Interspeech 2024
1-5 September 2024, Kos, Greece

Towards Speech Classification from Acoustic and Vocal Tract data in Real-time MRI

Yaoyao Yue[1], *Michael Proctor*[2], *Luping Zhou*[1], *Rijul Gupta*[1], *Tharinda Piyadasa*[1], *Amelia Gully*[3], *Kirrie Ballard*[1], *Craig Jin*[1]

[1]The University of Sydney, Australia [2]Macquarie University, Australia [3]University of York, UK

{yaoyao.yue, luping.zhou, rijul.gupta, tharinda.piyadasa, kirrie.ballard, craig.jin}@sydney.edu.au, michael.proctor@mq.edu.au, amelia.gully@york.ac.uk

Abstract

Real-time magnetic resonance image (rtMRI) data of the upper airway provides a rich source of information about vocal tract shaping that can inform phonemic analysis and classification. We describe a multimodal phonemic classifier that combines articulatory data with speech audio features to improve performance. A deep network model processes rtMRI video data using ResNet18 and speech audio using a custom CNN and then combines the two data streams using a Transformer layer to fully explore the correlation of the two streams towards better vowel-consonant-vowel classification via the Transformer's multi-head self-attention mechanism. The classification accuracy of both the unimodal and multimodal models show substantial improvement on previous work ($> 38\%$). The addition of audio features improves classification accuracy in the multimodal model by 7% compared with the unimodal model using articulatory data. We analyze the model and discuss the phonetic implications.

Index Terms: vocal tract, phonemic classification, multimodal networks, real-time MRI, Transformer

1. Introduction

Speech production is characterized by continuous dynamic reconfiguration of the upper airway arising from the coordination of speech gestures [1, 2]. The goals of production that give rise to different vocal tract configurations in anatomically diverse speakers and the relationships between different vocal tract geometries and the acoustic speech signal are still imperfectly understood and an active area of research [3, 4, 5]. Mapping vocal tract configurations to phonological structures has proven to be difficult because of the variability and complex interactions between articulators, because it remains difficult to accurately visualize the coordination of various articulators for speech production, and because the relationship between morphological features and phonological features is itself complicated [6, 7]. New insights into these issues are afforded by real-time imaging of the upper airway [8, 9], a key method providing temporally and spatially rich information about the vocal tract during speech production [10].

There are challenges with analyzing real-time images of the vocal tract, often related to the complexity of speech. Previous studies of vocal tract configuration using deep learning models based on mid-sagittal rtMRI data have primarily used image data alone. Saha et al. [11] classified 54 vowel-consonant-vowel (VCV) combinations with an accuracy rate of 42% using the USC Speech and Vocal Tract Morphology MRI Database [8]. Van Leeuwen et al. [12] classified 27 sustained phonemes with an accuracy rate of 57% using the same dataset. Multimodal co-learning [13] may offer improvements in the analysis of vocal tract rtMRI images, especially in regards to enrichment of the representation and/or latent space of the network models. In this light, the advantages of multimodal co-learning classification approaches have been demonstrated in many aspects of speech processing and automatic speech recognition [14, 15, 16], as models can leverage the complementarity between different modalities to improve classification performance and robustness. For example, Köse et al. [17] showed that integrating rtMRI video and speech audio data offers improved performance over a unimodal approach in 39 phone classification tasks, using the USC-TIMIT dataset [9].

In this work, we consider models that utilize the Transformer architecture. Because of their self-attention mechanism, Transformers [18] have gained prominence across various multimodal models, due to their excellent capacity to model long range dependency. For example, Transformer-based models have shown promise in various speech tasks like acoustic-to-articulatory inversion (AAI) [19, 20], spoken language understanding (SLU) [21], automatic speech recognition (ASR) [22] and speech translation (ST) [23]. Here, we develop and explore a multimodal neural network model for a phonemic VCV classification task that takes midsagittal image and acoustic speech data as input, and uses a Transformer network to effectively fuse the two streams. To the best of our knowledge, this work is the first study that applies a Transformer network for the analysis of rtMRI data corresponding to a VCV task and analyzes the attention matrices to obtain an improved understanding of the model behavior. The classification accuracy of both unimodal and multimodal models compares favourably with previous work, demonstrating the benefits of including a Transformer network. Furthermore, we demonstrate the improvements that can be obtained for VCV identification using a multimodal approach combining rtMRI data and acoustic speech data.

