



Development and Validation of a Wav2Vec 2.0-Based Cross-Language Methodology for Measurement of Articulatory Precision

Tanya Talkar, Kan Kawabata, Connor Higgins, Sean Tobyne

Linus Health, USA

ttalkar@linus.health

Abstract

Degraded speech intelligibility due to dysarthrias can result from changes inherent to diseases such as Amyotrophic Lateral Sclerosis (ALS). Tracking this degradation longitudinally can highlight treatment effects. Various approaches have been developed to automatically measure speech intelligibility, often focusing on output from automatic speech recognition systems. In this paper, we extract phonetic outputs from the wav2vec 2.0 model and compare them to expected phonetic outputs across 12 languages to derive Articulatory Precision (ArtP). The strongest correlations between ArtP and Azure Pronunciation Assessment were 0.93 for English, 0.85 in German, and 0.66 for Swedish. We additionally find that ArtP has a correlation of 0.77 with the ALSFRS-R speech subscore and is sensitive to perceptual measures of speech intelligibility. The measure shows promise as a reliable and clinically relevant tool that can be used in multiple languages and disorders to assess speech intelligibility.

Index Terms: speech biomarkers, healthcare, articulation

1. Introduction

Speech production involves coordination within and across four fundamental subsystems - respiratory, phonatory, articulatory, and resonatory [1]. Impairments in any of the subsystems can contribute to observable perceptual differences, such as a decrease in speech intelligibility [1, 2]. Speech intelligibility differences are commonly reported in conditions such as Amyotrophic Lateral Sclerosis (ALS) [3, 4], Alzheimer’s Disease (AD) [5, 6], and Parkinson’s Disease (PD) [7, 8]. This may be due to dysarthria or apraxia [9, 10], depending on the manifestation of the disorder. Reduced speech intelligibility can cause difficulties in communication, which can lead to feelings of isolation and a decrease in quality of life [11, 12]. Due to its presence in many disorders, as well as its effect on quality of life, speech intelligibility is of great interest as a potential clinical endpoint in clinical trials to measure treatment effects [13]. Tracking speech intelligibility over time could also be useful in understanding disease progression.

Although measures of speech intelligibility have been used in clinics and on scales such as the revised Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS-R) and the Unified Parkinson’s Disease Rating Scale (UPDRS), the measures are subjective and may not track subtleties in speech degradation [14, 15]. The scales also require the presence of a trained clinician, which limits the possible frequency of assessment and increases the burden on clinical trials or within a health system. Automated methods of measuring speech intelligibility have been developed and can aid in remote assessment and longitudinal collection [16, 17, 18, 19, 20]. One approach has been

to use automatic speech recognition (ASR) systems to measure intelligibility [18, 19] by measuring transcription errors from ASR systems that are hypothesized to occur due to impaired speech. Analysis of these errors can then be used to determine the impairment severity, which can be correlated with existing scales of speech impairment. There are a few different scales at which intelligibility can be calculated using ASR outputs (e.g. words, phonemes). For the purposes of this study, we will focus on analysis of phonological features [17, 20].

One approach to measuring speech intelligibility is Articulatory Precision (ArtP), or how accurately an individual produces the expected sounds given a specific transcription. We can compute the phone log-likelihood ratio [15, 17, 20, 21] by comparing the likelihood of an expected phone to the actual phone that was produced, after alignment through forced alignment methods. The average of the differences in the likelihoods across the transcript can provide insight into the precision of pronunciation and, subsequently, speech intelligibility. This approach has shown promise in the detection of mild cognitive impairment (MCI) [20] and correlates with perceptual scores of dysarthria [17]. It has also been shown to track changes in speech impairment more sensitively as compared to the ALSFRS-R speech subscore [15]. However, existing measures of ArtP are limited in the number of languages in which they have been implemented (primarily English). We developed a cross-language approach to generate a more universal measure of ArtP for a wider range of individuals to participate in clinical trials or longitudinal monitoring.

