal tract configuration and the phoneme.

Saha et al. [6] have previously attempted to classify vowel-consonant-vowel (VCV) combinations from the USC speech database using rtMRI [5], [7]. Image features of multiple frames were extracted and combined with a long short-term memory network to form a general prediction of the VCV 'video'. They reached an accuracy of 42% for 51 different VCV combinations.

This study aims to predict the corresponding phoneme from a static MR image using a convolutional neural network. We trained a neural network for three tasks: classification of 13 vowels, classification of 14 consonants, and classification of all the 27 phonemes. Secondly, we perform an extensive analysis of what the neural network has learned to gain insights in the relation between vocal tract configurations, speech, and phonetics. These techniques can help us gain a better understanding in what makes pathological speech abnormal.

This study is a preliminary work to demonstrate the feasibility of using rtMRI and deep learning for articulatory-to-acoustic mapping. The larger goal we have is to individualize the approach for people with impaired speech. We want to predict the impact of interventions like surgery

http://dx.doi.org/10.21437/Interspeech.2019-1173

and radiotherapy on functional outcomes, such as speech and swallowing, for a more personalized treatment plan. A methodology that is able to perform an articulatory-to-acoustic mapping is essential in order to reach this goal.

2. Method

2.1. Data

A published database, the USC Speech and Vocal Tract Morphology MRI Database from the University of Southern California, was used for the classification tasks [5]. The database consists of 2D rtMRI data (1.5 Tesla) including the recorded speech of the vocal tract and 3D volumetric MRI (3 Tesla) while subjects utter sustained vowels and continuant consonants. The latter set was used for phoneme classification. Data of 17 subjects (8 male, 9 female) were present with 13 vowels (234 images) and 14 consonants (255 images). Further details on the different phonemes are given in Table 1. For some subjects, a phoneme was missing or a duplicate was present. Overall the dataset was balanced with a mean of 18 ± 0.8 samples per class. For the classification task, only the midsagittal slice was extracted and used. Each image was unity-based normalized and resampled to 32 by 32 pixels. Data were randomly split on subject level between a train (14 subjects) and a test set (3 subjects).

The Cifar10 dataset [8], existing of 60,000 color images from 10 different classes, was used to pretrain the network for improved image feature extraction. The three color channels were averaged to create grayscale images.

2.2. Architecture and training

The neural network architecture consists of four convolutional blocks. Each block is made up of two convolutional layers with ReLU activations ending with a max pooling layer, as shown in Figure 1. In the last block, the pooling layer is replaced by flattening to transform the feature maps to vectorial feature space. A softmax classifier is used for the prediction of the phoneme resulting in a probability score for each class.

Because of the limited data in the speech dataset used, the network was pretrained with the Cifar10 dataset to learn general image filters. The training was performed with a batch size of 10 images, early stopping and restoring the best performing model. For the speech classification tasks, the dense layer and softmax layer were replaced by newly initialized layers with the number of output nodes (k) corresponding to the number of classes to predict. All layers remained trainable and able to be fine-tuned to the MRI data.

Table 1: *Sustained phonemes (in bold) in USC Speech and Vocal Tract Morphology MRI Database.*

Sustained vowels	b**i**:t (b**ee**t), b**ɪ**t (b**i**t), b**ei**t (b**ai**t), b**ɛ**t (b**e**t), b**æ**t (b**a**t), p**ɑ**:t (p**o**t), b**ʌ**t (b**u**t), b**ɔ**:t (b**ough**t), b**oʊ**t (b**oa**t), b**u**:t (b**oo**t), p**ʊ**t (p**u**t), b**ɻ**d (b**ir**d), æb**ʌ**t (abb**o**t)
Sustained consonants	a**f**a, a**v**a, a**θ**a, a**ð**a, a**s**a, a**z**a, a**ʃ**a, a**ʒ**a, a**h**a, a**m**a, a**n**a, a**ŋ**a, a**l**a, a**ɹ**a

For both the Cifar10 and the speech dataset loss was defined by the categorical cross-entropy. The network was optimized with Adam optimizer [9].

Six-fold cross-validation was performed for grid hyperparameter tuning with the training set (training on 12 subjects, testing on 2). Hyperparameters were drop-out (0.1, 0.3, 0.5, 0.7), batch-size (1, 4, 8, 16), and amount of data augmentation (with varying amounts of zoom, rotation, shift and shear). Early-stopping with a latency of 30 epochs was applied. For the test phase, the hyperparameter set with the minimum average loss of all folds on all tasks (vowels, consonants, vowels+consonants) was chosen. Because of the varying number of classes, the number of epochs to train was differed per task and was set to the mean number of epochs of all folds times the factor of 1.3. As the size of training data increases during the test phase, this scaling factor ensures the convergence of the network.

The final performance was determined on each test subject, by retraining the network on all subjects except for the test subject (leave-one-out cross-validation). With this method, as close to all the data could be exploited for the training, while retaining a strict division between train and test samples. This process was repeated ten times to minimize the variance caused by random initialization.

2.3. Analysis

Top-1, top-3, and top-5 classification accuracy with standard deviations were computed for each task of the train and test set. The Welch's t-test was performed in order to compare train and test set performance. Probability confusion matrices were computed for the test set with all iterations combined. Principal component analysis (PCA) was performed on the embedded space (output of the flattened layer) to visualize the mapping learned by the network's image feature extractor. Saliency maps highlight regions in the input image contributing most towards the predicted class. They were created by computing the change in the prediction to a small change in the input image, resulting in a sensitivity heatmap over the input image [10], [11].

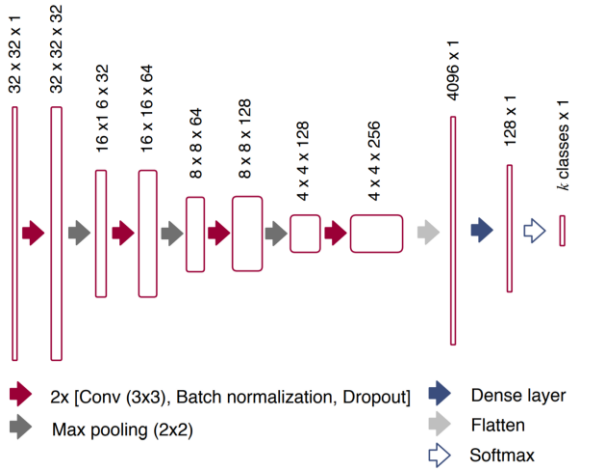

Figure 1: *Classification architecture existing of convolutional layers with kernel size 3 by 3 and max pooling of 2 by 2 for extracting meaningful image features, followed by a dense and softmax layer for phoneme classification with k classes. ReLU activations were present after each convolutional layer.*

Table 2: *Leave-one-out cross-validation accuracy of train and test set with batch-size of 4, drop-out of 0.3, and data augmentation with max zoom of factor 0.2, rotation of max 20 degrees, max shift of 0.2 and max shear of 0.2. Average of 10 iterations is given.*

	top-1 accuracy (stdev) %		**top-3 accuracy (stdev) %**		**top-5 accuracy (stdev) %**	
	train	test	train	test	train	test
Vowels	51.6 (12.7)	70.7 (14.1)	87.3 (7.5)	96.2 (2.9)	97.0 (3.4)	100.0 (0.0)
Consonants	52.1 (13.8)	61.7 (16.7)	85.7 (10.6)	93.6 (7.6)	94.2 (7.0)	99.1 (1.3)
Vowels + Consonants	43.0 (10.9)	57.0 (8.4)	76.3 (9.6)	89.2 (5.8)	89.2 (6.0)	97.4 (2.2)