2. Method

2.1. Data Preprocessing

This work [1] uses the USC Speech and Vocal Tract Morphology MRI Database [8] which consists of rtMRI videos of 17 speakers (9 female, 8 male) with synchronized audio recordings for 54 different VCV sequences. The videos capture the midsagittal posture of the entire upper vocal tract at a frame rate of 23.18 frames/sec and a spatial resolution of 68×68 pixels over a 200 mm $\times$ 200 mm field of view centred on the tongue body. The VCV sequences for each subject are spread across three recordings, each containing 18 utterances. Speech audio is recorded at 20 kHz sampling rate.

In the preprocessing phase, we evaluated 54 VCV utter-

[1]Research supported by ARC Grant DP220102933.

10.21437/Interspeech.2024-840

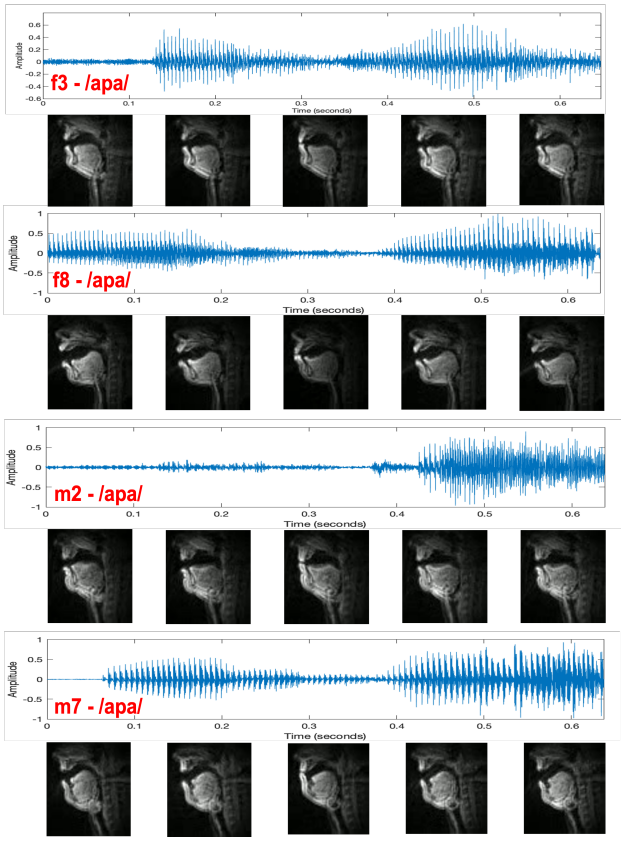

Figure 1: *Example utterances from the USC speech and Vocal Tract Morphology MRI Database [8]. Each 647 ms audio interval is centred on the video frame in which the articulatory target of the intervocalic consonant is achieved.*

ances from 17 subjects to identify a suitable analysis window duration. Analysis of utterance lengths indicated that a 15-frame window spanning 647 ms was sufficient to capture the articulation of all VCVs in the dataset. The 15-frame window containing each utterance was manualy located by inspecting the audio and video to identify the center frame corresponding to the the primary articulatory target of the intervocalic consonant. In cases where target articulatory postures spanned multiple frames, gestural trajectories of other articulators were used to select the frame corresponding to the consonantal target. The 647 ms audio window corresponding to the 15-frame video sequence was extracted for each VCV token. Fig. 1 illustrates audio analysis windows for four utterances of /apa/ by different speakers, with time-aligned video frames. Audio features were extracted for each audio frame by shifting a Hanning window across the speech signal, where the window length was 51.2 ms and the hopsize was 12.8 ms. The audio features consist of the first 513 Fourier transform magnitude values per frame obtained from applying a discrete-time Fourier transform of size 1024.

2.2. Multimodal Approach for VCV Classification

The proposed VCV classifier is shown in Fig. 2 and consists of a feature extraction model and a Transformer-based classification model. The feature extraction model consists of two components: an audio backbone and a video backbone. The unimodal model and multimodal model had 30 M and 37 M parameters, respectively.