In this paper, we first describe the architecture behind the cross-language ArtP measure and the datasets used to train and fine-tune the model. We then detail the verification of the model for 12 languages. We also perform analytical validation to compare the ArtP measure in English with the ALSFRS-R subscore. We finally describe a clinical validation to show that the ArtP measure is able to detect speech intelligibility changes more sensitively than the ALSFRS-R speech subscore in English. These steps follow the V3 framework provided by DiMe to evaluate the utility of the ArtP measure [22]. To our knowledge, this is the first version of the measure that has been developed and extended to this full language set.

2. Methods

2.1. Datasets

We used a data collection agency, Appen, to acquire data from 50 individuals in each of 11 languages (550 individuals total). The languages, their locales, the male/female count, and age statistics are provided in Table 1.

The data collection involved recordings of participants reading two single sentence prompts that had been translated

Table 1: List of languages, their locales, the male/female split, and the age distribution for participants in the collection.

| Language | Locales | M/F | Age |
|---------------|---|-------|---------------|
| Danish (da) | Denmark | 25/25 | 36.54 ± 13.89 |
| German (de) | Germany, Belgium, Switzerland | 25/25 | 37.24 ± 11.29 |
| Spanish (es) | Spain, US | 25/25 | 40.54 ± 9.95 |
| Finnish (fi) | Finland | 25/25 | 41.48 ± 14.82 |
| French (fr) | France, Belgium, Canada, Switzerland | 25/25 | 36.86 ± 11.53 |
| Italian (it) | Switzerland, Italy | 24/26 | 38.53 ± 13.21 |
| Japanese (ja) | Japan | 23/27 | 39.42 ± 12.69 |
| Korean (ko) | Korea | 23/27 | 37.48 ± 10.80 |
| Dutch (nl) | Netherlands, Belgium | 25/25 | 41.30 ± 11.99 |
| Polish (pl) | Poland | 23/27 | 34.74 ± 9.61 |
| Swedish (sv) | Sweden, Fin- land | 25/25 | 40.08 ± 12.93 |

into the language and locale of interest. For languages with multiple locales, we asked for data mostly from the primary country (e.g. France for French), but asked to have at least 20% of data collection come from each of the other locales (e.g. 20% from each of French-speaking Belgium, Canada, and Switzerland). The data was reviewed by native speakers to ensure that individuals adhered to the prompts, that there was no noise or other simultaneous talkers in the recording, and that participants read the sentences in a normal speaking rate and tone without nonverbal interruptions. For each sentence used in the protocol, we ran the prompted sentence text through the `phonemizer` module in Python using an eSpeak backend [23]. The phone outputs were verified to ensure they matched expected pronunciations for each language and locale. The data in this dataset is not available for public use, given commercial restrictions. Data collection was considered work-for-hire and did not require approval from an ethics committee.

We also performed analytical and clinical validation to ensure the model was able to track impairments in speech. The ALS at Home dataset consists of 110 individuals with ALS who had conducted multiple sessions of a sentence reading task in English over the course of a year through the Speech Vitals application. Recordings from 18 individuals had been rated on Dysarthria Severity and Listener Effort by trained individuals. The 126 sessions (consisting of 773 speech samples) from the 18 individuals were used in all English-based analyses presented in this paper. Additional details about collection of this dataset, the Speech Vitals application, the protocol, and access to the dataset are located in [15]. Ground truth phonetic transcriptions were generated based on the original prompted text, such that the measure would determine how closely an individual was able to adhere to the prompt, even with varying levels of dysarthria.

All audio was collected at 48000 Hz using a native app installed on personal mobile phones. Instructions were provided to have the microphone about an arms length distance away from an individual’s mouth. All audio was resampled to 16000

Hz for use with the model.

