



75-Speaker Annot-16: A benchmark dataset for speech articulatory rt-MRI annotation with articulator contours and phonetic alignment

Xuan Shi^{1*}, Yubin Zhang^{1*}, Yijing Lu^{2*}, Marcus Ma^{1*}, Tiantian Feng^{1*}, Asterios Toutios³, Haley Hsu¹, Louis Goldstein¹, Shrikanth Narayanan¹

¹University of Southern California, United States

²University of Potsdam, Germany

³EarliTec Diagnostics, United States

louisgol@usc.edu, shri@usc.edu

Abstract

High-quality speech articulatory databases are essential for advancing speech science and technology research. However, the lack of standardized annotations limits their full potential use and broad accessibility. In this context, we introduce 75-Speaker Annot-16, a comprehensive annotation dataset derived from the 75-Speaker vocal tract MRI database. Annot-16 provides phonetic alignments, articulator contour annotations, and handmade ground-truth articulator contours. Our annotation process integrates automated algorithms with expert verification to ensure accuracy and efficiency. To demonstrate its utility, we establish three benchmark tasks: speech phoneme recognition, articulatory contour segmentation, and articulatory phoneme recognition. Annot-16 can serve as a valuable resource for speech modeling, computer vision, and cross-modal learning, bridging engineering applications, speech science, and linguistic research. Database webpage: https://sail.usc.edu/span/75speakers_annot/.

Index Terms: speech articulatory database, articulator contour, phonetic alignment, annotation

1. Introduction

Speech articulatory databases are valuable resources for research in linguistic phonetics, speech-language pathology, and speech engineering. In particular, recently released real-time magnetic resonance imaging (rt-MRI) articulatory databases of American English [1, 2, 3], French [4] and Japanese [5, 6] have contributed significantly to these fields, as they provide much more detailed articulatory information in both spatial and temporal dimensions than other types of articulatory databases such as electromagnetic articulography (EMA) and ultrasound.

However, there is currently no general-purpose annotated rt-MRI database that can serve as a benchmark for future research. To this end, we aim to develop such a benchmark database for American English, which we hope can serve as both a resource for fundamental scientific research and as reference for evaluating future speech technologies. In this paper, we present an annotated articulatory dataset of American English: the 75-Speaker Annot-16 Database (Annot-16 hereafter) that includes detailed annotations for 16 representative speakers (8 native and 8 non-native speakers) from the widely recognized 75-Speaker vocal tract speech database [1]. In Annot-16, three types of annotations are created: phonetic alignments for the audio recordings, articulator contour annotations for the MRI videos, and high-quality handmade ground-truth annotations of key articulator contours for selected speech frames [1].

Our annotation method combines automated algorithms with expert verification and curation to balance efficiency and

accuracy. For phonetic alignment, phonetic labels are automatically generated using the Montreal Forced Aligner and manually checked by phoneticians. For the articulator contour extraction, we adopt a semi-automatic region-based segmentation algorithm [7], which with a small amount of initial human annotation, automatically generates articulator contour tracks. For creating the high-quality handmade ground truth here, several speech experts collaborate to manually annotate articulator boundaries from selected speech frames.

Annot-16 database complements the public 75-Speaker database and opens opportunities to develop novel multimodal engineering approaches in speech production modeling. We present three benchmarking efforts to demonstrate the promising application scenarios. Our first task is speech phoneme recognition: we employ a widely-used off-the-shelf acoustic model to infer the phoneme transcription, demonstrating the sensitivity of the acoustic model on speech from speakers of different linguistic backgrounds. The second task is the articulatory contours segmentation: we fine-tune SAM2 [8] on the Annot-16 database and show improved accuracy in segmentation and potential application in annotation tool development. Lastly, we propose classifying phonemes based on articulator contour input, a task bridging the physical production of speech and its abstract linguistic representations.