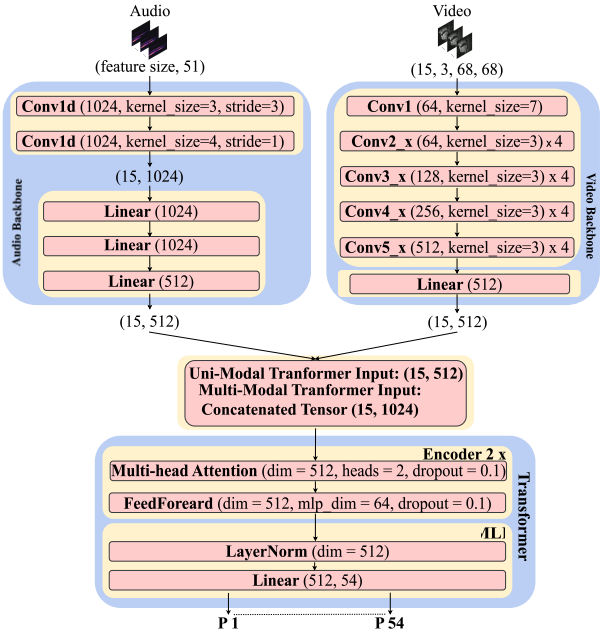

Figure 2: *The proposed network model for VCV classification combines a Convolutional network for acoustic data, ResNet18 for rtMRI data, and a Transformer model to integrate information across the data frames.*

Audio Backbone: To extract the relevant audio information, a convolutional neural network consisting of two sequential processing blocks is used: a convolutional block and a linear block. The convolutional block includes two 1D convolutional layers. The output of the audio backbone is a 2-D tensor of size [15, 512] matching the output size of the video backbone.

Video Backbone: The video backbone consists of a ResNet18 model which is applied to each image in the video and is used to extract a sequence of image features. In keeping with the structure of ResNet18, the rtMRI image data are formatted as 3-channel RGB data all with the same grayscale values. The output of ResNet18 (4-D tensor with size [15, 512, 7, 7]) is reshaped to [15, 25088] and is fed into a fully connected layer to obtain an image feature of size [15, 512].

Transformer: The final part of the proposed model is a Transformer network, using the the self-attention mechanism [19] which is known to improve the modeling of sequence data. The concatenation of the image and speech encodings is referred to as late data fusion and the Transformer network is well-suited to integrating such data using the attention mechanism. The output of the Transformer is the class probabilities for the 54 VCV classes.

3. Experiments

This study investigates models for the classification of 54 English VCV utterances, using midsagittal rtMRI sequences and companion speech audio data. These VCV sequences provide a rich dataset for speech classification, as they feature intervocalic consonants – some sharing place and manner of articulation – coarticulated with flanking context vowels potentially differentiated by tongue and lip posture. The proposed method is compared with several state-of-the-art techniques using the USC Speech and Vocal Tract Morphology MRI Database [24].

A training set was constructed from 14 speakers (7 female, 7 male; approximately 80% of total data), and the test set contained the remaining 3 speakers (2 female, 1 male; approximately 20% of total data).

In order to avoid overfitting and poor generalization given the limited data, we applied several data augmentation techniques. For image data, *RandomHorizontalFlip*(p=0.5), *RandomRotation*(15), and *RandomAdjustSharpness*(sharpness_factor=0.5, p=0.5) from torchvision were used during training. For audio data, *TimeMasking*(30), *TimeMasking*(10), and *FrequencyMasking*(30) from torchaudio (2.0.2) [25] were used. The ResNet18 model used was pretrained on ImageNet with a batch size of 16, 100 epochs, and an AdamW [26] optimizer with a learning rate of 0.001 for training. We used the Monte Carlo method for tuning hyperparameters while building our models and then used the same configuration for all subsequent models. Early-stopping with a latency of 20 epochs was applied. Testing was conducted with a batch size of four.

The model was initialised using the default random initialization mechanism included in PyTorch (2.0.1) [27]. To ensure reproducibility, we fixed the random seed of used libraries to 42 and configured PyTorch and CUDA to use deterministic implementation. Training was conducted on a P40 GPU, achieving an average duration of approximately 260 seconds per epoch.

3.1. Results

We evaluate the various neural network models for VCV classification, using categorical cross-entropy as the loss measure and top-1 categorical accuracy to analyze the performance of the models. It's important to note that our data are balanced across different categories, which means that overall metrics and balanced metrics are equivalent in this context. This balance ensures that our performance evaluations are not biased by disproportionate representation of any category, allowing for a more accurate assessment of model effectiveness.

The evaluation method includes the calculation of confidence intervals for model performance. We used a bootstrap method [28, 29] to conduct 1,000 sampling runs, each with a put-back from the dataset, to simulate different data distributions. With this method, we calculated confidence intervals for model accuracy, specifically by calculating the 5% and 95% quartiles of accuracy for these 1,000 sampling runs. This process provides additional insights into the stability of the model's performance and can help assess fluctuations in the model's performance on different subsets of data.

Table 1: *Comparison of the proposed model with several reference models on a VCV classification task with 54 different VCVs*

Model	Acc.	Confidence Interval
Unimodal Approach, Video [11]	42.04%	[40.81%, 44.3%]
Unimodal Approach, Video [17]	39.02%	[36.69%, 41.95%]
Unimodal Approach, Video [ours]	80.24%	[78.19%, 82.3%]
Multimodal Approach, Video and Audio [17]	21.34%	[19.01%, 23.68%]
Multimodal Approach, Video, and Audio [ours]	**87.86%**	[85.79%, 89.51%]

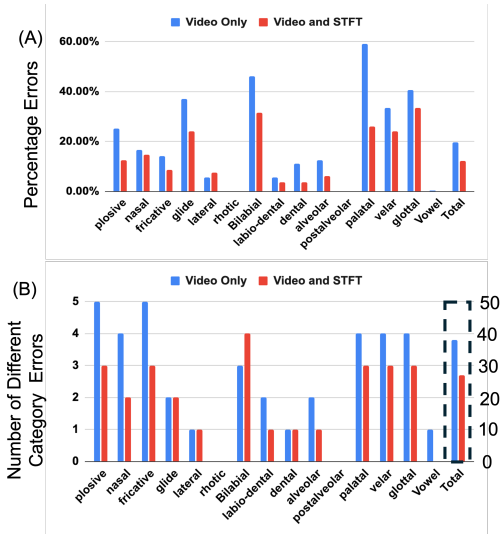

Figure 3: *Percentage of errors in test data analysis by place and manner of articulation (top). The number of different category errors (bottom).*

3.1.1. Comparison with state-of-the-art (SOTA) models

While work in [11] focuses on VCV classification like our study, the work in [17] focuses on a different classification task constructed out of categories that include 5 places of articulation and 7 manners of articulation, as well as a 39 phoneme classification. As a result, we compare our unimodal approach with the results of [11] directly but are unable to perform a direct comparison with the unimodal and multimodal approaches of [17]. As previous work did not release the models publicly, we replicated the models using the configuration provided in the literature. Our replication study of the work in [17] presents subpar results in comparison to our multimodal approach on the task of VCV classification using the USC-TIMIT [9] dataset. We present all of these results in Table 1.

The dataset [24] used in this study includes 54 classes of VCV combinations, consisting of vowels /a/, /u/, and /i/ paired with 18 consonants. However, relying solely on video data makes it challenging to distinguish certain consonants, particularly those sharing places and manners of articulation. For consonant pairs such as /p/-/b/, /t/-/d/, and /k/-/g/, for example, the only distinguishing phonological feature is voicing. A confusion matrix considering these challenges revealed a theoretical ceiling performance of 77.78%, which represents the highest achievable accuracy by a classifier using visual information alone, if voicing distinctions do not greatly influence midsagittal posture.

The results in Table 1 indicate that the Transformer network provides substantial improvement ($> 38\%$) compared with previous models. Comparing the performance of the proposed network model using unimodal and multimodal data indicates multimodal performance improves by about 7%. Interestingly, our unimodal model exceeds the theoretical performance ceiling for video-only data, suggesting that systematic differences do exist in these rtMRI images between voiced and unvoiced consonants with equivalent manner and place [30, 31].

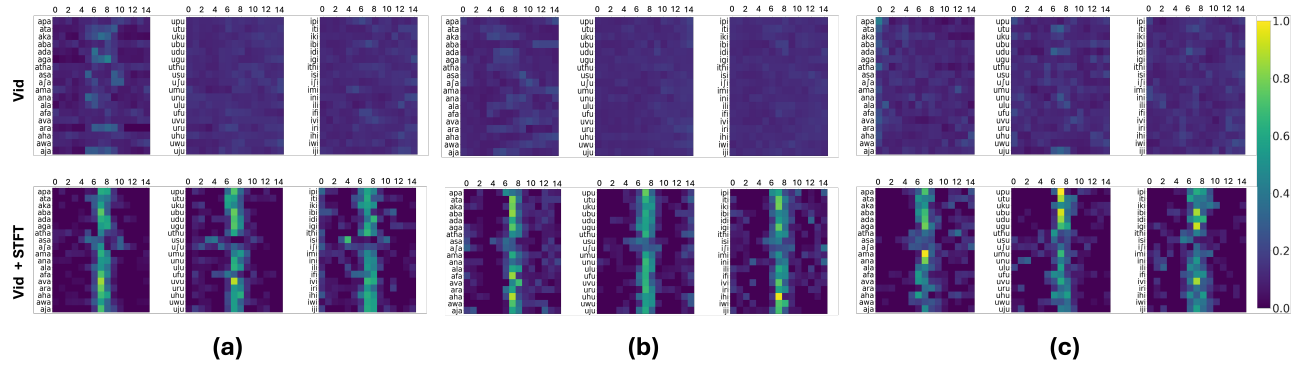

Figure 4: *The attention values for each frame in the entire VCV sequence obtained from the video-only model (top) and multimodal model (bottom) for 54 VCVs with respect to f6 (a), f7 (b), and m2 (c).*

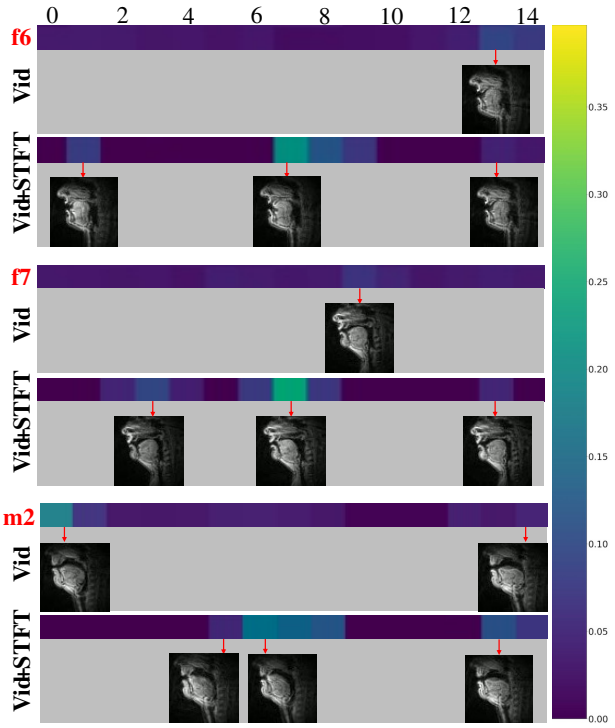

Figure 5: *The attention values and the corresponding video frames corresponding to subjects f6 on /ini/, f7 on /udu/, and m2 on /ata/ using the model with unimodal data (top) and multimodal data (bottom).*

More insights into the performance differences between the unimodal and multimodal models are provided through systematic analysis of classification errors (Fig. 3). Fig. 3A shows the percentage of total errors based on the specific place and manner of articulation. The general trend is for improved performance using multimodal data. Fig. 3B illustrates how many category errors are made, i.e. incorrectly classified voice, place, and/or manner. The results indicate that the multimodal model reduces the types of classification error made.

4. Analysis and Discussion

To better understand the performance of our network model, we explore the attention matrices from the Transformer and the frames of the rtMRI videos that receive the most attention from the model. As there are two encoders with eight multi-head attention units, we obtain 16 attention matrices. We perform a z-score normalization on each attention matrix and then threshold at one positive standard deviation, so that all values below the threshold are set to zero. We then calculate the mean of the normalized-and-thresholded attention matrices and apply a second threshold operation, again using one positive standard deviation as the threshold value (Fig. 5). We shall refer to the attention matrix values as 'the attention values'.