2.2. Articulatory Precision Derivation

Previous implementations of articulatory precision (ArtP) using the phonetic approach have used the Kaldi speech recognition toolkit [15, 21, 24]. While the model itself is customizable and can be augmented with additional languages, the most common models have been trained in English and Mandarin. Therefore, we decided to implement ArtP using the wav2vec 2.0 model. We utilized the `wav2vec2-xlsr-53-espeak-cv-ft` model¹, trained on the CommonVoice 13 dataset to output phonetic labels for 60 different languages [25]. The model also works on an eSpeak backend. Details about the dataset and the languages available can be found on the huggingface² and Mozilla CommonVoice websites³. The model outputs predicted phonetic labels as well as probabilities across all possible phonetic labels for each frame.

The algorithmic implementation of ArtP involves comparisons of expected phonetic transcriptions to the output of the wav2vec 2.0 model. We first generate an expected phonetic ground truth alignment between the phonemized sentence and the produced audio through an adaptation of the forced alignment pipeline in pyTorch⁴. The forced alignment gives us the frames of the audio where we would assign the expected phone. Subsequently, the generalized formula for calculating ArtP for a phonetic sequence x as compared to the transcribed sequence x^* across all aligned indices is as follows:

$$ArtP(x) = \frac{1}{|x|} \sum_{i=1}^{|x|} \log \frac{P(x_i)}{P(x_i^*)} \quad (1)$$

We average the difference between the log-likelihood of the expected phone (x_i) and the output phone (x_i^*) from the wav2vec 2.0 model across the entire sequence to return the final ArtP value for a particular recording. ArtP is then scaled from 0 and 10, with 0 indicating unintelligible speech. This was calculated on all language data, as well as on the ALS at Home data.

Initially, ArtP values for Danish, Korean, and Japanese were below 5, which typically indicates impaired speech. After listening to the recordings and looking at the wav2vec 2.0 outputs, it was clear that the transcriptions were not matching the expected perceptual output, primarily through vowel substitutions. We therefore augmented the wav2vec 2.0 model with additional data from these three languages, sourced from Mozilla CommonVoice. As the original model had been trained on Corpus 13, we downloaded data from the Delta Segments ranging from 14-20 for Japanese, Danish, and Korean. CommonVoice provides train and test splits from validated data. We additionally selected audio that was between 2-10s, to better match sentence durations from our dataset. Using the pretrained `wav2vec2-xlsr-53-espeak-cv-ft` model, we continued training for 100 epochs, with a layer dropout of 0.05 and `mask_time_prob` of 0.05. We started the learning rate at $1e-5$.

¹<https://huggingface.co/facebook/wav2vec2-xlsr-53-espeak-cv-ft>

²https://huggingface.co/datasets/mozilla-foundation/common_voice_13_0

³<https://commonvoice.mozilla.org/en/datasets>

⁴https://pytorch.org/audio/stable/tutorials/forced_alignment_for_multilingual_data_tutorial.html

We fine-tuned separate models for each language, and the results reported for Japanese, Danish, and Korean stem from the output of those fine-tuned models.

2.3. Verification

To validate the approach across multiple languages, we computed mixed effects correlations [26] between ArtP and a Pronunciation Assessment score derived from Azure Speech Services⁵. The Pronunciation Assessment is derived from Azure’s collection of training data to generate a final rating of how closely the pronunciation matches that of an expected native speaker. We chose the Pronunciation Assessment because it follows the approach of using ASR as a correlate for speech intelligibility and also provides phoneme-specific scores which we averaged to mimic the ArtP approach. We prefer to use the wav2vec 2.0 approach in future applications, however, given its customizability with additional data, particularly clinical data. We use the mixed effects correlation because of the presence of multiple recordings from each individual.

In addition to verification through comparison with the Pronunciation Assessment, we also operated under the assumption that any individual would have approximately the same pronunciation across the two sentences. We therefore also calculated the ICC between the ArtP values across the two sentences for the language data using a one-way random effects model.

2.4. Analytical Validation

We performed analytical validation using the ALS At Home dataset [15], which contains longitudinal ALSFRS-R speech subscores. We report the mixed effects correlation [26] between the ArtP averaged over all sentences in a data collection session and the ALSFRS-R speech subscores for that session. For analytical and subsequent clinical validation, we scaled all ArtP values to a range of 0 to 10, with 0 indicating completely unintelligible speech.

