



Seamless Dysfluent Speech Text Alignment for Disordered Speech Analysis

Zongli Ye¹, Jiachen Lian², Xuanru Zhou¹, Jinming Zhang¹, Haodong Li³, Shuhe Li¹, Chenxu Guo¹, Anaisha Das², Peter Park², Zoe Ezzes⁴, Jet Vonk⁴, Brittany Morin⁴, Rian Bogley⁴, Lisa Wauters⁴, Zachary Miller⁴, Maria Gorno-Tempini⁴, Gopala Anumanchipalli²

¹Zhejiang University, China ²UC Berkeley, United States

³Southern University of Science and Technology, China ⁴UCSF, United States

yezongli25@gmail.com, jiachenlian@berkeley.edu, gopala@berkeley.edu

Abstract

Accurate alignment of dysfluent speech with intended text is crucial for automating the diagnosis of neurodegenerative speech disorders. Traditional methods often fail to model phoneme similarities effectively, limiting their performance. In this work, we propose Neural LCS, a novel approach for dysfluent text-text and speech-text alignment. Neural LCS addresses key challenges, including partial alignment and context-aware similarity mapping, by leveraging robust phoneme-level modeling. We evaluate our method on a large-scale simulated dataset, generated using advanced data simulation techniques, and real PPA data. Neural LCS significantly outperforms state-of-the-art models in both alignment accuracy and dysfluent speech segmentation. Our results demonstrate the potential of Neural LCS to enhance automated systems for diagnosing and analyzing speech disorders, offering a more accurate and linguistically grounded solution for dysfluent speech alignment.

Index Terms: speech pronunciation, dysfluency, forced alignment, clinical

1. Introduction

The diagnosis and analysis of neurodegenerative speech disorders, such as primary progressive aphasia (PPA) [1], traditionally depend on real-time MRIs (rtMRIs) and manual speech transcripts by speech-language pathologists (SLPs). Recent automated approaches for diagnosing and analyzing dysfluent speech have focused on comparing uttered speech (actual spoken text) with lexical speech (intended text), with discrepancies termed dysfluencies like sound repetition, insertion, deletion, and substitution [2, 3]. Accurate identification of dysfluencies is essential for developing automated speech disorder diagnosis systems, relying on precise *dysfluent speech-text alignment*.

Speech-text alignment, or forced alignment, maps speech tokens to corresponding text and identifies their temporal boundaries. It is crucial for tasks like text-to-speech (TTS) synthesis, data segmentation, phonetic research, and speech assessment. Various methods exist for forced alignment [4, 5, 6, 7, 8]. Typically, it assumes strong monotonicity and element-wise alignability, where each speech token is monotonically mapped to a text token based on similarity. While this works for fluent speech, *aligning dysfluent or disordered speech requires different or stricter constraints*.

We illustrate an ideal dysfluent speech alignment and its role in enhancing automatic speech disorder diagnosis. Suppose a patient is instructed by an SLP to read "A pen on the table," with the phonetic ground truth: /AH . P EH N . AA N . DH AH . T EY B AH L./ . A possible dysfluent transcription might be: /UH. UH. EY. P EN K N. AH N. DH AH. DH AH. T T T EY B AH L./ .

An ideal alignment would map actual pronunciations to intended phonemes as: AH-(UH, UH, EY) . P-(P) EH-(EH K) N-(N) . AA-(AH) N-(N) . DH-(DH AH DH) AH-(AH) . T-(T, T, T) EY-(EY) B-(B) AH-(AH) L-(L) / . This alignment process implicitly performs dysfluency detection. For instance, EH-(EH K) marks an insertion, while AH-(UH, UH, EY) reveals repetition of acoustically similar sounds. Unlike traditional fluent speech alignment, there are three challenges in dysfluent speech text alignment, which we outline next.

