10.3217/978-3-99161-093-9-021

Deep learning based generalizable articulatory feature extraction for speech-BCI applications

Ruoling Wu[1], Julia Berezutskaya[1], Zachary V. Freudenburg[1], Nick F. Ramsey[1,2]

1 University Medical Center Utrecht Brain Center, Utrecht, The Netherlands

2 Donders Institute for Brain, Cognition and Behavior, Radboud University, Nijmegen, The Netherlands

E-mail: nick.ramsey@donders.ru.nl

ABSTRACT: *Objective*: Speech brain-computer interfaces (BCIs) often rely on paired neural-speech data, which is challenging to acquire from individuals with vocal tract paralysis. We aim to bridge this gap by proposing an across-subject framework that synthesizes intelligible speech using generalizable articulatory features extracted from an able-bodied cohort. *Approach:* We develop a subject-independent articulatory-to-speech model to extract generalizable articulatory features from real-time MRI (rtMRI) videos of a healthy cohort and use these features to synthesize text. Main Results: Our framework successfully synthesized text from the extracted generalizable articulatory features by achieving a Phoneme Error Rate (PER) of 18.9 % on unseen subjects. The results confirm that the extracted generalizable features remain robust and decodable across different subjects. Significance: By reducing the dependence on subject-specific articulatory data, our work offers a viable pathway for developing generalizable speech BCIs that could be used by those who can no longer articulate.

INTRODUCTION

The loss of speech can substantially compromise an individual's quality of life as it creates profound barriers to social interaction [1]. Medical conditions such as amyotrophic lateral sclerosis (ALS) or brainstem stroke can cause motor impairments that lead to the inability to control vocal tract muscles to produce speech. While recent advances in brain-computer interfaces (BCIs) show a promising way to restore speech, most current applications rely on paired neural-articulatory or neural-acoustic data from the same individual, which is challenging to acquire from individuals who cannot articulate or vocalize.

To circumvent this challenge, a viable solution is mapping articulatory features shared across a healthy cohort to the neural data of the target users [2], since motor representations remain preserved years after paralysis [3, 4]. Therefore, it is possible to map the motor representations of healthy people to neural data of people with paralysis. This allows for the decoding of intended speech for users with severe paralysis.

Our previous work has demonstrated the possibility of reconstructed generalizable articulatory features extracted from neural data of a different group of participants [2]. However, we did not demonstrate the linguistic effectiveness of the generalizable articulatory features. In this study, to investigate the generalizability and linguistic effectiveness of the extracted features, we propose an across-speaker speech BCI framework (Fig 1). First, generalizable articulatory features are extracted from articulatory data of healthy people. This step is the focus of this study. In the next step, the extracted articulatory features are fed into a neural feature extractor to reconstruct these features from the neural data of a target attempting to articulate the same speech output. The reconstructed articulatory features are then converted into speech via an articulatory-to-speech model.

While previous studies have explored various articulatory-to-speech synthesis approaches using modalities such as Electromagnetic Articulography (EMA) [5, 6], real-time MRI (rtMRI) [7-12], and Electromyography (EMG) [13], existing models are not able to achieve a balance between feature generalizability and robust speech synthesis performance.

To strike the balance, we introduce an across-subject articulatory-to-speech deep learning model to extract generalizable articulatory features from real-time sagittal MRI. Using data from multiple healthy individuals producing the same set of sentences, the model transforms them into text.