2.5. Clinical Validation

We follow our previous approach laid out in [15] to calculate the Minimally Detectable Change (MDC) and Minimal Clinically Important Difference (MCID) for ArtP on speech data from the ALS At Home study. The MDC describes the natural variation in the measure over time, while the MCID describes the minimal change in the measure that corresponds to one point of change in a perceptual clinical rating - Dysarthria Severity or Listener Effort for this dataset. [15] described that the MDC for our previous ArtP measure was lower than the two MCIDs, while the MDC of the ALSFRS-R Speech subscore was greater than the MCIDs, indicating that ArtP was more sensitive to clinically perceptual differences in intelligibility as compared to the ALSFRS-R Speech subscore. We calculated the MDC and MCIDs for the new ArtP measure to verify that it continued to be sensitive with the cross-language model.

3. Results

3.1. Verification

Table 2 lists the mixed-effects correlations between ArtP and Azure Pronunciation Assessment. We see strong correlations

⁵<https://learn.microsoft.com/en-us/azure/ai-services/speech-service/how-to-pronunciation-assessment?pivot=programming-language-python>

Table 2: Correlation and p-values for comparison between ArtP measure and Pronunciation Assessment from Azure. Strong correlations are bolded

| Language | Azure Corr. | p-value |
|----------|-------------|---------|
| English | 0.93 | < 0.001 |
| Danish | 0.55 | 0.001 |
| Dutch | 0.44 | 0.03 |
| Finnish | 0.53 | < 0.001 |
| French | 0.65 | < 0.001 |
| German | 0.85 | < 0.001 |
| Italian | 0.62 | < 0.001 |
| Japanese | 0.54 | 0.06 |
| Korean | 0.45 | 0.006 |
| Polish | 0.46 | 0.005 |
| Spanish | 0.59 | 0.006 |
| Swedish | 0.66 | < 0.001 |

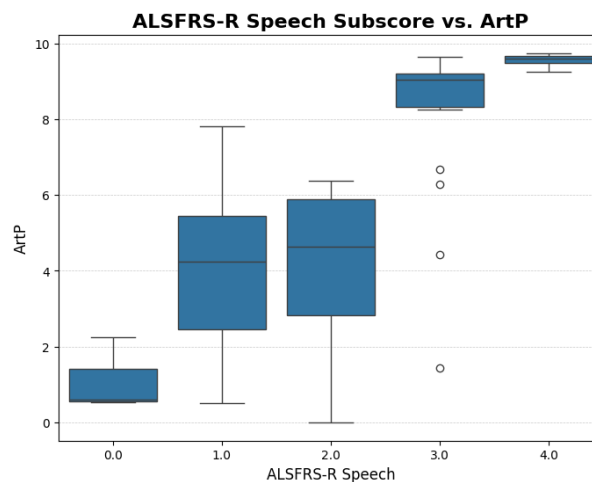


Figure 1: Boxplots of ArtP values across ALSFRS-R Speech subscores

for English and German, and moderate correlations for all others - Danish, Dutch, Finnish, French, Italian, Japanese, Korean, Polish, Spanish, and Swedish [27].

We calculated an ICC of 0.91 (95% CI: [0.89, 0.92], $p < 0.001$) for the language data when comparing ArtP values of the two sentences from each participant. Each sentence was marked as a rater.

3.2. Analytical & Clinical Validation

The mixed-effects correlation between the ALSFRS-R Speech subscore and the ArtP metric was 0.77 ($p < 0.001$) [26]. This is shown in Figure 1, with lower ArtP scores associated with greater impairment.

For clinical validation, we found that the MCID for ArtP was greater than the MDC in both Dysarthria Severity and Listener Effort, indicating that it sensitively detects speech intelligibility changes. It continued to be more sensitive as compared to the ALSFRS-R Speech subscore as shown in Table 3.

Table 3: Comparison of Minimally Detectable Change (MDC) and Minimal Clinically Important Difference (MCID) for ArtP and the ALSFRS-R Speech subscore, relative to ratings of Dysarthria Severity and Listener Effort

| Measure | Clinical Rating | MDC | MCID |
|---------------------------|---------------------|------|------|
| Articulatory Precision | Dysarthria Severity | 0.86 | 0.82 |
| | Listener Effort | 0.86 | 0.72 |
| ALSFRS-R Speech Sub-score | Dysarthria Severity | 0.97 | 2.27 |
| | Listener Effort | 0.97 | 1.99 |

4. Discussion

In this study, we present a cross-language derivation procedure and validation for Articulatory Precision (ArtP), a measure meant to capture speech intelligibility and pronunciation accuracy. As speech intelligibility is a common complaint within diseases such as ALS, Alzheimer’s Disease, and Parkinson’s Disease, it is of interest as both a clinical endpoint to measure treatment effects and a marker of disease progression. Existing techniques to measure goodness of pronunciation (GOP) or speech intelligibility have utilized automatic speech recognition outputs to analyze transcription outputs, as we have done in this paper, and compared expected phonetic sequences to ASR output sequences. Our approach builds upon existing approaches and extends to a set of 12 languages. In addition, we present the applicability of the measure through the V3 framework, focusing on verification and analytical and clinical validation on an ALS dataset in English.

4.1. Verification & Limitations of Verification Approach

In verification, we found strong correlations for English and German, but moderate correlations for all other languages. The numbers from our combined dataset with 568 individuals match or exceed other GOP evaluations [16, 28, 29]. There are, however, a few caveats to keep in mind. The dataset with 11 languages (excluding English) was collected such that there was little deviation in expected pronunciations. In English, however, the ALS at Home dataset contained data from individuals with varying levels of dysarthria as they progressed with ALS. We also found that for some languages, the ArtP distribution was wider than the Pronunciation Assessment distribution (Polish, Korean), while in others ArtP had a narrower distribution (Japanese), measured by percentage of the total scale, yet both ratings suggested little or no deviation from expected pronunciations. Inherently, there was a smaller range of values in both measures across the 11 languages, which can lead to correlations that may not reflect the correlations obtained when we include cross-language data from individuals with impaired speech. We utilized the Pronunciation Assessment for phonemes from Azure as a ground truth, but do not have access to the training set nor their ground truth for assessing accuracy of phonemes. In addition, Pronunciation Assessment may generalize multiple phones under a phoneme, and therefore could be less sensitive to the subtle changes of phone substitution that we assess through ArtP. Further validation work would involve the utilization of human raters and transcribers for phonetic ground truth, particularly as we move towards collection of individuals with varying levels of dysarthria and intelligibility.

We calculated an ICC of 0.91 when assessing the ArtP values across the two prompts for each participant in the collected

language dataset. This shows initial promise for test-retest reliability of the measure, but future work will ensure that we space out repeated measures of ArtP over time for individuals to provide the true reliability measure.

Common discrepancies between expected and predicted transcriptions across languages stemmed from detection of phones that were close in place and manner. For example, in Japanese, we saw a common substitution of the uvular phone [ŋ] predicted when the original phonetic transcription had the velar [ŋ]. In Polish, we witnessed the expected close-central vowel [i] substituted by the near-close-near-front vowel [i] by the model. In Spanish, some expected [r] were predicted as [r], while the opposite was true in Dutch. We don’t know if these potential substitutions were assessed within the Pronunciation Assessment score. Though these substitutions may not affect the overall perceptual meaning of the word, phone substitutions differentiate between native and non-native speakers [30]. Therefore, it will be important to proceed with the existing phonetic dictionary and collect additional data and human-labeled phonetic transcriptions to augment and test the model to ensure that it detects the distinctions between the phones across different languages more clearly. For example, augmenting the model with CommonVoice data in Japanese helped alleviate some, but not all, of the uvular/velar substitutions, perhaps because [ŋ] and [ŋ] are allophones in Japanese [31]. Collecting the right data and fine-tuning models, perhaps based on languages that are phonetically similar to each other, will also be important as we move to look at disordered and impaired speech across these languages, as degradations of specific phones over others could help characterize impairment.