Consider Fig. 5, in which attention values for the 15 frames of three example utterances are illustrated, along with video frames corresponding to the largest attention values. The top row of each panel shows attention values for the unimodal model, and the bottom row, attention values for the multimodal model. We can observe that the multimodal model better 'attends' to the central video frame, which corresponds to the consonant articulation. To further explore this observation, Fig. 4 shows the attention values for all of the 54 VCVs for the three test subjects. In order to simplify the image, we summed the attention values for the three repeats of each VCV. Fig. 4 clearly indicates that the multimodal model demonstrates improved attention to the frame with the consonant, and likely accounts for some of the improved performance of the multimodal model over the unimodal model.

5. Conclusion

In this work, we explore VCV classification based on rtMRI data and acoustic speech data, and propose unimodal models using rtMRI data only, and a multimodal model integrating rtMRI data and audio data. The multimodal model combines a ResNet18 network to process the image data and a CNN network to process the audio data, and a Transformer to integrate the image and audio data embeddings. Visualization of the Transformer attention matrices provides insight into how the model attends to different frames of an input sequence. It demonstrates that the model has the ability to adjust its focus to highlight important frames that may contain critical features for the VCV classification task. In the future, we will explore more detailed and dynamic analysis of the rtMRI data, which will hopefully facilitate our understanding and analysis of speech acoustics.

6. References

[1] E. L. Saltzman and K. G. Munhall, "A Dynamical Approach to Gestural Patterning in Speech Production," *Ecological Psychology*, vol. 1, no. 4, pp. 333–382, Dec. 1989.

[2] C. P. Browman and L. Goldstein, "Articulatory Phonology: An Overview," *Phonetica*, vol. 49, no. 3-4, pp. 155–180, May 1992.

[3] S. Adachi, H. Takemoto, T. Kitamura, P. Mokhtari, and K. Honda, "Vocal tract length perturbation and its application to male-female vocal tract shape conversion," *The Journal of the Acoustical Society of America*, vol. 121, no. 6, pp. 3874–3885, Jun. 2007.

[4] V. Ramanarayanan, M. Van Segbroeck, and S. S. Narayanan, "Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories," *Computer Speech & Language*, vol. 36, pp. 330–346, Mar. 2016.

[5] P. K. Ghosh and S. Narayanan, "Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. EL251–EL257, Oct. 2011.

[6] P. Ladefoged, R. Harshman, L. Goldstein, and L. Rice, "Generating vocal tract shapes from formant frequencies," *The Journal of the Acoustical Society of America*, vol. 64, no. 4, pp. 1027–1035, Oct. 1978.

[7] G. Fant, *Speech Acoustics and Phonetics: Selected Writings*, 1st ed. Berlin: Kluwer Academic Publishers, 2004.

[8] 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. S. Narayanan, "Database of Volumetric and Real-Time Vocal Tract MRI for Speech Science," in *Interspeech 2017*. ISCA, Aug. 2017, pp. 645–649.

[9] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein, D. Byrd, E. Bresch, P. Ghosh, A. Katsamanis, and M. Proctor, "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, Sep. 2014, number: 3.

[10] S. G. Lingala, B. P. Sutton, M. E. Miquel, and K. S. Nayak, "Recommendations for real-time speech MRI: Real-Time Speech MRI," *Journal of Magnetic Resonance Imaging*, vol. 43, no. 1, pp. 28–44, Jan. 2016, number: 1.

[11] P. Saha, P. Srungarapu, and S. Fels, "Towards Automatic Speech Identification from Vocal Tract Shape Dynamics in Real-time MRI," in *Interspeech 2018*. ISCA, Sep. 2018, pp. 1249–1253.

[12] K. V. Leeuwen, P. Bos, S. Trebeschi, M. V. Alphen, L. Voskuilen, L. Smeele, F. V. D. Heijden, and R. V. Son, "CNN-Based Phoneme Classifier from Vocal Tract MRI Learns Embedding Consistent with Articulatory Topology," in *Interspeech 2019*. ISCA, Sep. 2019, pp. 909–913.