First, dysfluent speech is often only partially aligned with the text. For example, when pronouncing pen (/P EH N/), disordered speech might produce /P K EN N/, where the randomly inserted sound /K/ should not be aligned. Here, we represent speech using its transcribed phoneme sequence. Second, robust, context-aware similarity mapping is crucial, as exhaustive enumeration fails to address complexities like /B B P EN N/, where /P-B/ substitution indicates a voicing error, suggesting /B/ should align with /P/ based on articulatory, acoustic, or semantic similarity. However, accurate phonetic transcription remains challenging, as state-of-the-art phoneme recognition systems [9, 10, 11] struggle with atypical speech.

SSDM [12, 13] framed dysfluent speech alignment as a local sequence alignment problem, proposing the longest common subsequence (LCS) algorithm [14] as a solution. Unlike global aligners like DTW [15], which consider all tokens, LCS focuses on matching tokens, ignoring irrelevant ones. SSDM introduced the connectionist subsequence aligner (CSA) as a differentiable LCS, but preliminary results show minimal improvement over vanilla LCS, especially in phoneme similarity modeling, as CSA treats phonetically similar sounds as distinct. In this context, a robust, and linguistically grounded subsequence aligner is on the verge of emerging.

To provide experiments with larger-scale data that can more realistically reflect acoustic characteristics (such as phoneme pronunciation similarities), new data simulation methods have been explored. We inject disfluencies into VCTK [16] text data based on similarity probabilities, generating a larger-scale disfluent dataset. Additionally, to ensure more natural text-speech data, we employ an LLM+TTS approach for synthesis.

In this work, we propose Neural LCS to handle the aforementioned problem. First, we focus on dysfluent text-to-text alignment. Our model outperforms significantly Hard LCS and DTW, especially at the phoneme level. Next, we introduce a speech-to-text alignment model based on Neural LCS and test it on dysfluent speech segmentation using simulated speech and PPA data. Our results surpass the current state-of-the-art models. To facilitate further research, we open-source our work at <https://github.com/Berkeley-Speech-Group/Neural-LCS.git>

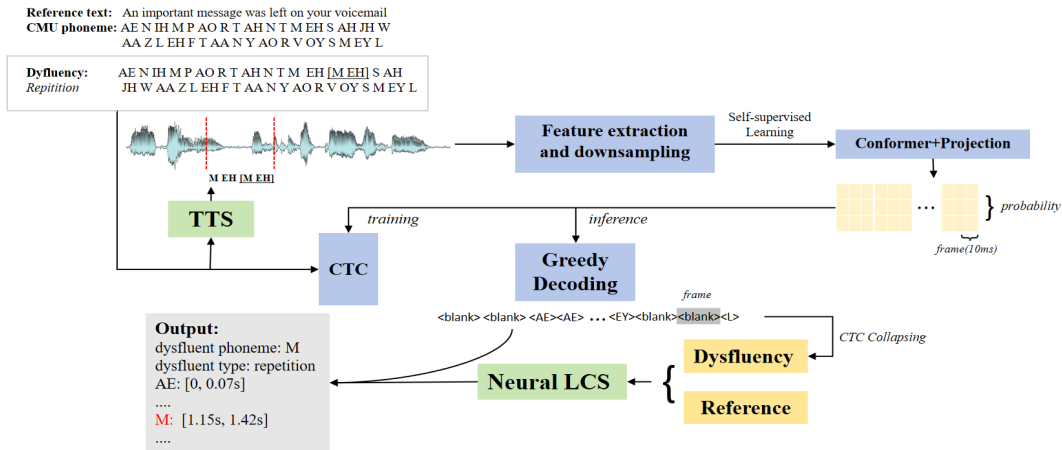


Figure 2: : Structure of Speech-text alignment model

2.3.1. Feature Encoder

We use the default T5 [25] feature encoder with a fully-visible mask, enabling all tokens to attend to each other. For phoneme-level tokenization, we implemented a custom tokenizer based on the CMU phoneme dictionary, while for word-level tokenization, we used the default T5-small tokenizer.

2.3.2. Training Objective

Our alignment label exhibits an imbalance in class distribution. If a text contains less dysfluency, label 1 accounts for most of its corresponding label sequence. Thus we employ Focal Loss [26] \mathcal{L}_{FL} , which is defined as:

$$\mathcal{L}_{FL}(p_t) = -\alpha(1 - p_t)^\gamma \log(p_t) \quad (1)$$

where p_t represents the predicted probability of the true class, α is a weighting factor for class balance, and γ controls the down-weighting of well-classified samples. By focusing more on hard-to-classify instances, Focal Loss mitigates the impact of class imbalance and enhances the model’s ability to learn from underrepresented classes. In this work, we let $[\alpha_0, \alpha_1, \alpha_2] = [0.5, 0.1, 0.8]$, $\gamma = 3$.

2.4. Speech-text alignment

In Sec.2.2.2, we obtained text-speech data generated by LLM. In this part, we propose STA(Speech-text alignment) model. we use Neural LCS as a basic component to implement an end-to-end framework for aligning audio and text, which can directly segment and detect dysfluencies from the speech signal. The entire paradigm is shown in Fig. 2.

For phoneme level process, wax2vec2.0 feature extractor [27] accepts the speech signal generated by LLM and TTS model, it converts the feature dimension of each frame of signal to the length of CMU dictionary + 1 via basic conformer and projection layer, and is trained with the source dysfluent sequence of the generated speech through CTC-Loss [4].

During inference, we apply greedy decoding on CTC emission matrix. This gives us the alignment between the dysfluent sequence and the audio timeline. Next, we align the collapsed dysfluent sequence with the reference text, which provides the alignment between the reference and the audio timeline. This allows us to accurately segment the audio based on the reference and speech.

The advantage of using Neural LCS as the aligner is that when the dysfluent unit transcription is inaccurate, the reference and speech can still be matched through soft alignment to achieve more accurate segmentation.

3. Experiments

3.1. Dataset

(1) **VCTK [16]**: it includes 109 native English speakers with accented speech. It’s text is used in our text-text data simulation as mentioned in Sec.2.2.1.(2) **LLM disorder**: We use LLM+TTS to generate large scale more natural dysfluent text-speech data. The detail is shown in Sec.2.2.2.(3) **PPA Speech [1]**: it is collected in collaboration with clinical experts and includes recordings from 38 participants diagnosed with Primary Progressive Aphasia (PPA). Participants were asked to read the ”grandfather passage,” resulting in approximately one hour of speech in total.

3.2. Training Details

We performed a randomized 90/10 train/test split on both text-text and text-speech data. The Neural LCS model was trained with a batch size of 32 for 15 epochs(both phoneme level and word level), totaling 16 hours on an RTX A6000, using Adam optimization with a learning rate of 1e-4, without dropout or weight decay. The phoneme-level speech-text alignment model, with a batch size of 1, was trained for 10 epochs, totaling 50 hours on the same GPU, under the same optimization settings.

3.3. Results

3.3.1. Speech-text alignment

Comparison diagram between Neural LCS algorithm and Hard LCS is shown in Fig. 3.It can be seen that our model performs well in handling the alignment of similar phonemes or words. For the phoneme level, the model can more accurately capture the pronunciation similarity between vowels and the pronunciation similarity between consonants. For the word level, the model can focus on the similarity of pronunciation or letter composition between words to achieve a more reasonable alignment. Compared with the Hard LCS algorithm, the Neural LCS model can better combine the related features of phonemes or words to achieve alignment. In addition, in the example shown

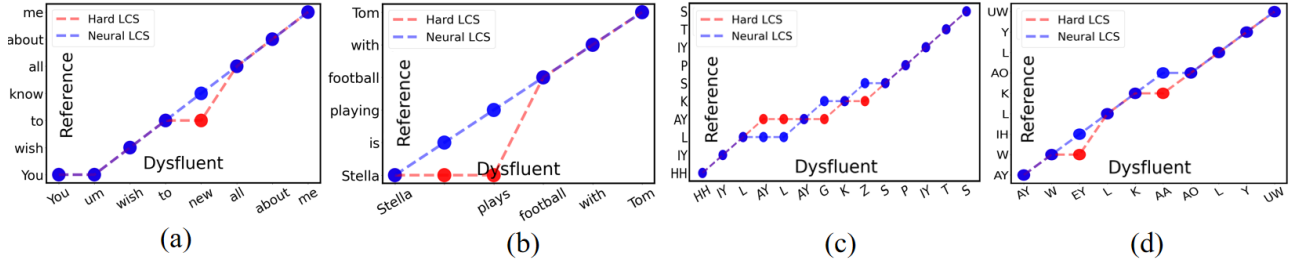


Figure 3: (a) shows that our word-level Neural LCS model captures acoustic similarities between words, even if they contain different letters, like *swiftly-wishy*. (b) demonstrates the model’s ability to capture morphological similarities, even without similar pronunciations, like *plays-playing*. (c) highlights the phoneme-level model’s ability to capture consonant similarities, such as *K-G, S-Z*. (d) shows that the model captures vowel similarities, like *IH-EY, AO-AA*.

in Fig. 3, our model can also accurately identify various dysfluency types (talked in Sec. 2.2.1) through soft alignment.

We apply DTW, Hard LCS, and Neural LCS to the test set of text-text data and LLM-generated text to compare their alignment accuracy with the reference text. Results are shown in Table 3. On both test sets, Neural LCS significantly improves alignment of dysfluent phonemes and words over traditional DTW and Hard LCS, consistent with our analysis. It also detects dysfluency types more accurately. Additionally, Neural LCS performs better at the phoneme level, likely due to the added complexity of word-level alignment, which involves both morphological and acoustic features.

Level	Method	text-text data	LLM text
Phoneme	DTW	33.47%	54.80%
	Hard LCS	24.78%	43.53%
	Neural LCS	72.55%	90.96%
Word	DTW	58.65%	62.42%
	Hard LCS	58.47%	60.67%
	Neural LCS	68.44%	75.07%

Table 3: Comparison of Different Methods

3.3.2. Speech-text alignment

We conduct our STA (Speech-text Alignment) model inference on our proposed LLM disorder data and PPA Speech, using YOLO-Stutter [18], an open-sourced state-of-the-art model for dysfluency boundary detection in speech, as the baseline. We evaluate the models using **Boundary Loss (BL)**: the mean squared error between the predicted and actual boundaries of the dysfluent regions.

Methods	Evaluated Dataset	Rep BL	Del BL	Sub BL	Ins BL
YOLO-Stutter	LLM disorder	27ms	13ms	10ms	50ms
STA model	LLM disorder	10ms	27ms	8ms	23ms
YOLO-Stutter	PPA Speech			21ms	
STA model	PPA Speech			17ms	

Table 4: Boundary Loss (BL) of the four dysfluency types

As indicated in Table 4, except deletion detection, our STA model outperforms YOLO-Stutter in terms of the BL metric. In particular, there have been significant improvements in repetition and insertion, which means that our STA model can more accurately match the dysfluent parts of speech with timestamps. Notably, our STA model adopts full sequence alignment, so except for the unfluent parts, we can perform speech alignment on all phonemes in the reference.

3.4. Ablation experiments

To investigate the impact of the proportions of different dysfluency types quantities on training results, we selected four different proportions except for average on our text-text data, as follows: $P = [\text{Repetition, Insertion, Deletion, Substitution}]$, $P_1 = [1:1.5:1:1.5]$, $P_2 = [1:1.5:1.5:1]$, $P_3 = [1:1:1.5:1.5]$, $P_4 = [1:1:1.2:1]$. Table 5 shows the type-specific accuracy on LLM disorder text (Mix refers to multiple dysfluency types in a sentence). Despite proportion adjustments, repetition accuracy remained high and stable, while substitution stayed relatively low. Pairwise comparison reveals that increasing the substitution proportion improves its accuracy but lowers insertion accuracy, and vice versa. Increasing the deletion proportion has a minimal impact on other types.

Proportions	Rep	Ins	Del	Sub	Mix
Average	96.23%	83.85%	93.42%	91.25%	90.96%
P_1	96.40%	81.95%	92.09%	92.20%	90.64%
P_2	95.79%	83.93%	93.74%	86.61%	89.68%
P_3	96.10%	81.37%	95.39%	93.22%	91.56%
P_4	96.51%	82.55%	95.04%	92.77%	90.14%

Table 5: Type-specific accuracy of different proportions

4. Conclusion and Future Work

In this work, we propose Neural LCS, a novel approach to dysfluent speech alignment that addresses limitations in existing methods. Our system operates in two modes: (1) aligning transcribed phonemes/words with reference sequences using acoustic or morphological characteristics, and (2) directly segmenting speech signals according to the reference. By incorporating acoustic phonetic similarity, Neural LCS significantly improves accuracy in detecting and segmenting various dysfluencies. We contribute two simulated dysfluent corpora: *text-text data* and *LLM Disorder*, containing acoustic features and naturally dysfluent text/audio with high-quality annotations. Experiments on both simulated and real disordered speech datasets demonstrate Neural LCS outperforms existing models in speech-text alignment, making it valuable for clinical and research applications. Neural LCS in essence tackles the word and phoneme allophony issue [28, 29], and it would be helpful to also explore better phoneme similarity models either in kinematics systems [30, 31] or gestural systems [12, 32, 33].

5. Acknowledgements

Thanks for support from UC Noyce Initiative, Society of Hellman Fellows, NIH/NIDCD, and the Schwab Innovation fund.

6. References

- [1] M. L. Gorno-Tempini, A. E. Hillis, S. Weintraub, A. Kertesz, M. Mendez, S. F. Cappa, J. M. Ogar, J. D. Rohrer, S. Black, B. F. Boeve *et al.*, “Classification of primary progressive aphasia and its variants,” *Neurology*, vol. 76, no. 11, pp. 1006–1014, 2011.
- [2] J. Lian, C. Feng, N. Farooqi, S. Li, A. Kashyap, C. J. Cho, P. Wu, R. Netzorg, T. Li, and G. K. Anumanchipalli, “Unconstrained dysfluency modeling for dysfluent speech transcription and detection,” in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8.
- [3] J. Lian and G. Anumanchipalli, “Towards hierarchical spoken language disfluency modeling,” in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, 2024, pp. 539–551.
- [4] 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.
- [5] J. Tian, B. Yan, J. Yu, C. Weng, D. Yu, and S. Watanabe, “Bayes risk ctc: Controllable ctc alignment in sequence-to-sequence tasks,” *ICLR*, 2022.
- [6] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi *et al.*, “Scaling speech technology to 1,000+ languages,” *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.
- [7] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal forced aligner: Trainable text-speech alignment using kaldii,” in *Interspeech*, vol. 2017, 2017, pp. 498–502.
- [8] J. Zhu, C. Zhang, and D. Jurgens, “Phone-to-audio alignment without text: A semi-supervised approach,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8167–8171.
- [9] “Wavlm-ctc-huggingface,” <https://huggingface.co/microsoft/wavlm-large>.
- [10] Q. Xu, A. Baevski, and M. Auli, “Simple and effective zero-shot cross-lingual phoneme recognition,” *Interspeech*, 2022.
- [11] X. Li, S. Dalmia, J. Li, M. Lee, P. Littell, J. Yao, A. Anastasopoulos, D. R. Mortensen, G. Neubig, A. W. Black *et al.*, “Universal phone recognition with a multilingual allophone system,” in *ICASSP*. IEEE, 2020, pp. 8249–8253.
- [12] J. Lian, X. Zhou, Z. Ezzes, J. Vonk, B. Morin, D. P. Baquirin, Z. Miller, M. L. Gorno Tempini, and G. Anumanchipalli, “Ssdm: Scalable speech dysfluency modeling,” in *Advances in Neural Information Processing Systems*, vol. 37, 2024.
- [13] J. Lian, X. Zhou, Z. Ezzes, J. Vonk, B. Morin, D. Baquirin, Z. Mille, M. L. G. Tempini, and G. K. Anumanchipalli, “Ssdm 2.0: Time-accurate speech rich transcription with non-fluencies,” *arXiv preprint arXiv:2412.00265*, 2024.
- [14] D. S. Hirschberg, “Algorithms for the longest common subsequence problem,” *Journal of the ACM (JACM)*, vol. 24, no. 4, pp. 664–675, 1977.
- [15] H. Sakoe, “Dynamic-programming approach to continuous speech recognition,” in *1971 Proc. the International Congress of Acoustics, Budapest*, 1971.
- [16] J. Yamagishi, C. Veaux, and K. MacDonald, “Cstr vctk corpus: English multi-speaker corpus for estr voice cloning toolkit (version 0.92),” 2019, [sound], University of Edinburgh, The Centre for Speech Technology Research (CSTR).
- [17] C. M. University, “Cmu phoneme dictionary.” [Online]. Available: <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>
- [18] X. Zhou, A. Kashyap, S. Li, A. Sharma, B. Morin, D. Baquirin, J. Vonk, Z. Ezzes, Z. Miller, M. Tempini, J. Lian, and G. Anumanchipalli, “Yolo-stutter: End-to-end region-wise speech dysfluency detection,” in *Interspeech 2024*, 2024, pp. 937–941.
- [19] X. Zhou, C. J. Cho, A. Sharma, B. Morin, D. Baquirin, J. Vonk, Z. Ezzes, Z. Miller, B. L. Tee, M. L. Gorno-Tempini *et al.*, “Stutter-solver: End-to-end multi-lingual dysfluency detection,” in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 1039–1046.
- [20] X. Zhou, J. Lian, C. J. Cho, J. Liu, Z. Ye, J. Zhang, B. Morin, D. Baquirin, J. Vonk, Z. Ezzes, Z. Miller, M. L. G. Tempini, and G. Anumanchipalli, “Time and tokens: Benchmarking end-to-end speech dysfluency detection,” 2024. [Online]. Available: <https://arxiv.org/abs/2409.13582>
- [21] Anthropic, “Model card addendum: Claude 3.5 haiku and upgraded claude 3.5 sonnet,” 2024. [Online]. Available: <https://www.anthropic.com>
- [22] J. Kim, J. Kong, and J. Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” *International Conference on Machine Learning*, 2021.
- [23] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, “Signature verification using a Siamese time delay neural network,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 6, 1994, pp. 737–744.
- [24] Y. Li, C. L. P. Chen, and T. Zhang, “A survey on siamese network: Methodologies, applications, and opportunities,” *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 6, pp. 994–1014, 2022.
- [25] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” 2023. [Online]. Available: <https://arxiv.org/abs/1910.10683>
- [26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” 2018. [Online]. Available: <https://arxiv.org/abs/1708.02002>
- [27] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition,” 2019. [Online]. Available: <https://arxiv.org/abs/1904.05862>
- [28] A. Boomershin, K. C. Hall, E. Hume, and K. Johnson, “The impact of allophony versus contrast on speech perception,” *Contrast in phonology*, pp. 143–172, 2008.
- [29] K. Choi, E. Yeo, K. Chang, S. Watanabe, and D. Mortensen, “Leveraging allophony in self-supervised speech models for atypical pronunciation assessment,” in *NAACL*, 2025.
- [30] C. J. Cho, P. Wu, T. S. Prabhune, D. Agarwal, and G. K. Anumanchipalli, “Coding speech through vocal tract kinematics,” in *IEEE JSTSP*, 2025.
- [31] 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,” in *Proc. INTER-SPEECH 2023*, 2023, pp. 5132–5136.
- [32] J. Lian, A. W. Black, L. Goldstein, and G. K. Anumanchipalli, “Deep Neural Convolutional Matrix Factorization for Articulatory Representation Decomposition,” in *Proc. Interspeech 2022*, 2022, pp. 4686–4690.
- [33] 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.