单发言人长格式实时 MRI 语音数据集和基准测试

Sean Foley[1,2], Jihwan Lee[1], Kevin Huang[1], Xuan Shi[1],
Yoonjeong Lee[1], Louis Goldstein[2], Shrikanth Narayanan[1,2]

[1]Signal Analysis and Interpretation Lab, University of Southern California
[2]Department of Linguistics, University of Southern California

ABSTRACT

我们发布了 USC 长单发音人 (LSS) 数据集，其中包含了发声过程中实时 MRI 视频的声道动态以及同时获取的声音数据。这一独特的数据集包含大约一小时来自单一美式英语母语者的视频和音频数据，使其成为公开可用的实时 MRI 语音数据中较长的单发音人数据集之一。除了发音和声学原始数据外，我们还发布了适用于多种下游任务的数据派生表示形式。这包括裁剪至声道区域的视频、按句子级别划分的数据切分、恢复和降噪后的音频以及感兴趣区域的时间序列。我们还在发音合成和音素识别任务上对该数据集进行了基准测试，提供了未来研究可以在此基础上改进的任务基线性能。数据集网站：https://sail.usc.edu/span/single_spk

Index Terms— 实时 MRI，语音产生，数据集，基准测试

1. 介绍

语音生成需要在声道中形成狭窄区域，通过多个发音器官协同工作以高效和有效地实现这些狭窄区域 [1, 2]。在语音生成过程中使用的至关重要的主动发音器官至少包括嘴唇、下颌、舌头、软腭和喉部。虽然存在多种方法来捕捉语音生成过程中的各种发音器官，但实时 (rt) MRI 提供了对声道最深入的视角，提供了一个完整的中矢状面视图，包括通常难以非侵入性成像的喉、咽和软腭。当然，rtMRI 也带来了昂贵的获取、运行和维护成本的问题，使得公开发布的数据集对语音科学和技术社区极为有益。

已经有一些先前的 rtMRI 语音数据集被公开发布。USC-TIMIT 数据集 [3] 包含来自 10 名说话者的 rtMRI 视频和音频，其中包括 5 名男性和 5 名女性，产生 460 句语音丰富的句子，以 23 帧每秒的速度捕获。对于每位说话者，大约有 37 分钟的语音。USC 75-Speaker 数据集 [4] 包含来自 75 名说话者的读出和自发性语音的混合内容，其中包括美式英语、印度英语以及中文等其他语言的母语使用者。每位说话者的大约有 17 分钟的语音，MRI 视频以 86 帧每秒的速度重建。这里发布的数据集在这些早期的数据集基础上进行了扩展，重点是为单个说话者捕捉更多数据。

使用 rtMRI 收集的语音数据已被应用于一系列语音处理任务和音系分析，如发音合成和语音逆向工程。虽然电磁发音图描记术和超声波发音数据也用于这些任务中，但它们通常需要额外的声学特征来捕捉鼻音和声音 [5]。基于 rtMRI 的方法可以直接实现类似的效果，无需任何附加内容 [6]。从发音数据进行自我监督表示学习 [7] 已应用于脑机接口 [8] 和自动化发音评估 [9]，其中 rtMRI 数据集已被部署于这些发展中 [7]。最后，从实时磁共振成像及其对应的音频中进行音素识别已经允许了多模态建模和表征学习 [10, 11]。

我们发布了 USC 单说话人长语音 (LSS) 数据集，其中包含大约一小时的单一美式英语母语者的 rtMRI 语音数据。这提供了比之前发布的数据集多得多的单说话人语音数据，特别是自发性语音数据。因此，我们预计该数据集可以帮助推进诸如发音合成和语音逆向等难以在多个说话人间进行的任务的模型开发。

数据集	读（分钟）	自发。(分钟)	每秒
USC TIMIT [3]	∼ 37	0	2
USC 75 Speaker [4]	∼ 12	∼ 5	8
USC LSS	37	17	9

Table 1. 阅读和自发言语中最长单一发言时长以及先前数据集与当前数据集的 FPS 对比。

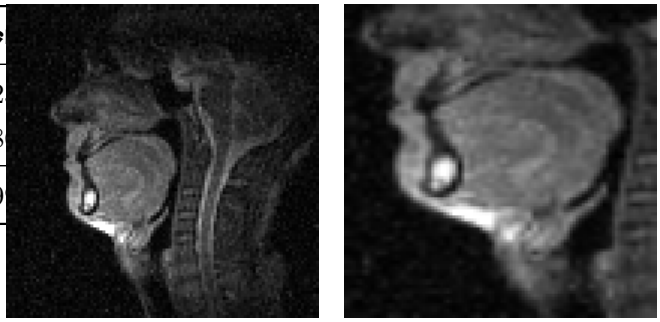

Fig. 1. 原始视频的示例帧（左）和裁剪到发声器官区域的视频（右）

2. USC 单说话人长语音（LSS）数据集

2.1. 概览

USC LSS 数据集包含使用 rtMRI 采集的单个说话者近一小时的语音。据我们所知，这使得该数据集成为目前公开可用的最长的单说话者 rtMRI 数据集。该数据集包含语音产生过程中声腔的 rtMRI 视频和同步音频。此外，虽然以前的数据集通常只发布原始音频和视频，但我们还包括从原始数据派生的一些可能更适合各种研究形式的表示，包括裁剪到声腔区域的视频、两种处理过的音频——去噪和恢复、兴趣区域时间序列以及按句子级别划分的数据。

2.2. 数据采集

语料库包含一位 32 岁的美国英语男性母语者的演讲数据，其中包括朗读和自发性讲话的组合。朗读部分包括 USC TIMIT 语料库中使用的 460 个句子以及在 USC 75-Speaker 数据集中使用过的祖父、彩虹和北风段落各两次重复。自发性讲话部分则包含了对五个图片描述和关于食物、旅行、音乐及电影话题的提示各两次重复。说话人的声道通过一台 0.55T MRI 扫描仪用定制上呼吸道接收线圈 [12] 在中矢状面上进行了成像。获取的数据以每秒 99 帧（FPS）的速度重建，请参见 [13] 了解详情。音频以 16 kHz 的采样率进行采集。总共收集了 71 次扫描中的 54 分钟数据，其中包括 37 分钟的朗读讲话和 17 分钟的自发性讲话（请参见表 1 对比先前的数据集）。初步音素对齐是使用蒙特利尔强制对齐器（MFA）[14] 提取，并由语音学家在 Praat 中进行了手动校正。除了原始视频外，我们还发布了裁剪到声道区域的视频，上颚硬骨、喉部和咽壁作为边界（请参见图 1），使得视频帧内仅包含与讲话相关的关键信息。

2.3. 恢复与去噪音频

除了原始音频外，我们还提供了一个最近的语音恢复模型 Miipher [15][1] 恢复的版本，该模型减少了背景 MRI 设备噪声。与通常仅使用语音声学作为输入的典型语音降噪模型不同，Miipher 将文本表示和语音声学一起用作输入，最大限度地利用文本信息处理不可听部分，并输出更清晰的输入语音版本。我们建议在使用此版本时要谨慎，因为某些原本不可听的部分恢复后可能不匹配相应的发音运动学。只有当不需要严格对应于语音声学与发音运动学之间的精确对应关系时，这个恢复版本才可能是有用的。我们还包括了使用降噪器模型 [16] 处理后的音频。

2.4. 句子级拆分

除了完整的原始文件外，我们还包括句子级别的视频和音频文件，包括原始的、去噪的以及恢复后的音频。通过将原始刺激作为脚本语音的基础，并使用 Whisper-large[17] ASR 作为自发语音的基础，半自动地创建了音频转录文本。这些基础线根据音频手动进行了调整。从这些转录中，我们使用一个定制脚本来按照标点符号分割音频和视频文件，总计得到 684 个句子。我们将这些句子进一步划分为训练、验证和测试集（比例为 0.85/0.05/0.1），以用于任何下游任务。

[1] https://github.com/Wataru-Nakata/miipher

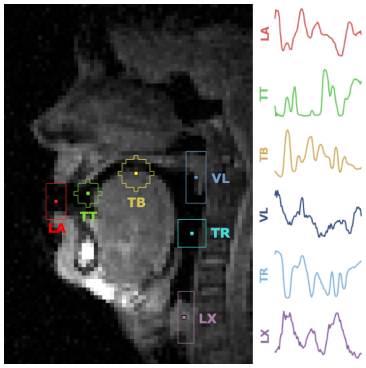

Fig. 2. 示例帧带有标记的兴趣区域（ROI）（左）及其对应的时间序列（右）。

2.5. 发音区域兴趣区（ROI）时间序列

如图 2 所示，选择了六个可解释的代表性感兴趣区域（ROI）来自发音道，这些区域最小限度地代表了英语中的语音，如下所示 [2]：唇开口度（LA）、舌尖（TT）、舌身（TB）、软腭（VL）、舌根（TR）和喉部（LX）。每个区域由语音研究人员手动标注以最好地捕捉该区域内每部分的运动，使用 VocalTract ROI 工具箱 [2]。对于每个 ROI，计算像素强度作为狭窄程度的代理度量。我们包括了 ROI 的时间序列数据供研究人员探索这个低维空间中的语音处理任务。

3. 基准基线

我们的目标不是优化任务性能，而是建立可重现的基线，供未来的工作在此基础上进行。这些结果突显了在使用 USC LSS 数据集训练时，常见的架构能够“开箱即用”实现的效果。

3.1. 发音合成

模型．类似于先前的工作 [6, 18]，我们使用神经声码器从发音特征合成波形。具体来说，我们使用由一个生成器和两个判别网络（多周期判别器 (MPD) 和多尺度判别器 (MSD)）组成的 HiFi-GAN 模型 [19]。对于我们的实验，我们使用了一个预训练的生成器、MPD 和 MSD，这些是在 VCTK [20]，Librispeech [21] 和 LJSpeech [22] 数据集上训练得到的。我们将生成器的四个模块中的最后一个以及 MPD 和 MSD 都冻结起来，以防止判别器过于强大而压制住生成器。

[2]https://github.com/reedblaylock/VocalTract-ROI-Toolbox

音频	CER (%) ↓	**误码率**（%）↓	**最大公约数** ↓
Denoised	31.3 ± 6.0	52.6 ± 10.1	4.3 ± 0.1
Restored	24.3 ± 4.6	45.0 ± 8.2	4.8 ± 0.2

Table 2. 所有当前研究中训练的模型的语音合成质量结果，以句子平均值和 95%置信区间表示。

预处理．我们尝试使用与第 2.3 节中描述的恢复音频相比去噪后的音频进行实验。所有音频的采样率为 16 kHz。视频帧被评分并调整为 z，尺寸改为 128×128，这比原始帧大小表现更好。视频重塑为 $[H \times W, t]$，其中 t 是帧数，并以灰度加载以保持生成器中使用的 1D 卷积。

实现。生成器的输入大小更改为 128×128，而所有其他维度保持不变。我们使用了跳数为 162、上采样率为 [6, 3, 3, 3] 以及四个块中的上采样核为 [12, 6, 6, 6]。训练期间使用了批量大小为 2 和学习率为 1e-3。其他超参数与 V1 预训练的 HiFi-GAN 相同。我们使用句子级别的分割来划分训练集、验证集和测试集，并报告保留的测试集的结果。

结果。音韵合成结果见表 2。有趣的是，基于恢复语音训练的模型在字符错误率 (CER) 和词错误率 (WER) 上优于基于去噪语音训练的模型，但在梅尔谱系数距离 (MCD) 上后者表现更好。这可能归因于恢复语音本身的某些方面，而不是实际的模型性能。总体而言，恢复音频模型的主观表现优于之前尝试直接将 rtMRI 音韵特征转换为语音 [6, 18] 的研究，而去噪音频模型与这些先前的工作相当。

模型在恢复语音上训练后预测的示例可以在图 3 中看到。与地面真实频谱图（底部）相比，预测的频谱图（顶部）明显缺乏自然语音特有的精细共振结构。虽然可以观察到预测语音中的一般音节和共振结构，但它表现出相当宽厚的带状，近似于共鸣峰结构。这仍然使语音听起来是可理解的，但使其显得非常不自然。

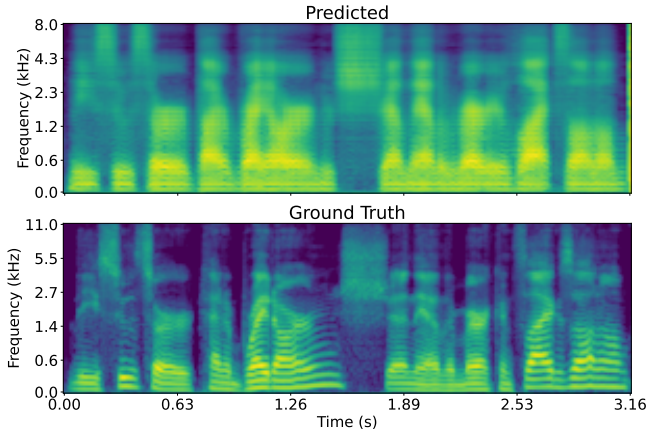

Fig. 3. 示例预测（顶部）和地面真实（底部）从恢复的语音模型中得到的频谱图。

3.2. 音素识别

我们遵循在 [11] 中报告的一般设置，目标是比较输入模态不同下的音素识别性能差异，即比较单一模式的音频和视频模型与多模态模型的差异，请参阅 [11] 以获取更多详细信息。

模型 。主要架构是 Conformer[23]。Conformer 的输出由单个 LSTM 层解码，使用最终线性层进行预测。输入到 Conformer 中的单一模式数据包括声学特征或视频特征。对于多模态模型，音频和视频特征沿时间维度连接在一起。所有模型均使用连接主义时序分类（CTC）损失 [24] 进行训练。作为基准线，我们通过 Hugging Face 使用 Wav2Vec2Phoneme 对音频进行了零样本音素识别。

预处理 . We use the sentence-level splits described in Section 2.4 and the denoised audio for training. Audio features were extracted from the 9^{th} layer of the pre-trained base WavLM model [25] and video features comprised of the classification (CLS) token for each frame from the last hidden layer of a fine-tuned base ViT model [26].

实现。输入维度被设定为 768 用于单模态模型，而多模态模型则设置为 2*768，并且前馈维度设为 256。注意力头的数量设定为 4，卷积核大小为 31，dropout 设为 0.3。Conformer 层数设定为 3。LSTM 的隐层大小为 128，这作为最终线性层的输入尺寸。对于所有模型，均使用 Adam 优化器，批量大小设为 16，学习率设为 1e-4，并且每 20 个周期衰减 0.9。

模型	**周期** ↓
Wav2Vec2Phoneme	0.39 ± 0.03
Audio	0.22 ± 0.03
Video	0.50 ± 0.02
Multimodal	0.30 ± 0.03

Table 3. 每个发言的平均 PER 及 95%置信区间，当前研究中的模型（底部）与基线模型（顶部）相比。

结果 . 音素错误率（PER）结果见表 3。单模态音频模型表现最佳，优于其他两种模型和基线。仅视频模型表现最差，其 PER 仅为 0.50。有趣的是，多模态模型的表现劣于仅音频模型，这表明在这种情况下，简单的特征拼接并不是跨模态组合的理想方法。我们鼓励未来的研究探索其他多模态学习的方法，如跨模态注意力 [27] 和多模态专家混合 [28]。

图 4 显示了三种模型中每种发音方式和位置类别的 PER。值得注意的是，当 rtMRI 视频中的发音区别细微时（如塞音和摩擦音），或者高度变化时（如元音 [29]），多模态的表现明显比只有音频的情况要差得多。当声学和发音特征都清晰时，例如在 liquids 和 nasals 中，多模态表现最佳。实际上，在 velars 方面，多模态模型在这三种模型中表现最好，因为它们通常与 labials 和 coronals 在发音上很容易区分，并且有明显的共振峰过渡。

4. 研究应用

USC LSS 数据集可用于多种语音处理任务和音系分析。基于 rtMRI 的语音反转和发音综合等任务受到了长时间单个说话人数据可用性的限制 [6]。该数据集有助于开发学习声学与发音之间映射的模型。此外，声道轮廓可以作为许多任务 [30, 6, 31] 的稳健发音表示，但从生成 99 FPS 视频的 0.55T 磁铁中自动提取轮廓的方法尚不完善。该数据集可以帮助进一步推进这些工作。

对于音系学家而言，USC LSS 数据集提供了研究

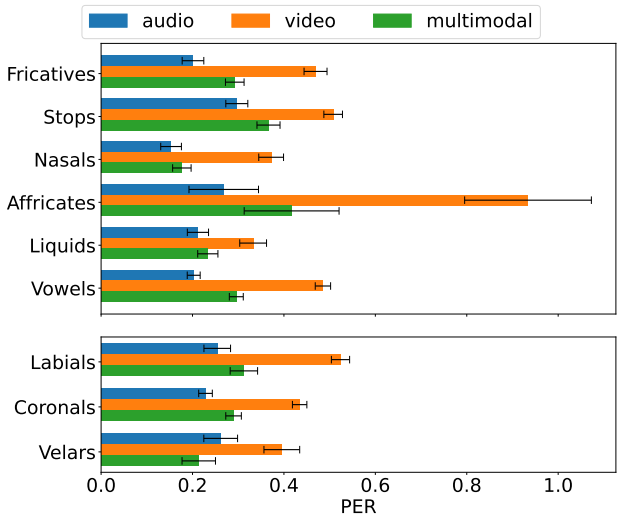

Fig. 4. 按发音类别分组的单模态和多模态模型的 PER 结果，按发音方式（上方）和发音部位（下方）分组。

单个说话人发音的机会，特别是关于变异性与统一性 [32] 的问题。说话人在生成某个特定音段 [33] 时表现出显著的变异性，同时在“重用”某些发音动作 [34, 32] 方面也显示出一定程度的一致性。虽然语料库语音学的发展使得这些领域的研究可以基于数据驱动进行，但这里呈现的大量配对音频和发音数据肯定有助于理解说话人内部的生成机制。

5. 致谢

本工作得到了 NIH 拨款 T32 DC009975 和 NSF 2311676 的支持。

6. REFERENCES

[1] Michael T Turvey, “Preliminaries to a theory of action with reference to vision,” Perceiving, acting and knowing, vol. 2, 1977.

[2] Catherine P Browman and Louis Goldstein, “Articulatory phonology: An overview,” Phonetica, vol. 49, no. 3-4, pp. 155–180, 1992.

[3] Shrikanth 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.

[4] Yongwan Lim 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, pp. 187, 2021.

[5] Cheol Jun Cho, Peter Wu, Tejas S Prabhune, Dhruv Agarwal, and Gopala K Anumanchipalli, “Articulatory encodec: Vocal tract kinematics as a codec for speech,” arXiv preprint arXiv:2406.12998, 2024.

[6] Peter Wu, Tingle Li, Yijing Lu, Yubin Zhang, Jiachen Lian, Alan W Black, Louis Goldstein, Shinji Watanabe, and Gopala K Anumanchipalli, “Deep speech synthesis from mri-based articulatory representations,” arXiv preprint arXiv:2307.02471, 2023.

[7] Jiachen Lian, Alan W Black, Yijing Lu, Louis Goldstein, Shinji Watanabe, and Gopala 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.

[8] Sean Metzger et al., “A high-performance neuroprosthesis for speech decoding and avatar control,” Nature, pp. 1–10, 2023.

[9] Jiachen Lian et al., “Ssdm: Scalable speech dysfluency modeling,” Advances in neural information processing systems, vol. 37, pp. 101818–101855, 2024.

[10] Xuan Shi, Tiantian Feng, Kevin Huang, Sudarsana Reddy Kadiri, Jihwan Lee, Yijing Lu, Yubin Zhang, Louis Goldstein, and Shrikanth Narayanan, “Direct articulatory observation reveals phoneme recognition performance characteristics of a self-supervised speech model,” JASA Express Letters, vol. 4, no. 11, 2024.

[11] Sean Foley et al., “Towards disentangling the contributions of articulation and acoustics in multimodal phoneme recognition,” arXiv preprint arXiv:2505.24059, 2025.

[12] Felix Muñoz, Yongwan Lim, Sophia X Cui, Helmut Stark, and Krishna S Nayak, “Evaluation of a novel 8-channel rx coil for speech production mri at 0.55 t,” Magnetic Resonance Materials in Physics, Biology and Medicine, vol. 36, no. 3, pp. 419–426, 2023.

[13] Prakash Kumar et al., “State-of-the-art speech production mri protocol for new 0.55 tesla scanners,” in Interspeech 2024, 2024, pp. 2590–2594.

[14] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, “Montreal forced aligner: Trainable text-speech alignment using kaldi.,” in Interspeech, 2017, vol. 2017, pp. 498–502.

[15] Yuma et al. Koizumi, “Miipher: A robust speech restoration model integrating self-supervised speech and text representations,” in 2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, 2023, pp. 1–5.

[16] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi, “Real time speech enhancement in the waveform domain,” arXiv preprint arXiv:2006.12847, 2020.

[17] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, “Robust speech recognition via large-scale weak supervision,” in International conference on machine learning. PMLR, 2023, pp. 28492–28518.

[18] Peter Wu, Bohan Yu, Kevin Scheck, Alan W Black, Aditi S Krishnapriyan, Irene Y Chen, Tanja Schultz, Shinji Watanabe, and Gopala K Anumanchipalli, "Deep speech synthesis from multimodal articulatory representations," arXiv preprint arXiv:2412.13387, 2024.

[19] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," Advances in neural information processing systems, vol. 33, pp. 17022–17033, 2020.

[20] Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2012.

[21] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015, pp. 5206–5210.

[22] Keith Ito and Linda Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[23] Anmol Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," arXiv preprint arXiv:2005.08100, 2020.

[24] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in Proceedings of the 23rd International Conference on Machine Learning, New York, NY, USA, 2006, ICML '06, pp. 369–376, Association for Computing Machinery.

[25] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1505–1518, 2022.

[26] Dosovitskiy Alexey, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[27] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh, "Hierarchical question-image co-attention for visual question answering," Advances in neural information processing systems, vol. 29, 2016.

[28] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby, "Multimodal contrastive learning with limoe: the language-image mixture of experts," Advances in Neural Information Processing Systems, vol. 35, pp. 9564–9576, 2022.

[29] Douglas H Whalen, Wei-Rong Chen, Mark K Tiede, and Hosung Nam, "Variability of articulator positions and formants across nine english vowels," Journal of phonetics, vol. 68, pp. 1–14, 2018.

[30] Erik Bresch and Shrikanth 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.

[31] Xuan Shi et al., "75-speaker annot-16: A benchmark dataset for speech articulatory rt-mri annotation with articulator contours and phonetic alignment," in Proc. Interspeech 2025, 2025, pp. 2175–2179.

[32] Eleanor Chodroff and Colin Wilson, "Uniformity in phonetic realization: Evidence from sibilant place of articulation in american english," Language, 2022.

[33] Sarah Harper, "Individual-and group-level associations between articulatory and acoustic variability for english consonants," in ICPhs 2023, 2023.

[34] Matthew Faytak, "Place uniformity and drift in the suzhounese fricative and apical vowels," Linguistics Vanguard, vol. 8, no. s5, pp. 569–581, 2022.