4.2. Analytical and Clinical Validation

Correlations between ArtP derived from the ALS at Home recordings and ALSFRS-R speech subscores matched and exceeded observed values from other studies [32, 33, 34]. While the sample size was small, the use of 126 sessions from those individuals indicates that the ArtP trajectory likely follows the trajectory mapped by ALSFRS-R.

In clinical validation, we were additionally able to show that the ArtP metric not only correlates with the ALSFRS-R speech subscore, but is also more sensitive to perceptual measures of Dysarthria Severity and Listener Effort. Combined with our precedent for this in [15], our results here emphasize the utility of the ArtP approach, as well as its continued applicability in English with an underlying model that has been trained on 60 different languages.

Moving forward with this work, we will need to collect longitudinal clinical data within different languages and different populations to further validate ArtP as a measure sensitive to speech intelligibility changes. The language data collected in this study, while collected at home, was reviewed to ensure that data was collected in a non-noisy environment. As we move towards utilizing ArtP as a metric for monitoring treatment effects and an individual’s progression over time, we will likely have to make adjustments to our pre-processing algorithms to account for noise and other factors that may arise through a device-agnostic data collection platform.

5. References

- [1] P. Rong, Y. Yunusova, J. Wang, and J. R. Green, “Predicting early bulbar decline in amyotrophic lateral sclerosis: A speech subsystem approach,” *Behavioural Neurology*, vol.

- 2015, no. 1, p. 183027, 2015. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1155/2015/183027>
- [2] S. d. S. Barreto and K. Z. Ortiz, "Intelligibility measurements in speech disorders: a critical review of the literature," *Pró-Fono Revista de Atualização Científica*, vol. 20, pp. 201–206, 2008.
 - [3] B. Tomik and R. J. Guiloff, "Dysarthria in amyotrophic lateral sclerosis: A review," *Amyotrophic Lateral Sclerosis*, vol. 11, no. 1-2, pp. 4–15, 2010.
 - [4] R. D. Kent, J. F. Kent, G. Weismer, R. L. Sufit, J. C. Rosenbek, R. E. Martin, and B. R. Brooks, "Impairment of speech intelligibility in men with amyotrophic lateral sclerosis," *Journal of Speech and Hearing Disorders*, vol. 55, no. 4, pp. 721–728, 1990.
 - [5] J. Jiang, J. C. Johnson, M.-C. Requena-Komuro, E. Benhamou, H. Sivasathiseelan, A. Chokesuwattanaskul, A. Nelson, R. Nortley, R. S. Weil, A. Volkmer *et al.*, "Comprehension of acoustically degraded speech in alzheimer's disease and primary progressive aphasia," *Brain*, vol. 146, no. 10, pp. 4065–4076, 2023.
 - [6] O. Ivanova, I. Martínez-Nicolás, and J. J. G. Meilán, "Speech changes in old age: methodological considerations for speech-based discrimination of healthy ageing and alzheimer's disease," *International Journal of Language & Communication Disorders*, vol. 59, no. 1, pp. 13–37, 2024.
 - [7] T. Khan, J. Westin, and M. Dougherty, "Classification of speech intelligibility in parkinson's disease," *Biocybernetics and Biomedical Engineering*, vol. 34, no. 1, pp. 35–45, 2014.
 - [8] J. Ruzs, R. Cmejla, T. Tykalova, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, and E. Ruzicka, "Imprecise vowel articulation as a potential early marker of parkinson's disease: effect of speaking task," *The Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 2171–2181, 2013.
 - [9] M. J. McAuliffe, A. R. Fletcher, S. E. Kerr, G. A. O'Beirne, and T. Anderson, "Effect of dysarthria type, speaking condition, and listener age on speech intelligibility," *American Journal of Speech-Language Pathology*, vol. 26, no. 1, pp. 113–123, 2017.
 - [10] K. V. Chenausky, D. Gagné, K. L. Stipancic, A. Shield, and J. R. Green, "The relationship between single-word speech severity and intelligibility in childhood apraxia of speech," *Journal of Speech, Language, and Hearing Research*, vol. 65, no. 3, pp. 843–857, 2022.
 - [11] E. Langlois, H. Desaeayer, M. Petrovic, K. Van Lierde, and L. De Visschere, "The influence of oral health status on speech intelligibility, articulation and quality of life of older community-dwelling people," *Gerodontology*, vol. 36, no. 4, pp. 352–357, 2019.
 - [12] S. Y. Chu and C. L. Tan, "Subjective self-rated speech intelligibility and quality of life in patients with parkinson's disease in a malaysian sample," *The Open Public Health Journal*, vol. 11, no. 1, 2018.
 - [13] S. Pinto, A. Nebel, J. Rau, R. Espesser, P. Maillachon, O. Niebuhr, P. Krack, T. Witjas, A. Ghio, M.-c. Cuartero *et al.*, "Results of a randomized clinical trial of speech after early neurostimulation in parkinson's disease," *Movement Disorders*, vol. 38, no. 2, pp. 212–222, 2023.
 - [14] L. J. Evers, J. H. Krijthe, M. J. Meinders, B. R. Bloem, and T. M. Heskes, "Measuring parkinson's disease over time: the real-world within-subject reliability of the mds-updrs," *Movement Disorders*, vol. 34, no. 10, pp. 1480–1487, 2019.
 - [15] G. Stegmann, C. Krantsevich, J. Liss, S. Charles, M. Bartlett, J. Shefner, S. Rutkove, K. Kawabata, T. Talkar, and V. Berisha, "Automated speech analytics in als: higher sensitivity of digital articulatory precision over the alsfrs-r," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 25, no. 7-8, pp. 767–775, 2024.
 - [16] J. Tröger, F. Dörr, L. Schwed, N. Linz, A. König, T. Thies, J. R. Orozco-Arroyave, and J. Ruzs, "An automatic measure for speech intelligibility in dysarthrias—validation across multiple languages and neurological disorders," *Frontiers in Digital Health*, vol. 6, p. 1440986, 2024.
 - [17] C. Middag, J.-P. Martens, G. Van Nuffelen, and M. De Bodt, "Automated intelligibility assessment of pathological speech using phonological features," *EURASIP Journal on advances in Signal Processing*, vol. 2009, pp. 1–9, 2009.
 - [18] M. Karbasi and D. Kolossa, "Asr-based speech intelligibility prediction: A review," *Hearing Research*, vol. 426, p. 108606, 2022.
 - [19] S. E. Gutz, K. L. Stipancic, Y. Yunusova, J. D. Berry, and J. R. Green, "Validity of off-the-shelf automatic speech recognition for assessing speech intelligibility and speech severity in speakers with amyotrophic lateral sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 65, no. 6, pp. 2128–2143, 2022.
 - [20] L. Xu, K. Chen, K. D. Mueller, J. Liss, and V. Berisha, "Articulatory precision from connected speech as a marker of cognitive decline in alzheimer's disease risk-enriched cohorts," *Journal of Alzheimer's Disease*, p. 13872877241300149, 2024.
 - [21] P. Kadambi, T. Mahr, L. Annear, H. Nomeland, J. Liss, K. Hustad, and V. Berisha, "How does alignment error affect automated pronunciation scoring in children's speech?" in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2024, pp. 5133–5137.
 - [22] J. C. Goldsack, A. Coravos, J. P. Bakker, B. Bent, A. V. Dowling, C. Fitzer-Attas, A. Godfrey, J. G. Godino, N. