Interspeech 2024
1-5 September 2024, Kos, Greece

Examining Vocal Tract Coordination in Childhood Apraxia of Speech with Acoustic-to-Articulatory Speech Inversion Feature Sets

Nina R. Benway [1], *Jonathan L. Preston* [2], *Carol Espy-Wilson* [1]

[1]University of Maryland, College Park, MD, USA
[2]Syracuse University, Syracuse, NY, USA
benway@umd.edu, jopresto@syr.edu, espy@umd.edu

Abstract

Childhood apraxia of speech is a genetically driven, neurodevelopmental speech sound disorder with speech deficits theorized to reflect difficulty in the spatiotemporal programming of speech movements. Therefore, this work examined how well articulatory coordination features generated from audio-estimated kinematic data distinguished speakers with childhood apraxia of speech versus non-apraxic speech sound disorder. Two correlation-based feature sets motivated by recent literature demonstrated high performance in replicated 6-fold nested cross validated studies, with no statistically significant difference between feature sets (mean AUROC = .90, σ = .04). An ablation study emphasized the importance of source-filter coordination in this population of apraxic speakers, with the source-ablated feature set performing significantly worse than the lip-ablated, the tongue-ablated, and full feature set (Δ_{AUROC} = -.19, SE = 0.01, $p < .001$).

Index Terms: clinical speech technology, childhood apraxia of speech, articulatory coordination, speech inversion

1. Introduction

1.1. Childhood apraxia of speech

Childhood apraxia of speech is a neurodevelopmental speech sound disorder with a genetic basis in deficient chromatin remodeling or transcriptional regulation that impacts white matter tracts in speech-associated regions of the brain, manifesting as speech, language, and/or literacy difficulties [1]. Both segmental and suprasegmental speech errors are associated with childhood apraxia of speech [2]. These speech errors are theorized to primarily reflect difficulty in programming the spatiotemporal parameters of sequential speech movements [3]. Differential diagnosis and treatment planning, however, is complicated by lack of direct operationalization for speech features (for review: [4]) and low reliability of perceptual diagnosis [5] for childhood apraxia of speech. These factors motivate the continued development and validation of quantitative speech analysis systems for this speech disorder subtype. Because childhood apraxia of speech is theorized to impact spatiotemporal speech sequencing, this work approaches differential diagnosis through the lens of vocal tract gesture coordination.

2. Related Works and Contributions

2.1. Acoustic-to-articulatory speech inversion

Vocal tract gesture coordination is best evaluated through kinematic methods [4], but no public kinematic dataset exists for childhood apraxia of speech. Collecting traditional kinematic data (e.g., motion capture, electromagnetic articulography) is time and resource intensive. Recent signal processing developments for speech inversion, however, may be able to offset kinematic dataset scarcity in this population. Speech inversion is the task of estimating articulatory information from the continuous speech waveform [6]. Specifically, the Speech Inversion System described in [7] (Figure 1) estimates the *location* and *degree* of vocal tract constrictions posited by Articulatory Phonology [8]: lip aperture (LA), lip position (LP), tongue tip constriction location (TTCL), tongue tip constriction degree (TTCD), tongue body constriction location (TBCL), and tongue body constriction degree (TBCD). Modeling vocal tract constrictions rather than the absolute positions of individual articulators more accurately describes the vocal tract, in a speaker-independent fashion, than modeling individual articulator positions that depend on speaker dimensions [9]. Recent work has also shown that vocal tract gesture estimation is improved when glottal source characteristics (F0; periodicity, PER; and aperiodicity, APER) are included as input to the task [7, 10].

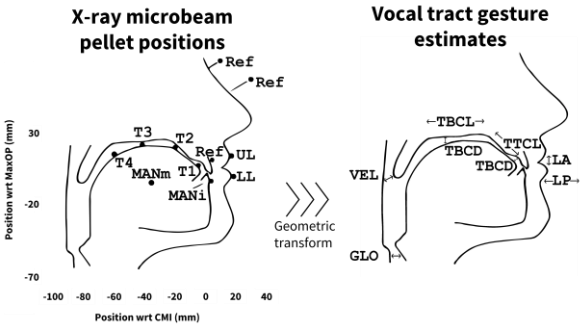

Figure 1: *Speech Inversion System estimates (right) are speaker-independent machine learning predictions of geometrically transformed, speaker-dependent ground-truth kinematic data (left).*