[13] F. Pahde, M. Puscas, T. Klein, and M. Nabi, "Multimodal Prototypical Networks for Few-shot Learning," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Waikoloa, HI, USA: IEEE, Jan. 2021, pp. 2643–2652.

[14] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, no. 3-4, pp. 303–319, Jul. 2002.

[15] V. Mitra, H. Nam, C. Y. Espy-Wilson, E. Saltzman, and L. Goldstein, "Articulatory Information for Noise Robust Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1913–1924, Sep. 2011.

[16] T. Kanamaru, T. Arakane, and T. Saitoh, "Isolated single sound lip-reading using a frame-based camera and event-based camera," *Frontiers in Artificial Intelligence*, vol. 5, p. 1070964, Jan. 2023.

[17] O. D. Köse and M. Saraclar, "Multimodal Representations for Synchronized Speech and Real-Time MRI Video Processing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1912–1924, 2021.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[19] S. Hu, X. Xie, Z. Jin, M. Geng, Y. Wang, M. Cui, J. Deng, X. Liu, and H. Meng, "Exploring Self-Supervised Pre-Trained ASR Models for Dysarthric and Elderly Speech Recognition," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Rhodes Island, Greece: IEEE, Jun. 2023, pp. 1–5.

[20] S. Udupa and P. K. Ghosh, "Real-Time MRI Video Synthesis from Time Aligned Phonemes with Sequence-to-Sequence Networks," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Rhodes Island, Greece: IEEE, Jun. 2023, pp. 1–5.

[21] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal Transformer for Unaligned Multimodal Language Sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 6558–6569.

[22] F.-J. Chang, M. Radfar, A. Mouchtaris, B. King, and S. Kunzmann, "End-to-End Multi-Channel Transformer for Speech Recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, ON, Canada: IEEE, Jun. 2021, pp. 5884–5888.

[23] Z. Wu, O. Caglayan, J. Ive, J. Wang, and L. Specia, "Transformer-based Cascaded Multimodal Speech Translation," in *Proceedings of the 16th International Conference on Spoken Language Translation*. Hong Kong: Association for Computational Linguistics, Nov. 2019.

[24] Y. Lim, A. Toutios, Y. Bliesener, Y. Tian, S. G. Lingala, C. Vaz, T. Sorensen, M. Oh, S. Harper, W. Chen, Y. Lee, J. Töger, M. L. Monteserin, C. Smith, B. Godinez, L. Goldstein, D. Byrd, K. S. Nayak, and S. S. Narayanan, "A multispeaker dataset of raw and reconstructed speech production real-time MRI video and 3D volumetric images," *Scientific Data*, vol. 8, no. 1, p. 187, Dec. 2021.

[25] J. Hwang, M. Hira, C. Chen, X. Zhang, Z. Ni, G. Sun, P. Ma, R. Huang, V. Pratap, Y. Zhang, A. Kumar, C.-Y. Yu, C. Zhu, C. Liu, J. Kahn, M. Ravanelli, P. Sun, S. Watanabe, Y. Shi, and Y. Tao, "TorchAudio 2.1: Advancing Speech Recognition, Self-Supervised Learning, and Audio Processing Components for Pytorch," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. Taipei, Taiwan: IEEE, Dec. 2023, pp. 1–9.

[26] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*. New Orleans: OpenReview.net, May 2019.

[27] 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, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Dec. 2019, pp. 8026–8037.

[28] B. Efron and R. Tibshirani, *An introduction to the bootstrap*, ser. Monographs on statistics and applied probability. New York: Chapman & Hall, 1993, no. 57.

[29] P. Ferrer, L. and Riera, "Confidence Intervals for evaluation in machine learning." [Online]. Available: https://github.com/luferrer/ConfidenceIntervals

[30] R. D. Kent and K. L. Moll, "Vocal-Tract Characteristics of the Stop Cognates," *The Journal of the Acoustical Society of America*, vol. 46, no. 6B, pp. 1549–1555, Dec. 1969.

[31] M. I. Proctor, C. H. Shadle, and K. Iskarous, "Pharyngeal articulation in the production of voiced and voiceless fricatives," *The Journal of the Acoustical Society of America*, vol. 127, no. 3, 2010.