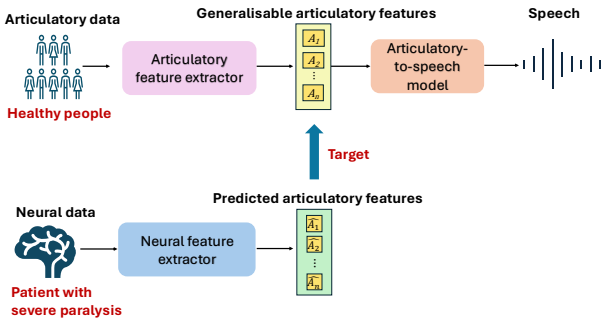

Figure 1: The across-speaker speech BCI framework. Articulatory data from healthy people is firstly used to

10.3217/978-3-99161-093-9-021

train an articulatory feature extractor. Then, the extracted generalizable articulatory features then serve as training targets for a neural feature extractor, which learns to map the neural data of patients with severe paralysis to the generalizable articulatory features. Finally, an articulatory-to-speech model synthesizes speech from the articulatory features predicted from neural data. We here report on the components on the upper panel.

MATERIALS AND METHODS

Across-speaker articulatory feature extraction model: Our proposed model is designed with dual objectives: (1) to extract generalizable articulatory features from different subjects' rtMRI videos (2) to facilitate high-quality text synthesis from these generalizable articulatory features. Fig 2 shows our proposed model architecture. To extract generalizable articulatory features, we used AV-HUBERT, which is known for its robust performance in visual speech recognition [14, 15]. The standard AV-HUBERT consists of four parts: independent video and audio front-end encoders, a shared transformer backbone, and a cluster prediction head. As our study only focuses on the video modality, we disabled the audio front-end encoder and replaced the cluster-prediction head with a fully-connected layer as the phoneme prediction head, as in a previous study [10]. To prevent overfitting on our relatively small dataset, we replaced the original transformer backbone with a smaller transformer consisting of six layers and eight heads. To remove the subject-specific information from the articulatory features, we integrated an adversarial subject classifier inspired by the Domain-Adversarial Neural Networks (DANN) architecture [16]. The subject classifier is connected to the AV-HUBERT encoder via a Gradient Reverse Layer (GRL) [17]. During backpropagation, the GRL flips the sign of the gradients, forcing the AV-HUBERT encoder to learn articulatory features that minimize Connectionist Temporal Classification (CTC) loss [18] $\mathcal{L}_{ctc}$, while maximizing the subject classification loss $\mathcal{L}_s$, defined as:

$$\mathcal{L}_{ctc} = -\ln \sum_{\pi \in B^{-1}(1)} \prod_{t=1}^{T} P(\pi_t | x, t)$$

$$\mathcal{L}_s = -\sum_{i=1}^{K} y_i \log(\hat{y}_i)$$

In the CTC loss $\mathcal{L}_{ctc}$, $\prod_{t=1}^{T} P(\pi_t | x, t)$ is the probability of a single specific predicted phoneme path across all time steps. In the subject classification loss $\mathcal{L}_s$, K is the number of subjects, y_i is the ground truth subject label and $\hat{y}_i$ is the predicted probability for subject label i.

To obtain robust subject-invariant articulatory features, we employed a multi-stage training strategy. In the warm-up stage, we trained the AV-HUBERT encoder and the phoneme predictor together to converge only using $\mathcal{L}_{ctc}$ to ensure the AV-HUBERT encoder can extract linguistically effective subject-dependent articulatory features. Next, without influencing the encoder and the phoneme prediction head, we froze their weights and trained only the subject classifier with $\mathcal{L}_s$ until convergence. After that, we unfroze all layers and trained the model with a combined loss function $\mathcal{L}_{combine}$, defined as

$$\mathcal{L}_{combine} = \mathcal{L}_{ctc} + \alpha \mathcal{L}_s$$

we regulated GRL using the "soft-start" strategy by gradually increasing the GRL's scaling factor α from 0 to 1 via a sigmoid-based schedule. We used AdamW [19] as the optimizer with a learning rate that peaks at 0.001 after 30% updates, following the optimization strategy used [10].