10.21437/Interspeech.2024-1114

2.2. Articulatory coordination

Vocal tract gesture coordination for the differential diagnosis of childhood apraxia of speech has previously been quantified through cross-correlational analysis of spatiotemporal coupling. Moss and Grigos [11] found, in children aged 3-7 years, that cross-correlational coefficients of variation (CV) significantly distinguished childhood apraxia of speech versus non-apraxia speech sound disorders, although there was no difference between disorder subtypes for measures of spatial coupling (peak correlation; PC) and temporal coupling (lag). However, the cross-correlation analysis was limited to jaw-lower lip, jaw-upper lip, and upper lip-lower lip pairings due to limitations inherent in optical motion capture instrumentation. It is not known how the PC, lag, and CV summary metrics would perform for the differential diagnosis of childhood apraxia of speech with additional kinematic channels.

Broadly, articulatory coordination features have been used in machine learning studies to detect speech changes in the timing of speech gestures related to mental health status (for review, see: [12]), including from Speech Inversion System estimates of vocal tract gestures. In [13], channel delay correlation matrices generated from kinematic estimates (AUROC = .82) outperformed channel delay correlation matrices generated from formants or Mel-frequency cepstral coefficients during the detection of psychomotor slowing in speakers with depression. However, this performance was obtained using a dilated CNN architecture, and it is not known how adult-trained Speech Inversion feature sets would perform with child speakers using machine learning architectures emphasizing parsimony and robustness for use with small clinical datasets.

2.3. Motivation and Contributions

A robust, reliable speech analysis system that can quantify vocal tract (in)coordination in the context of childhood apraxia of speech from audio could provide theoretically-motivated information assisting clinician differential diagnosis. Therefore, the purpose of this study is to lay groundwork for this task by examining the coordination feature sets that best differentiate childhood apraxia of speech and non-apraxia speech sound disorders. Our first contribution tests the hypothesis that vocal tract gesture estimates can be used in place of ground-truth kinematic data to describe articulatory coordination, as seen in our previous work with adults [13]. We show this is indeed the case (AUROC = .90, σ = .04). Our second contribution tests the hypothesis that speech systems that are not typically analyzed kinematically (i.e., other than lips [4]) will contribute relevant information to the differentiation of childhood apraxia of speech from non-apraxia speech sound disorder. We also show this is the case in an ablation paradigm, where eliminating source-filter pairings from the feature set significantly lowers classifier performance (Δ_{AUROC} = -.19, SE = 0.01, $p < .001$).

3. Methods

This study is a binary classification experiment reanalyzing data collected under Syrcause University IRB approval (# 14-117 and 17-177). All experiments were programmed in Python SciKit Learn and statistical analyses were conducted in R. For all analyses, AUROC was selected as the outcome metric to provide information on the balance between false positive and true negatives as is relevant in a clinical setting.

3.1. Dataset

Data come from a multisyllabic word repetition task administered during eligibility evaluations for four separate clinical trials of childhood apraxia of speech and non-apraxia speech sound disorder [14-17]. Participants are described in detail in each of these citations. Each speaker produced 20 multisyllabic American English words. Multisyllabic word repetition challenges the speech motor planning system by incorporating more phone and lexical stress sequences than traditional diadochokinetic testing. Stimuli include, e.g., “specificity”, “abominable”, and are further described in [18].

3.2. Features

3.2.1. Ground Truth Labels

This reanalysis reflects all speakers meeting the inclusionary criteria for the four original studies: 96 participants aged 7-23 years (Table 1). Speaker-level ground truth labels of either *childhood apraxia of speech* (CAS) or *non-apraxia speech sound disorder* (SSD) were assigned at the time of the speaker’s original enrollment. All clinical trials required that participants were speakers of American English, passed a hearing screening, scored below the 7th percentile on the Goldman Fristoe Test of Articulation [19], and had average/near average receptive vocabulary/nonverbal intelligence. Studies for non-apraxia speech sound disorder required that participants had distortions on a speech sound relative to the perceptual standard for their linguistic community. Studies for childhood apraxia of speech required that participants demonstrate at least two of the following: sound additions in word repetitions, slow/errored diadochokinetic /pɑtɑkɑ/, lexical stress errors in multisyllabic picture naming, and evidence of syllable segregation in multisyllabic picture naming. Inclusionary details are described fully in each study’s original publication.