2. 75-Speaker Annot-16 Database

The 75-Speaker Annot-16 benchmark database is created from the existing public 75-speaker speech MRI database [1]. We would note that the public 75-speaker data adopts the CC BY 4.0 license, allowing us to build upon the material for any purpose. This larger source corpus includes rt-MRI data from 49 native and 26 non-native American-English speakers, with each participant contributing approximately 17 minutes of recorded speech. The speech materials include read speech, consisting of consonant-vowel sequences, phonetically balanced sentences and passages, as well as spontaneous speech elicited by picture description and topic discussion tasks. Audio and rt-MRI data are collected simultaneously. RT-MRI is performed in a 6 mm slice in the mid-sagittal plane of the vocal tract, parameterized as TR = 6.04 ms to capture real-time articulator movements. The image sequences are reconstructed at a high frame rate of 83 Hz [1].

2.1. Subject Selection

To develop the Annot-16 database, we selected 16 representative speakers from the public 75-Speaker dataset for detailed annotation. The selection of 16 speakers is stratified mainly based on their first language (L1): 8 speakers are native speakers of American English, while the other 8 speakers speak languages of India (including in some cases Indian English) as their first

*These authors contributed equally to this work.

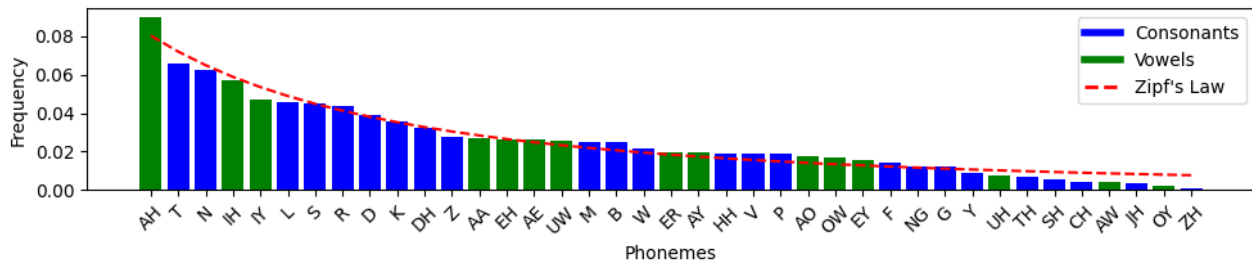


Figure 1: Aggregated phoneme frequency distribution in the 75-Speaker Annot-16 Database.

language. The 16 selected speakers are gender-balanced (7F, 9M) and their age distribution (27.8 ± 7.4) closely matches the distribution of the 75-speaker database (26.3 ± 7.2).

2.2. Phonetic Alignment

We first create transcriptions to generate the phoneme- and word-level phonetic transcriptions of the audio recording. More specifically for spontaneous speech, we use the Microsoft Azure Speech-to-Text model to obtain text transcripts. Then, phonetic alignment is applied by the Montreal Forced Aligner [9] using the English (US) ARPA dictionary. Two phoneticians manually inspect the phonetic alignments and correct alignment errors.

The phoneme distribution of the Annot-16 is presented in Figure 1. Overall, the phoneme distribution pattern is aligned with the estimated phoneme distribution of standard English [10]. As reported in [10], the fitted Zipf’s law line does not perfectly describe the distribution frequency of phonemes. The high- and low-frequency phonemes are overestimated while the mid-frequency phonemes are underestimated. The phonemes AH, T, N, and IH occur most frequently, whereas phonemes like AW, JH, OY, and ZH are less common.

2.3. Semi-Automatic Articulator Contour Segmentation

We apply an open-source semi-automatic MRI segmentation algorithm [7] to extract articulator contours, which is a robust and stable algorithm on temporal data, while offering flexibility to inject expert knowledge. This method, based on an unsupervised segmentation approach, transforms magnetic resonance image acquisitions and geometrical object model vertices into the Fourier domain. It then predicts object vertices in frequency

space by minimizing data differences between the resonance image and object vertices. Further details are available in [7].