Dataset and preprocessing: We used rtMRI data in the publicly available USC-TIMIT dataset [20]. The dataset consists of rtMRI videos recorded from ten American English speakers (five males M1-M5, five females F1-F5,), each reading the same 460 sentences. The video resolution is 68 x 68 pixels, and the frame rate is 23.18 frame-per-second (fps). To prepare for the sentence-level analysis, we used the transcription API Whiperx [21] to generate precise time-aligned transcriptions. Using these timestamps, each video recording was segmented into sentence-level clips. Following the sentence segmentation, a data preprocessing pipeline (shown in Fig 3) was run on the sentence-level clips: first, to isolate the vocal tract from the background, we defined the Region of Interest (ROI) for each speaker. Followed the pipeline in the previous work [22], we calculated a pixel variance heat map per speaker and identified the outermost borders (top, left, right, bottom) where pixel variance exceeded the mean pixel variance. Each frame was then cropped from its original size of 68 x 68 pixels to 47 x 47 pixels by the borders. After cropping, we rescaled pixel values to the range of a [0, 1] range and applied z-score normalization per clip to reduce flickering. After that we resized the video frame to 96 x 96 pixels and upsampled the temporal resolution from 23.18 fps to 25 fps to align with the model's input requirements [15].

Evaluation metrics: To assess our dual objectives, we use a two-fold evaluation strategy focusing on linguistic effectiveness and generalizability of the extracted features. First, to evaluate the linguistic effectiveness, we quantify the phoneme decoding performance. We implemented a CTC decoder with a pretrained 4-gram language model [15] to decode text from the generalizable articulatory features using Phoneme Error Rate (PER) as the evaluation metric. PER measures the distance between the decoded and the ground-truth phoneme sequence. It is defined as,

$$PER = (S + D + I)/N,$$

where S, D, I represent the total number of substitutions, deletions, and insertions required to align the two sequences, respectively, and N represents the total number of phonemes in the ground-truth phoneme sequence. Second, to quantify feature generalizability, we trained a k-nearest neighbors (KNN) classifier on the extracted articulatory features to predict subject labels. The model is cross-validated across all features using a 10-fold cross-validation scheme, providing a

10.3217/978-3-99161-093-9-021

measurement of how effectively subject-specific information was removed across the entire cohort.

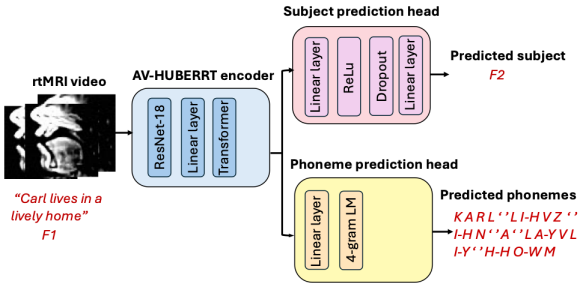

Figure 2: Model architecture. Our proposed model is composed of three main components: an AV-HUBERT Encoder that extracts articulatory features, followed by two parallel task heads: the phoneme prediction head and the subject prediction head. The phoneme prediction head consists of a linear layer and a 4-gram pretrained language model (LM). It converts the extracted features into phonemes to ensure the linguistic effectiveness, while the subject prediction head attempts to identify the speaker's identity.

Figure 3: rtMRI preprocessing pipeline. Raw rtMRI videos were cropped from 68 x 68 pixels to 47 x 47 pixels ROIs. Next, pixels were rescaled to a range of [0-1], normalized by z-score. Finally, the frames are spatially upsampled to 96 x 96 pixels and temporally upsampled to 25 fps to meet the input requirement of AVHUBERT.

RESULTS

Feature generalizability across subjects: We evaluated the subject-invariance of features extracted from both models by using a KNN classifier to assess their ability to classify subject labels. As shown in Figure 4, the mean subject classification accuracy for features from the subject-variant AV-HUBERT model is 0.98. In contrast, the accuracy for the subject-invariant model is significantly lower at 0.51 ($p<0.05$, paired t-test). This means that the subject-invariant model effectively lost the ability to differentiate individual subjects, and thus gained generalizability.