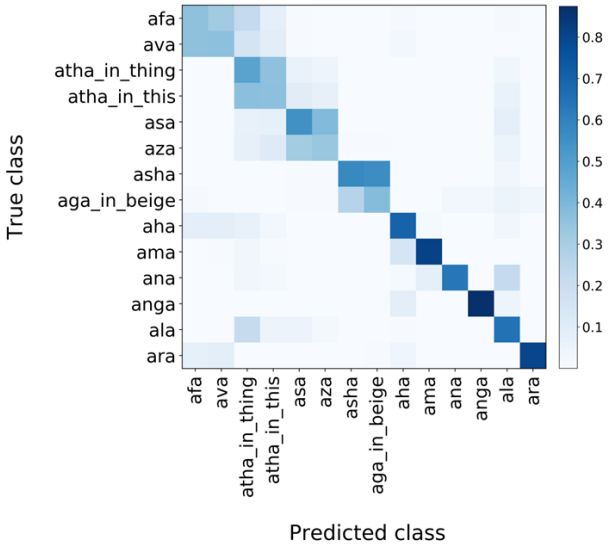

Figure 2: *Confusion matrix of the consonants classification task. Predicted probabilities of the three test subjects of all iterations are combined. Especially where a voiced and unvoiced variant of the phoneme exist confusion can be seen.*

3. Results

3.1. Performance

During hyperparameter tuning, the mean loss was minimized when a batch-size of 4 and drop-out of 0.3 were used, and the data augmentation was performed with a maximum zoom of factor 0.2, a maximum rotation of 20 degrees, a maximum shift of 0.2 and a maximum shear of 0.2.

Table 2 shows the mean accuracy and variance of all iterations of the different tasks over the train and test subjects. The vowel classification task shows the highest accuracy, 70.7%, on the test set. Correctly classifying the dataset with both vowels and consonants was the most difficult task with 27 classes but, with 57.0% accuracy in the test set, performs well above random chance (±4%). Surprisingly, for all tasks, the test set performance is better than the train set performance, though not significant for most metrics. Only for the top-5 accuracies and the top-3 accuracy for the vowel-task, the difference between the two sets is significant (Welch's t-test, p-value <0.05).

In Figure 2, the confusion matrix of the consonants shows how voiced and unvoiced consonants with similar articulation are most easily confused, such as *afa*/*ava* and *asa*/*aza*.

3.2. Embedding

In Figure 3a, the first and second principal components of the embedded space of all samples of a vowel model are plotted.

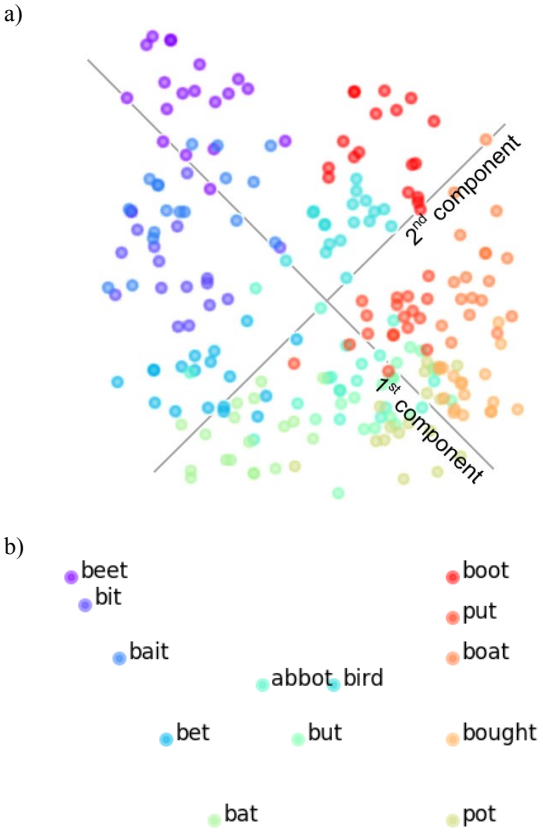

Figure 3: *a) First and second principal component of the output of the flattened layer of all samples based on one of the vowel models. Axes are oriented to align with (b). b) The legend of (a) with the samples mapped in the vowel diagram space. Spatial relations can be seen between the principal components (a) and the well-studied vowel diagram (b). Best viewed in color.*

We mapped the vowels from the USC speech dataset to the well-studied vowel diagram serving as the legend [12] (Figure 3b). The PCA plot is rotated to match the orientation of the vowel diagram. Visually, the embedding shows congruence with the orientation of vowels in the vowel diagram, demonstrating that the neural network learned a similar relation between samples as known to phoneticians. It can be seen that the lower vowels, like *bat* and *pot*, are oriented at the bottom, as opposed to the higher vowels *beet* and *boot* that are projected at the top.

3.3. Saliency

Figure 4 shows the saliency map of six examples from the same subject derived from the consonants+vowels

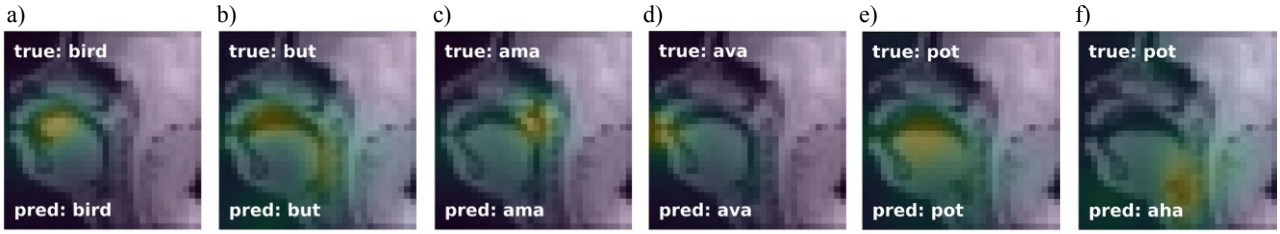

Figure 4: *Saliency maps of one of the subjects. Different input samples are shown with the true class (true) and predicted class (pred). A yellow shade indicates high sensitivity, thus small changes in these pixels in the input have a large effect on the predicted class. Best viewed in electronic form.*

classification model. Regions in the image light up where changes in the input have the most impact on the prediction.

For continuant consonants *atha*, *afa* and *ava* the lips and tongue are of high importance, and for *ama*, *ana* and *anga* the oropharynx. Vocal tract configurations that are less differentiable from the other samples, like *bit* and *bat*, show more widespread attention maps, opposed to well differentiable images like, *bird*, *ama* and *aha*. Most vowels show a more widespread field between the tongue and palate. Figure 4e and 4f show the saliency map of the same input image *pot* at different iterations, which is once classified correctly (e) and once misclassified as *aha* (f). The saliency maps differ accordingly and help explain why misclassification took place as the sensitivity in figure 4f is very local and not most meaningful for predicting the true class *pot*.

The confusion matrices and embedding for the classification tasks not shown here are included in the media files. Also, more examples of saliency maps are provided.

4. Discussion

In this study, we demonstrate that 27 sustained phonemes can be classified from MR images by a convolutional neural network with an accuracy of 57.0% on the test set. Our findings suggest that deep learning represents a viable tool for articulatory-to-acoustic mapping from rtMRI.

Apart from the top-1 accuracy, we considered the top-3 and top-5 accuracy. Accuracy increased between 26% and 33% depending on the task for the top-3 accuracy, which indicates that related samples are misclassified more often. The confusion matrix confirmed this effect. The classification task of both vowels and consonants showed the lowest performance, which can be explained by the fact that twice as many classes were included making the task more difficult.

Most noteworthy, the test set consistently outperformed the train set results. During modeling, it is expected that the dataset, on which hyperparameter tuning was performed, will result in a better performance. One-versus-all cross-validation was performed on each subject to analyze the differences between subjects, while being trained on all other subjects. It appears that the random train/test split gave us three test subjects that outperformed the training set on average. With only 17 subjects in the dataset the risk of having a biased test set is not negligible. The Welch's test shows that the differences in the top-1 accuracy do not significantly differ between the train and test set, thus they come from the same distribution.

The image features learned and visualized in 2D using PCA, demonstrate that the network has indeed learned "sensible" information that resembles the vowel diagram. Furthermore, the saliency maps reveal that the network has learned to focus on the parts of the image that represent the crucial articulatory positions needed to distinguish the different phonemes, such as the lips, tongue and the oropharynx. The saliency maps were not always similar between subjects since the vocal tract configurations differ with each subject. Moreover, it seems that mistakes were made more often when the saliency maps showed places of sensitivity that were not expected to be important for classification. It would be interesting to apply similar methodology to data of subjects with impaired speech to compare to the healthy subjects. The vowel embedding might reveal insight in the way the different phonemes are related to each other. Saliency maps could aid in explaining which articulators are involved in the impairment of phoneme production.

It is expected that the addition of data of more subjects will improve this research. The risk of getting a biased test set from a random split could be avoided and the model would generalize better. Furthermore, limited experiments have been done on the model architecture due to endless options. The network architecture used in this paper is simple and can be trained with limited computing resources and time. Improvements are possible by using other pre-trained image classification networks as Xception or ResNet.

This research aims to use the speech model in combination with a biomechanical tongue model to better understand and predict the changes in articulation due to oral surgery or radiotherapy. The results of this study give us the confidence to proceed in this direction. The next steps are to develop a method for vocal tract segmentation and use the output to train an articulatory-to-acoustic model.

5. Conclusions

The results of this study show the potential of MRI and deep learning as a viable methodology to create a speech model. Analyses of the network provide new insights on what it is that the neural network has learned and 'sees'. This can be used to gain a better understanding of articulation in general and impaired speech in particular.

6. Acknowledgements

First author is a master's student of Technical Medicine.

The authors would like to thank their colleagues Kilian Kappert, Bence Halpern, Rita Simões and Fons Balm for the valuable discussions.

The Department of Head and Neck Oncology and surgery of the Netherlands Cancer Institute receives a research grant from Atos Medical AB (Malmö, Sweden), which contributes to the existing infrastructure for quality of life research.

7. References

[1] S. Narayanan *et al.*, "A multimodal real-time MRI articulatory corpus for speech research," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2011, no. August, pp. 837–840.

[2] S. Silva and A. Teixeira, "Unsupervised segmentation of the vocal tract from real-time MRI sequences," *Comput. Speech Lang.*, vol. 33, no. 1, pp. 25–46, 2015.

[3] K. Mády *et al.*, "Use of real-time MRI in assessment of consonant articulation before and after tongue surgery and tongue reconstruction," in *4th Inter. Speech Motor Conf.*, 2001, pp. 142–145.

[4] E. Bresch, Y.-C. Kim, K. Nayak, D. Byrd, and S. Narayanan, "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging [Exploratory DSP]," *IEEE Signal Process. Mag.*, vol. 25, no. 3, 2008.

[5] T. Sorensen *et al.*, "Database of volumetric and real-time vocal tract MRI for speech science," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017, vol. 2017–Augus, pp. 645–649.

[6] P. Saha, P. Srungarapu, and S. Fels, "Towards Automatic Speech Identification from Vocal Tract Shape Dynamics in Real-time MRI," in *INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association, Proceedings*, 2018, pp. 1249–1253.

[7] S. Narayanan *et al.*, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," *J. Acoust. Soc. Am.*, vol. 136, no. 3, pp. 1307–1311, 2014.

[8] A. Krizhevsky, V. Nair, and G. Hinton, "The CIFAR-10 dataset," *online http//www. cs. toronto. edu/kriz/cifar. html*, 2014.

[9] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv Prepr. arXiv1412.6980*, 2014.

[10] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," pp. 1–8, 2013.

[11] R. Kotikalapudi, "keras-vis." GitHub, 2017.

[12] D. M. Decker, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.