As the algorithm requires some manually drawn reference templates for automatically tracking articulator contours, we create a custom MATLAB GUI interface for making templates (Figure 2, left panel). The GUI allows the user to navigate through the rt-MRI videos and draw or edit the contours of each anatomical structure to generate reference templates from the selected frames. First, multiple candidate frames during the speech are selected as the reference from the rt-MRI videos of each participant. Then, four phoneticians divide the task and use the GUI to create several initial reference templates (e.g., 3-8 for each speaker) for these reference frames. The contour annotations are arranged as vertices along the boundary of anatomical structures defined as three large polygonal regions of tissue (R1, R2 and R3, Figure 2, right panel). The annotated anatomical structures are the epiglottis, tongue, lower teeth, lower lip, chin and neck in R1, arytenoid, pharyngeal wall, back and trachea in R2, and hard plate, velum, nasal cavity, nose and upper lip in R3. Based on the reference templates, the algorithm automatically tracks the contours of the anatomical structures for each frame of a video using a hierarchically optimized gradient descent procedure. Depending on the quality of the segmentation results, the algorithm may be rerun several times after adding more templates.

2.4. Handmade Articulator Contour Ground Truths

In addition to the articulator contours generated by the semi-automatic segmentation algorithm, we create 160 frames of

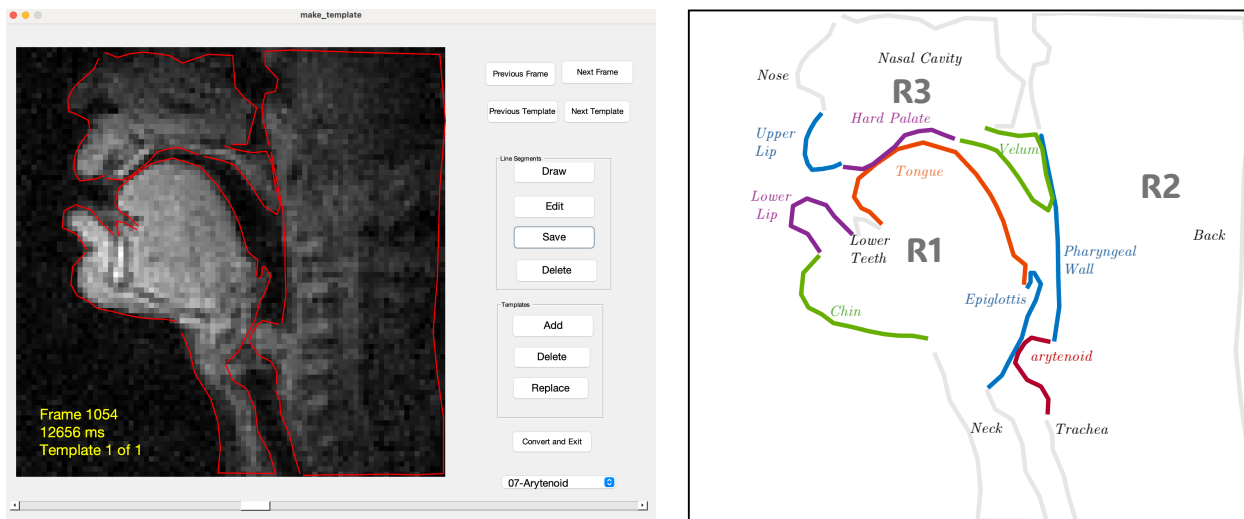


Figure 2: Illustration of articulator annotations. Left: MATLAB GUI for creating reference templates for the semi-automatic segmentation algorithm described in § 2.3. An example reference template is made for frame 1054 of the selected video; Right: Anatomical structure contours with labels of the reference template. The colored lines illustrate the key articulator contours included for creating high-quality handmade ground truths described in § 2.4.

Table 1: Hausdorff distance between the semi-automatically generated articulator contours and high-quality handmade ground truths.

Articulator	Upper Lip	Lower Lip	Hard Palate	Tongue	Velum	Pharyngeal	Epiglottis	Arytenoid
Distance	1.913	1.868	1.664	3.576	2.436	3.340	3.985	4.676

high-quality handmade annotations of articulator contours. These ground truths are used to evaluate the accuracy of the semi-automatic contour extraction method in § 2.3.