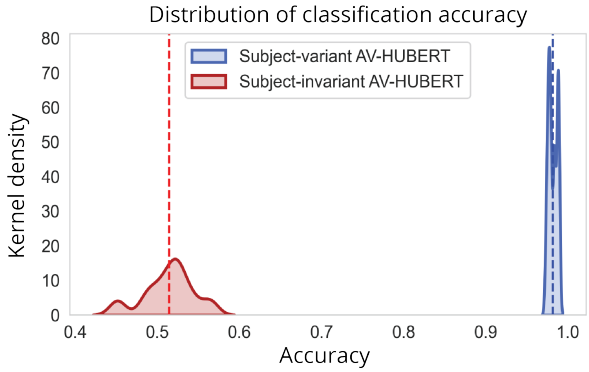

Figure 4: Distribution of subject-label classification accuracy. Curves represent the Kernel Density Estimation (KDE) of classification accuracies across subjects for both models. The subject-variant AV-HUBERT model (blue) achieves a mean accuracy of 0.98, while the subject-invariant AV-HUBERT model (red) shows a significantly reduced mean accuracy of 0.51. Vertical dashed lines indicate the respective mean values for each distribution.

Feature effectiveness evaluation: As previously described, we evaluated linguistic effectiveness of articulatory features based on the phoneme decoding performance. Table 1 shows the PER on each unseen subject. The subject-invariant AV-HUBERT consistently outperformed the subject-variant, achieving a mean PER of 18.9 ± 1.5, compared to 32.1 ± 1.9 for the subject-variant model.

Table 1: PER (%) on each unseen subject and their average

Subject	subject-invariant model	subject-variant model
F1	18.2	32.6
F2	17.7	31.1
F3	19.1	29.6
F4	20.6	32.5
F5	16.1	28.6
M1	21.2	35.5
M2	18.5	32.9
M3	19.6	33.5
M4	20.7	34.0
M5	17.4	30.9
AVG	18.9 ± 1.5	32.1 ± 1.9

DISCUSSION

In this work, we proposed an across-speaker MRI-to-speech framework that extracts generalizable articulatory features across subjects. Our framework outperformed the baseline model [9] with a phoneme error rate of 18.9% on the synthesized sentences, compared to 32.9 % PER of the subject-variant model. These results indicate that the extracted articulatory features are both generalizable and linguistic effective.
Regarding the improved performance of the subject-invariant model, we hypothesize that the linguistic

10.3217/978-3-99161-093-9-021

effectiveness benefits from the feature generalizability. Since the target outputs, i.e., the sentence text, are invariant across subjects, we speculate that forcing the model to 'unlearn' the subject-specific information allows it to focus on the articulatory information most relevant to the phoneme prediction. To explore this speculation, we need to further examine the model's learned behavior through a saliency study [23].

To quantify the feature generalizability, we trained a KNN classifier to identify the subject labels from the extracted articulatory features. Our result shows that the classification accuracy for features extracted from the subject-invariant model is significantly lower than that of the subject-variant model. This suggests that the subject-invariant model effectively eliminated a large part of subject-specific information during the encoding process. However, the classification accuracy for the invariant model remains above the chance level (0.10), indicating that some residual subject information persists within the extracted features. Future work will focus on further eliminating the residual subject-specific information while preserving linguistic information for text decoding.

Taken together, our proposed model extracts articulatory features from rtMRI videos that are both generalizable and linguistically effective. Next, we will map these features to neural data from a different subject, allowing us to develop across-speaker speech BCIs.

CONCLUSION

We demonstrated that our proposed framework can extract generalizable articulatory features from which speech can be synthesized with a low phoneme error rate. Using these articulatory features offers great potential for developing generalizable speech BCIs and reduces the need for a paired articulatory-neural or articulatory acoustic data, which is challenging to obtain from people with severe paralysis.

ACKNOWLEDGEMENT

This work is supported by Dutch Brain Interface Initiative (DBI2), project number 024.005.022 of the research programmed Gravitation, which is financed by the Dutch Ministry of Education, Culture, and Science (OCW) via the Dutch Research Council (NWO).