Table 1:*Dataset. CAS = childhood apraxia of speech; SSD = non-apraxia speech sound disorder.*

Subset	Speakers	Age $\bar{x}$ (σ_x) [*range*]	Utterances
CAS	45	10.9 (1.9) [9-17]	854
SSD	51	11.3 (3.4) [7-23]	969

3.2.2. Estimating kinematic data with Speech Inversion

Estimates of vocal tract gesture location and degree were generated using the Self-Supervised Learning Speech Inversion tool described in [7] and available at https://github.com/Yashish92/SSL-SI-tool. This Speech Inversion model was initially trained with ground-truth data from adult speakers of the Wisconsin X-Ray Microbeam dataset [20], and estimates LA, LP, TTCL, TTDC, TBCL, TBCD. Three additional source features, APER, PER, and F0, were extracted to represent phonatory characteristics as described in [10] from the study audio. The tract variable and source estimates were combined, resulting in feature representations of 9 estimated kinematic channels sampled every millisecond for 2 seconds (9x200). Each channel was z-normalized during the estimation process.

3.2.3. Cross-correlational PC, lag, CV

Measures of spatiotemporal coupling – PC, lag, and CV – were calculated following the methods of Moss and Grigos [11], using the Speech Inversion System estimates in place of traditional kinematic data. Fisher transformed cross-

correlations (z) were generated for each of the nine estimated kinematic channels ($r_{i,j}$). PC was defined as in [11], as the maximum value in the Fisher transformation for a given Fischer-transformed correlation $z_{i,j}$:

$$\left| z_{peak}^{i,j} \right| = max\left(\left| z_{i,j}^{d} \right| \right) \quad \forall d \in [0, D] \tag{1}$$

Lag was also defined as in [11], as the sample position in the Fisher-transformed correlation $z_{i,j}$ containing the PC value.

$$z_{peak}^{i,j} = \arg\max_{d}\left(\left|z_{i,j}^{d}\right|\right) \tag{2}$$

The absolute value of PC and lag were retained [11]. CV was calculated from the mean (μ) and standard deviation (σ) of the Fisher-transformed correlation $z_{i,j}$:

$$CV_{i,j} = \frac{\sigma_{z_{i,j}}}{\mu_{z_{i,j}}} \tag{3}$$

3.2.4. Channel-delay correlation matrices

Articulatory coordination features were derived as in [13]. Given an *M*-dimensional feature vector *X*, *N* samples long with delay *d*:

$$r_{i,j}^{d} = \frac{\sum_{t=0}^{N-d-1} x_i[t]x_j[t+d]}{N-|d|} \tag{4}$$

Each word was represented as a channel-delay correlation matrix (CDCM) constructed from each of the estimated kinematic channels, with redundant cross-correlational pairings discarded from the analysis. The value of *d* was specified in 5 ms increments, resulting in 11 delays 0ms to 50 ms, inclusive.

3.3. Experimental Design

Although the clinical trial in [14] is the largest to date for childhood apraxia of speech, it is still a relatively small dataset. Therefore, a k-fold nested cross validation experimental design was chosen to maximize statistical confidence and power while providing unbiased accuracy estimates in models where feature selection and hyperparameter tuning are employed [21]. We used 6 outer folds (16% of data) so each participant was represented once in the held-out test performance metrics. Outer folds were stratified by speaker class, such that speakers were drawn 1:1 at random from groups of SSD and CAS speakers, with every 7th draw including one more speaker from the SSD group. For each inner fold, 85% of speakers were selected for the inner training set and 15% of speakers were selected for the inner validation set, with 10 randomly-selected inner folds per outer fold (Figure 2). The entire 6-fold nested cross validation experiment was replicated 3 times to demonstrate the distribution of experimental results, with each replication evidencing performance if different combinations of participants were included in the inner folds and outer folds.

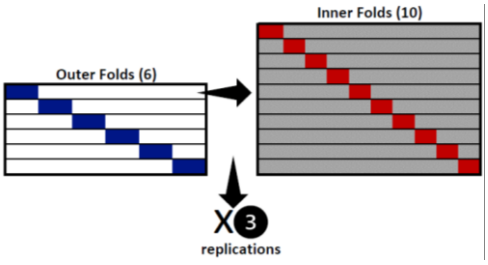

Figure 2: *6-fold nested cross validation experimental design. Blue = test (16%), red = validation (12.6%), gray = training (71.4%)*

3.4. Model Architecture and Tuning

Because there are relatively small number of participants in this clinical dataset, we chose classifiers emphasizing parsimony and clinical interpretability. First, a logistic regression classifier was used to predict word-level predictions relative to the speaker ground truth label. Speaker-level predictions were made by a metaclassifier that evaluated the average word-level logistic regression probability against a threshold of 0.5.

Principal component analysis (PCA) was used to reduce the dimensionality of each dataset, as has been customary in studies of articulatory coordination [22]. Each PCA was tuned such that the components describing 90% of the estimated variance in the training set were retained, with the model fit on training data used to transform validation and test data. When initial results indicated that the two feature sets were performing well, it raised the question of whether high performance could also be seen without the PCA, as direct representations of features would be most clinically interpretable. So, the experiments were re-run without PCA and all final results are reported based on models without PCA to emphasize clinical interpretability.

For all studies, hyperparameters were tuned with Optuna using the inner fold training and validation sets, and the best performing inner fold was used to determine the hyperparameters for testing the outer fold. Tuned hyperparameters relevant for logistic regression included: (inverse) regularization strength, regularization type, solver, whether to fit an intercept, the maximum number of iterations, class weightings, and tolerance for stopping.

3.5. Ablation Studies

Ablation studies using the best performing feature set and model architecture were performed to systematically assess changes in model performance due to the spatiotemporal correlations between lips, tongue, and source channels. The following parameters were selected for ablation: lip correlations (preserving tongue and source correlations), tongue correlations (preserving lip and source correlations), and source correlations (preserving lip and tongue correlations).

4. Results

4.1. Feature Set and Model Architecture Comparison

Results from the first set of experiments are shown in Table 3, and are reported as the mean (standard deviation) from all test set shuffles in the three study replications.

Table 3: *Mean (standard deviation) of performance. Word results reflect linear regression predictions for 1823 words. Speaker results reflect metaclassification of word-level probabilities for 96 speakers. Table continues next page.*

Analysis	Goal	AUROC	Precision	Recall	F1
PC/lag/CV	Word	.89 (5e-2)	.85 (4e-3)	.77 (1e-1)	.80 (6e-2)
+ PCA	Speaker	.99 (2e-3)	.95 (1e-3)	.87 (2e-2)	.91 (1e-2)
CDCM	Word	.87 (6e-2)	.78 (1e-1)	.80 (9e-2)	.78 (7e-2)
+ PCA	Speaker	.93 (8e-3)	.87 (5e-2)	.86 (3e-2)	.86 (3e-2)

PC/lag/CV	**Word**	**.90 (4e-2)**	**.86 (6e-2)**	**.75 (1e-1)**	**.79 (1e-1)**
	Speaker	**.97 (1e-2)**	**.96 (2e-2)**	**.84 (7e-2)**	**.89 (4e-2)**
CDCM	Word	.90 (5e-2)	.84 (9e-2)	.76 (2e-1)	.77 (2e-1)
	Speaker	.95 (2e-2)	.95 (2e-2)	.81 (1e-1)	.86 (8e-2)

Linear mixed effects models with random intercepts for each 6-fold shuffle were used to test the null hypotheses that the distributions of participant performances did not differ across analyses and across replications. Removing the PCA significantly increased AUROC by an average of .03 (SE = .01, $t = 2.655$, $p < .01$). The choice of feature set did not significantly influence AUROC, nor did the experimental replications. Therefore, the PC/lag/CV feature set was chosen for further analysis as it represents a lower-dimension feature set which may improve robustness in clinical use cases [23].

4.2. Reliability of Word-Level Predictions

Intraclass correlation coefficients assessed the replication-to-replication reliability of word level predictions in the endorsed PC/lag/CV feature set, as analytical stability is required for eventual clinical use. The intraclass correlation coefficient for word-level predictions across replications was high (ICC = .99, 95% CI = [.99,.99], random raters, average agreement). Of 1823 words in the analysis, 136 had non-unanimous class agreements across the 3 replications; therefore, we also examined the reliability of the thresholded word predictions. This intraclass correlation coefficient was calculated by estimating the repeatability statistic with the rptR package in R, using a link-scale approximation for binary data, and was judged to be high (ICC = .99, 95% CI = [0.995, 0.996]).

4.3. Ablation Studies

Ablation results are shown in Table 4, and are reported as the mean (standard deviation) of all test set shuffles in the three study replications using the endorsed PC/lag/CV feature set.

Table 4: *Mean (standard deviation) of word-level classification*

Ablated Correlations	AUROC	Precision	Recall	F1
Full PC/lag/CV	.90 (4e-2)	.86 (6e-2)	.75 (1e-1)	.79 (1e-1)
Lips	.89 (5e-2)	.86 (6e-2)	.77 (1e-2)	.81 (7e-2)
Tongue	.88 (5e-2)	.85 (6e-2)	.76 (9e-2)	.80 (7e-2)
Source	.71 (6e-2)	.65 (7e-2)	.60 (1e-1)	.61 (9e-2)

Linear mixed effects models with random intercepts for each replication-shuffle were used to test the null hypotheses that the distributions of participant performances in models trained with ablated feature sets did not differ from the non-ablated (full) baseline feature set, and that performance did not differ significantly across replications. Ablating training data related to the lips (i.e., LA, LP) or tongue (i.e., TTCL, TTCD, TBCL, TBCD) did not significantly change performance compared to the PC/lag/CV baseline containing all lip, tongue, and source channels. However, ablating training data related to source information (i.e., APER, PER, F0) did significantly decrease average AUROC performance by .19 (SE = .01, $t = -15.4$, $p < .001$). The distribution of receiver operating curves for the models trained on the non-ablated baseline feature set and the models trained on the source-ablated feature sets is shown in Figure 3.

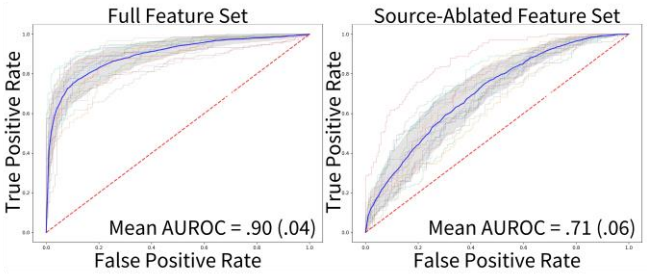

Figure 3: *Eighteen replicated receiver operating curves for the full feature set and source-ablated feature set, showing the average curves (blue) and standard deviations (gray).*

5. Discussion and Conclusions

This study provides evidence supporting our two hypotheses. First, we demonstrated that adult-trained Speech Inversion estimates can capture kinematic information relevant for the differential diagnosis of childhood apraxia of speech versus non-apraxia speech sound disorder. It is noteworthy that this was seen even with low-dimensional feature sets which are specifically desired for clinical applications [23]. Second, we demonstrated that signal channels not typically considered in traditional kinematic studies [4] capture relevant information for the differential diagnosis of childhood apraxia of speech. Specifically, the coordination of source-filter gestures may be a salient kinematic marker, replicating the importance of voicing errors seen in perceptual studies of childhood apraxia of speech [18]. Together, this study's findings motivate the continued development of a clinical speech analysis system to quantify vocal tract coordination in individual words, with no specialized kinematic instrumentation required.

Work is underway to overcome limitations related to the scope of this preliminary study. First, ongoing work examines the relative importance of specific vocal tract couplings. Second, ongoing work examines how estimated kinematic data aligns with perceptually identified speech features in this dataset [18]. Third, we are seeking to replicate these results in another suitable dataset.

Even though adult-trained Speech Inversion estimates performed well in this classification task, it is unknown how valid and reliable the adult-trained estimates are when compared to a kinematic ground truth for child vocal tracts. Validating a Speech Inversion System for children would provide a methodological tool that could benefit child speech research, broadly. A child-validated tool may ultimately elucidate a pathway for widespread child kinematic data collection, averting barriers associated with electromagnetic articulography or optical motion capture. With such a tool, (estimated) kinematic data might represent a wider range of children, because not all children might tolerate sensor placement for kinematic studies. This consideration is particularly salient given that childhood apraxia of speech can co-occur with disorders, such as autism spectrum disorder, where sensory adversion to sensors might be prevalent. Additionally, estimated kinematic data might be able to increase the number of signal channels that are analyzed, particularly for child speech where lingual sensor placement is limited by tongue size. For these reasons, validation of a child-trained Speech Inversion System is an important future direction of this line of research.

6. Acknowledgements

Funding for this work was provided by NIH T32DC000046-28 (N. Benway, Trainee, C. Espy-Wilson, Mentor, and M. Goupell & C. Carr, PIs), NSF 5-239382 (C. Espy-Wilson, PI), and NIH R15DC016426 (J. Preston, PI).

7. References

[1] Morgan, A.T., et al., *Genetic architecture of childhood speech disorder: a review*. Molecular Psychiatry, 2024.

[2] American Speech-Language-Hearing Association, *Childhood apraxia of speech [technical report]*. 2007: Available from www.asha.org/policy.

[3] American Speech-Language-Hearing Association, *Childhood apraxia of speech [Position Statement]. Available from www.asha.org/policy.* 2007.

[4] Terband, H., et al., *Assessment of childhood apraxia of speech: A review/tutorial of objective measurement techniques.* Journal of Speech, Language, and Hearing Research, 2019. 62(8S): p. 2999-3032.

[5] Murray, E., et al., *The Reliability of Expert Diagnosis of Childhood Apraxia of Speech.* Journal of Speech, Language, and Hearing Research, 2023.

[6] Papcun, G., et al., *Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data.* The Journal of the Acoustical Society of America, 1992. 92(2): p. 688-700.

[7] Siriwardena, Y.M. and C. Espy-Wilson. *The Secret Source: Incorporating Source Features to Improve Acoustic-To-Articulatory Speech Inversion*. in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2023.

[8] Browman, C.P. and L. Goldstein, *Articulatory phonology: An overview.* Phonetica, 1992. 49(3-4): p. 155-180.

[9] Mitra, V., et al., *Retrieving Tract Variables From Acoustics: A Comparison of Different Machine Learning Strategies.* IEEE Journal of Selected Topics in Signal Processing, 2010. 4(6): p. 1027-1045.

[10] Deshmukh, O., et al., *Use of temporal information: detection of periodicity, aperiodicity, and pitch in speech.* IEEE Transactions on Speech and Audio Processing, 2005. 13(5): p. 776-786.

[11] Moss, A. and M.I. Grigos, *Interarticulatory coordination of the lips and jaw in childhood apraxia of speech.* Journal of Medical Speech-Language Pathology, 2012. 20(4): p. 127.

[12] Cummins, N., et al., *A review of depression and suicide risk assessment using speech analysis.* Speech Communication, 2015. 71: p. 10-49.

[13] Seneviratne, N. and C. Espy-Wilson. *Multimodal Depression Classification using Articulatory Coordination Features and Hierarchical Attention Based text Embeddings*. in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022.

[14] Preston, J.L., et al., *A Randomized Controlled Trial of Treatment Distribution and Biofeedback Effects on Speech Production in School-Age Children With Apraxia of Speech.* Journal of Speech, Language, and Hearing Research, 2023. ePub Ahead of Issue.

[15] Preston, J.L., et al., *Variable practice to enhance speech learning in ultrasound biofeedback treatment for childhood apraxia of speech: A single case experimental study.* American Journal of Speech-Language Pathology, 2017. 26(3): p. 840-852.

[16] Sjolie, G.M., M.C. Leece, and J.L. Preston, *Acquisition, retention, and generalization of rhotics with and without ultrasound visual feedback.* Journal of Communication Disorders, 2016. 64: p. 62-77.

[17] Preston, J.L., M.C. Leece, and E. Maas, *Intensive treatment with ultrasound visual feedback for speech sound errors in childhood apraxia.* Frontiers in Human Neuroscience, 2016. 10: p. 1-9.

[18] Benway, N.R. and J.L. Preston, *Differences between school-age children with apraxia of speech and other speech sound disorders on multisyllable repetition.* Perspectives of the ASHA Special Interest Groups, 2020. 5(4): p. 794-808.

[19] Goldman, R. and M. Fristoe, *Goldman Fristoe Test of Articulation - Third Edition*. 2015: Pearson.

[20] Westbury, J.R., G. Turner, and J. Dembowski, *X-ray microbeam speech production database user's handbook.* University of Wisconsin, 1994.

[21] Ghasemzadeh, H., R.E. Hillman, and D.D. Mehta, *Toward Generalizable Machine Learning Models in Speech, Language, and Hearing Sciences: Sample Size Estimation and Reducing Overfitting.* arXiv e-prints, 2023: p. arXiv: 2308.11197.

[22] Huang, Z., J. Epps, and D. Joachim. *Exploiting Vocal Tract Coordination Using Dilated CNNS For Depression Detection In Naturalistic Environments*. in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020.

[23] Berisha, V., et al., *Digital medicine and the curse of dimensionality.* NPJ Digital Medicine, 2021. 4(1).