The validation experiment is performed on a small representative subset of the rt-MRI data, consisting of speech frames from a spontaneous speech task across all 16 speakers. In total, 160 frames are selected, with 10 frames chosen per speaker. For these ground truths, as illustrated by the colored contours in Figure 2 right panel, only 9 key articulators are selected for the annotation, i.e., the epiglottis, tongue, lower lip, chin, arytenoid, pharyngeal wall, hard palate, velum, and upper lip. One trained phonetician creates initial annotations for these frames by carefully verifying the articulator contours generated by the segmentation algorithm described in § 2.3. Then, several other experts review the initial curation and provide comments and edits for discussion. In cases of discrepancies regarding boundary placement, they finalize the modifications after discussing the results until a consensus is reached.

Table 1 demonstrates the differences in pixel, represented by Hausdorff distance [11], between the automatically generated contours by the segmentation algorithm and the high-quality handmade ground truths. The tongue, epiglottis, and arytenoid show the largest differences, possibly due to their high deformability in images. More specifically for laryngeal structures, the larger Hausdorff distance might be caused by the sub-optimal image quality in the larynx region. In many images, the intricate laryngeal structures appear muddled, presumably because the imaging protocol does not focus on the larynx, and the midsagittal slice selection in data collection is less optimal for this region.

3. Benchmark Baselines

In this section, we introduce three engineering applications, aiming to provide illustrative tasks to inspire future research within the community.

3.1. Speech Phoneme Recognition Baseline

In this baseline, we predict phoneme sequences from the speech recorded in an MRI scanner i.e., phoneme recognition. We employ a well-established open-source acoustic model, a fine-tuned wav2vec 2.0 [12]¹, to infer phoneme sequences in a zero-shot manner. To mitigate accuracy degradation caused by MRI scanner noise, we use a denoiser [13] as a front end to enhance the speech input. We evaluate phoneme recognition performance by computing the phoneme error rate (PER) on the entire dataset as well as on subsets categorized by specific demographic characteristics and linguistic backgrounds.

Table 2: Phoneme error rate (PER) from zero-shot inference.

	Overall	Gender		First Language	
		Female	Male	L1-Ame	L1-Ind
PER	0.323	0.324	0.321	0.288	0.336

As shown in Table 2, first language has a significant impact on phoneme recognition error rates, despite the acoustic model

¹In this study, we use the <https://huggingface.co/facebook/wav2vec2-xl-sr-53-espeak-cv-ft> from Hugging Face.

being fine-tuned on a large multilingual corpus. In contrast, the model performs similarly across different genders. The overall PER of 0.323 highlights the challenges of MRI speech recognition, compared with PER as 0.113 from a same architecture model [14] on the TIMIT corpus [15]. This may be partially due to the suboptimal acoustic conditions in the noisy MRI scanner.

3.2. Articulator Contour Segmentation Baseline

Manual segmentation of vocal tract articulators faces many challenges, including annotator subjectivity, noise artifacts, and inefficient annotation tools. Recent work has been successful using fully automatic segmentation models trained from scratch (e.g., [16, 17]), but these models are limited by the low quantity of publicly-available labeled vocal tract rt-MRI data. One solution to this problem is to fine-tune a general-use image segmentation model for the MRI modality. Using the 75-Speaker Annot-16 database, we create SPAM, a fine-tuned version of Meta’s SAM2 [8] for Speech Production research.

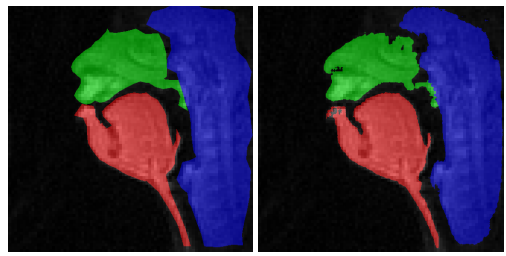
Segmentation Task We define the segmentation task as identifying the three regions of vocal articulators on the 2D rt-MRI sagittal plane stated in § 2.3. SPAM is both trained and evaluated on Intersection over Union (IoU) loss [18], which directly measures the overlap between a predicted and ground truth mask. We reserve 4 subjects (male native American English, female native American English, male native Indian English, female native Indian English) to be used as validation and test subjects and train on annotations of the remaining 12 subjects. For each subject, we select MRI videos of 22 unique speech tasks and select a single random frame per training batch for a total of 264 frames per batch.

Model We upscale each MRI image to 512 by 512 pixels via bilinear interpolation. SAM2 requires an input point for each desired segmentation, which we randomly select over all pixels in the ground truth mask. During training, we fine-tune SPAM on Intersection over Union (IoU) loss between the predicted and annotation masks for each of the three regions. We repeat this process 50 times until we have trained on 50 frames per video for a total of 13,200 total frames trained.

Evaluation We validate and test SPAM in two ways – the “Seen Subjects” setup contains 40 frames per video (stratified equally between validation and test sets) from the 12 training subjects on MRI videos of 3 unseen speech tasks; the “Unseen Subjects” setup contains 40 frames per video from the 4 held-out subjects on these 3 unseen tasks as well as on 3 speech tasks seen during training. We train and evaluate SPAM on three SAM2 model sizes: small, base, and large.

Results We present the IoU performance of SPAM compared to off-the-shelf SAM2 in Table 3. SPAM dramatically increases segmentation performance compared to off-the-shelf SAM2, which we attribute to the lack of medical imagery in SAM2’s training dataset. We also see that performance drops about 10 points between seen and unseen subjects for the back (R2) and upper (R3) regions. These regions tend to be static for a single subject, so SPAM may be partially overfitting these regions during training. Interestingly, we see little effect of model size on performance. We hypothesize that because the original MRI image is so small (82 by 82 pixels before upscaling), the modeling capabilities of the tiny model are sufficient for this task.

Table 3: (a) Segmentation example of an unseen speaker using 75-Speaker Annot-16 annotations. (b) Predicted segmentation example by SPAM. (c) Intersection over Union (IoU) segmentation scores of off-the-shelf SAM2 compared with SPAM for segmenting the mid-sagittal image into R1, R2, and R3 regions.



		Test IoU (Subjects Seen in Training)				Test IoU (Unseen Subjects)			
		R1	R2	R3	All	R1	R2	R3	All
SAM2	Small	35.55	1.19	6.06	14.27	24.87	2.78	4.27	10.64
	Base	15.67	25.99	7.98	16.55	14.41	29.35	8.39	17.38
	Large	32.40	1.83	5.19	13.14	24.11	2.77	4.30	10.39
SPAM	Small	93.22	94.61	89.91	92.58	91.60	80.14	80.44	84.06
	Base	93.43	94.79	90.11	92.78	92.04	79.78	81.95	84.59
	Large	93.09	94.51	89.6	92.40	91.31	79.57	79.98	83.62

(a) Annotation.

(b) SPAM Segmentation.

(c) Comparison of off-the-shelf SAM2 and SPAM Segmentation Performance.

Table 4: (a) Phoneme classification using the articulatory contours, evaluated under unseen speech and subject. (b) The confusion matrix for consonant classification.

	Model	Unseen Speech		Unseen Subject	
		Macro-F1	Top-3 Acc	Macro-F1	Top-3 Acc
Phonemes	RNN	0.339	63.55	0.219	47.11
	S4	0.443	73.09	0.253	51.04
Consonants	RNN	0.306	69.39	0.217	47.82
	S4	0.414	78.26	0.274	63.90

(a) Phoneme classification results using articulator contours.

	Stop						Fricative						Liquid			
	Bilabial		Alveolar		Velar		Labiodental		Dental		Alveolar		Palatal		Glottal	Palatal
	P	B	T	D	K	G	F	V	TH	DH	S	Z	SH	ZH	HH	R
P	49.3	22.3	1.4	2.7	0.7	0.0	4.1	10.8	2.7	0.0	1.4	2.7	0.0	0.0	0.0	2.0
B	26.1	51.0	0.0	2.7	0.4	0.4	4.2	6.5	4.6	0.4	0.4	0.8	0.0	0.0	1.2	1.5
T	3.3	3.6	16.4	16.9	3.3	3.6	3.4	5.2	3.1	4.2	7.7	10.2	4.2	3.3	5.2	4.5
D	1.6	7.9	7.1	26.9	13.4	6.3	7.9	11.1	1.6	1.2	2.0	1.6	4.4	1.6	4.4	1.2
K	0.0	1.3	2.3	1.0	43.8	30.9	4.5	0.5	0.3	2.5	4.8	2.0	1.8	0.5	1.8	2.8
G	0.0	0.0	1.0	0.0	14.6	69.8	1.0	0.0	0.0	1.0	0.0	0.0	0.0	2.1	9.4	1.0
F	7.3	6.8	0.0	2.1	0.0	0.0	51.0	19.3	2.1	1.6	0.0	1.0	0.0	1.6	0.5	6.8
V	4.9	4.1	0.0	2.0	0.0	0.4	18.6	56.3	3.2	1.2	0.0	1.2	0.0	0.4	0.4	7.3
TH	0.0	2.9	2.9	12.9	0.0	0.0	4.3	0.0	28.6	24.3	1.4	18.6	1.4	0.0	0.0	2.9
DH	0.0	2.9	0.0	2.2	0.0	0.7	2.9	4.4	10.1	63.0	2.2	5.8	0.0	0.0	5.1	0.7
S	1.7	0.3	2.7	6.6	3.1	0.0	1.0	0.9	6.8	3.9	25.6	44.8	1.4	0.7	0.3	0.3
Z	1.7	0.6	2.6	5.4	0.9	0.2	0.4	1.1	6.0	3.7	20.0	50.9	4.7	1.5	0.2	0.2
SH	0.0	0.0	2.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.4	0.0	73.8	21.4	0.0	0.0
ZH	3.1	0.0	12.5	6.3	0.0	0.0	0.0	6.3	0.0	0.0	3.1	0.0	6.3	62.5	0.0	0.0
HH	0.8	0.8	0.8	8.0	0.6	0.6	1.7	0.3	0.6	13.3	0.0	0.3	1.4	0.3	69.3	1.4
R	5.5	1.9	3.6	7.4	2.2	6.0	8.1	6.2	0.5	0.2	0.2	1.7	1.7	1.7	1.2	51.9

(b) The confusion matrix for consonant classification (unseen speech).

3.3. Articulator-based Phoneme Classification

Model and Evaluation In addition to recognizing phonemes from audio, we undertake to classify phonemes using annotated articulator contours. Similar to the segmentation experiments, the unseen subject evaluation involves training and testing phoneme classification on different subjects. Moreover, we perform an unseen speech evaluation, where training and testing utilize recordings from different elicitation conditions. Specifically, we select two recordings from the read speech (one from the VCV passage and another from the Grandfather passage) and one recording from the spontaneous speech (discussion about topics) as the test set. We compare two model architectures for phoneme classification using the articulator contours: Recurrent Neural Networks (RNNs) and the Structured State Space Sequence Model (S4) [19]. We apply a hidden size of 128 in both RNNs and S4. We perform phoneme classification using both the complete set of phonemes and a subset of selected consonants, including stops, fricatives, and liquids.

Results The phoneme classification results under unseen speech and subjects are presented in Table 4, evaluated using Macro-

F1 and Top-3 Accuracy. Overall, the S4 model consistently outperforms RNNs in phoneme classification across all conditions. Moreover, classifying phonemes from unseen subjects is more challenging than from unseen speech when using articulatory contours. Nonetheless, the S4 model performs well in both phoneme and consonant classification under unseen speech. To further analyze the trained S4 phoneme classifier, we visualize the confusion matrix for consonant classification under the unseen speech condition. The plot indicates that the classifier struggles to distinguish consonants with the same place of articulation but differ only in voicing. For example, the model often confuses /p/ with /b/ (bilabial), /k/ with /g/ (velar), /f/ with /v/ (labiodental) and /s/ with /z/ (alveolar). Again, this might be due to the sub-optimal image quality in the larynx region. Moreover, speakers may differ as to the exact placement of the MRI slice with respect to true mid-sagittal. Informal observation suggests that in a true mid-sagittal slice, the vocal folds can partially disappear as they are abducted for devoicing. These findings also highlight the importance of cross-modal learning that combines audio signals with articulator movement to better understand the dynamics of speech articulation.