REFERENCES

[1] M. Walshe and N. Miller, "Living with acquired dysarthria: the speaker's perspective," *Disability and rehabilitation,* vol. 33, no. 3, pp. 195–203, 2011.

[2] R. Wu, J. Berezutskaya, Z. V. Freudenburg, and N. F. Ramsey, "Across-speaker articulatory reconstruction from sensorimotor cortex for generalizable brain-computer interfaces," *bioRxiv,* p. 2025.10. 08.680888, 2025.

[3] M. L. Bruurmijn, I. P. Pereboom, M. J. Vansteensel, M. A. Raemaekers, and N. F. Ramsey, "Preservation of hand movement representation in the sensorimotor areas of amputees," *Brain,* vol. 140, no. 12, pp. 3166–3178, 2017.

[4] Z. V. Freudenburg *et al.*, "Sensorimotor ECoG signal features for BCI control: a comparison between people with locked-in syndrome and able-bodied controls," *Frontiers in neuroscience,* vol. 13, p. 1058, 2019.

[5] P. Wu, S. Watanabe, L. Goldstein, A. W. Black, and G. K. Anumanchipalli, "Deep speech synthesis from articulatory representations," *arXiv preprint arXiv:2209.06337,* 2022.

[6] F. Bocquelet, T. Hueber, L. Girin, C. Savariaux, and B. Yvert, "Real-time control of an articulatory-based speech synthesizer for brain computer interfaces," *PLoS computational biology,* vol. 12, no. 11, p. e1005119, 2016.

[7] Y. Otani, S. Sawada, H. Ohmura, and K. Katsurada, "Speech Synthesis from Articulatory Movements Recorded by Real-time MRI," in *Interspeech*, 2023, pp. 127–131.

[8] T. G. Csapó, "Speaker dependent articulatory-to-acoustic mapping using real-time MRI of the vocal tract," *arXiv preprint arXiv:2008.00889,* 2020.

[9] L. Pandey and A. Sabbir Arif, "Silent speech and emotion recognition from vocal tract shape dynamics in real-time MRI," in *ACM SIGGRAPH 2021 Posters*, 2021, pp. 1–2.

[10] N. Shah, A. Kashyap, S. Karande, and V. Gandhi, "MRI2Speech: Speech Synthesis from Articulatory Movements Recorded by Real-time MRI," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025: IEEE, pp. 1–5.

[11] P. Wu *et al.*, "Deep speech synthesis from MRI-based articulatory representations," *arXiv preprint arXiv:2307.02471,* 2023.

[12] Y. Yu, A. H. Shandiz, and L. Tóth, "Reconstructing speech from real-time articulatory MRI using neural vocoders," in *2021 29th European Signal Processing Conference (EUSIPCO)*, 2021: IEEE, pp. 945–949.

[13] K. Scheck *et al.*, "Cross-speaker training and adaptation for electromyography-to-speech conversion," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024: IEEE, pp. 1–4.

[14] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496,* 2018.

[15] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech

10.3217/978-3-99161-093-9-021

representation by masked multimodal cluster prediction," *arXiv preprint arXiv:2201.02184,* 2022.

[16] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *Journal of machine learning research,* vol. 17, no. 59, pp. 1–35, 2016.

[17] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*, 2015: PMLR, pp. 1180–1189.

[18] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[19] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101,* 2017.

[20] S. Narayanan *et al.*, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," *The Journal of the Acoustical Society of America,* vol. 136, no. 3, pp. 1307–1311, 2014.

[21] M. Bain, J. Huh, T. Han, and A. Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," *arXiv preprint arXiv:2303.00747,* 2023.

[22] E. Stolwijk, "Towards speech-based brain-computer interfaces: finding most distinguishable word articulations with autoencoders," 2023.

[23] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034,* 2013.