4. Research and Technical Applications

The Annot-16 offers several possibilities for both speech technology and linguistic-phonetic research. From a speech technology perspective, it can be employed to enrich speech recognition and synthesis systems with articulatory representations. This can further advance data-driven speech technology by offering insights into speech production mechanisms and improving model interpretability and generalization ability. For instance, new techniques to model articulatory representations, such as articulatory gestures, can be developed [20]. Researchers can also leverage articulatory knowledge to develop customized speech synthesis systems in low-resource scenarios [21]. For speech/speaker recognition, articulatory information can be added to improve performance [22, 23, 24, 25].

For linguistic-phonetic research, the database can be used for corpus-based articulatory analyses, such as studies on the articulatory basis of accents [26, 27], and the spatiotemporal properties of specific articulatory gestures, whose presence can be indexed by the associated phonetic alignments [28, 29]. For instance, a possible study using this database is to use the provided articulatory contours to obtain time series of vocal tract constrictions associated with particular gestures. Spatiotemporal landmarks can then be detected from these time series in order to analyze gestural dynamics. The phonetic alignments can be used to find intervals representing the linguistic variables of interest, such as specific phonological units, accents, and prosodic contexts.

5. Acknowledgements

This work was supported by The U.S. National Science Foundation under Grant NSF (IIS-2311676, BCS-2240349, RI-2106930). The authors alone are responsible for the content and conclusions.

6. References

- [1] Y. Lim, A. Toutios, Y. Bliesener, Y. Tian, S. G. Lingala, C. Vaz, T. Sorensen, M. Oh, S. Harper, W. Chen *et al.*, “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, 2021.
- [2] J. Kim, A. Toutios, Y.-C. Kim, Y. Zhu, S. Lee, and S. Narayanan, “Usc-emo-mri corpus: An emotional speech production database recorded by real-time magnetic resonance imaging,” in *International Seminar on Speech Production (ISSP), Cologne, Germany*, vol. 226, 2014.
- [3] T. Sorensen, Z. I. Skordilis, A. Toutios, Y.-C. Kim, Y. Zhu, J. Kim, A. C. Lammert, V. Ramanarayanan, L. Goldstein, D. Byrd *et al.*, “Database of volumetric and real-time vocal tract mri for speech science,” in *Interspeech*, 2017, pp. 645–649.
- [4] K. Isaieva, Y. Laprie, J. Leclère, I. K. Douros, J. Felblinger, and P.-A. Vuissoz, “Multimodal dataset of real-time 2d and static 3d mri of healthy french speakers,” *Scientific Data*, vol. 8, no. 1, p. 258, 2021.
- [5] K. Maekawa, “Real-time mri articulatory movement database and its application to articulatory phonetics,” *Acoustical Science and Technology*, vol. 46, no. 1, pp. 45–54, 2025.
- [6] —, “Introduction to the real-time articulatory movement database-version 2,” 2024. [Online]. Available: https://rtmridb.ninjal.ac.jp/Introduction-to-the-rtMRIDB_v2-en.pdf
- [7] E. Bresch and S. Narayanan, “Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images,” *IEEE transactions on medical imaging*, vol. 28, no. 3, pp. 323–338, 2008.
- [8] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, “Sam 2: Segment anything in images and videos,” 2024. [Online]. Available: <https://arxiv.org/abs/2408.00714>
- [9] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi,” in *Proc. Interspeech 2017*, 2017, pp. 498–502.
- [10] A. C. Lammert, J. Melot, D. E. Sturim, D. J. Hannon, R. DeLaura, J. R. Williamson, G. Ciccarelli, and T. F. Quatieri, “Analysis of phonetic balance in standard english passages,” *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 4, pp. 917–930, 2020.
- [11] F. Hausdorff, *Grundzuge der mengenlehre*. American Mathematical Soc., 1978, vol. 61.
- [12] Q. Xu, A. Baevski, and M. Auli, “Simple and effective zero-shot cross-lingual phoneme recognition,” in *Interspeech 2020*, 2020.
- [13] A. Défossez, G. Synnaeve, and Y. Adi, “Real time speech enhancement in the waveform domain,” in *Interspeech 2020*, 2020, pp. 3291–3295.
- [14] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, “Unsupervised speech recognition,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 27 826–27 839, 2021.
- [15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, D. S. Pallett, N. L. Dahlgren, V. Zue, and J. G. Fiscus, “Timit acoustic-phonetic continuous speech corpus,” *Philadelphia: Linguistic Data Consortium*, 1993.
- [16] S. Erattakulangara, K. Kelat, K. Burnham, R. Balbi, S. E. Gerard, D. Meyer, and S. G. Lingala, “Open-source manually annotated vocal tract database for automatic segmentation from 3d mri using deep learning: Benchmarking 2d and 3d convolutional and transformer networks,” 2025. [Online]. Available: <https://arxiv.org/abs/2501.06229>
- [17] M. Ruthven, A. M. Peplinski, D. M. Adams, A. P. King, and M. E. Miquel, “Real-time speech mri datasets with corresponding articulator ground-truth segmentations,” *Scientific Data*, vol. 10, no. 1, p. 860, Dec 2023. [Online]. Available: <https://doi.org/10.1038/s41597-023-02766-z>
- [18] M. A. Rahman and Y. Wang, “Optimizing intersection-over-union in deep neural networks for image segmentation,” in *Advances in Visual Computing*, G. Bebis, R. Boyle, B. Parvin, D. Koracin, F. Porikli, S. Skaff, A. Entezari, J. Min, D. Iwai, A. Sadagic, C. Scheidegger, and T. Isenberg, Eds. Cham: Springer International Publishing, 2016, pp. 234–244.
- [19] A. Gu, K. Goel, and C. Ré, “Efficiently modeling long sequences with structured state spaces,” *arXiv preprint arXiv:2111.00396*, 2021.
- [20] J. Lian, A. W. Black, Y. Lu, L. Goldstein, S. Watanabe, and G. K. Anumanchipalli, “Articulatory representation learning via joint factor analysis and neural matrix factorization,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [21] P. Wu, T. Li, Y. Lu, Y. Zhang, J. Lian, A. W. Black, L. Goldstein, S. Watanabe, and G. K. Anumanchipalli, “Deep speech synthesis from mri-based articulatory representations,” *arXiv preprint arXiv:2307.02471*, 2023.
- [22] P. K. Ghosh and S. S. Narayanan, “Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion,” *J. Acoust. Soc. Am. Express Letters*, vol. 130, no. 4, pp. EL251–EL257, sep 2011.
- [23] M. Li, J. Kim, A. Lammert, P. K. Ghosh, V. Ramanarayanan, and S. S. Narayanan, “Speaker verification based on the fusion of speech acoustics and inverted articulatory signals,” *Computer, Speech, and Language*, vol. 36, pp. 196–211, mar 2016.
- [24] 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, 2010.
- [25] A. Wrench and K. Richmond, “Continuous speech recognition using articulatory data,” in *Sixth International Conference on Spoken Language Processing (ICSLP 2000)*. International Speech Communication Association, 2000, pp. 145–148.
- [26] X. Shi, T. Feng, K. Huang, S. R. Kadiri, J. Lee, Y. Lu, Y. Zhang, L. Goldstein, and S. Narayanan, “Direct articulatory observation reveals phoneme recognition performance characteristics of a self-supervised speech model,” *JASA Express Letters*, vol. 4, no. 11, 2024.
- [27] K. Huang, J. Goldberg, L. Goldstein, and S. Narayanan, “Analysis of articulatory setting for l1 and l2 english speakers using mri data,” in *Interspeech 2024*, 2024, pp. 1020–1024.
- [28] Y. Zhang and L. Goldstein, “Stop voicing and devoicing as articulatory tasks: A cross-linguistic rt-mri study,” in *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS)*, 2023, pp. 1011–15.
- [29] Y. Lu, H. Hsu, L. Goldstein, and A. Toutios, “Effect of individual vocal tract geometry on the tongue shaping for american english /t/,” in *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS)*, 2023, pp. 1042–1046.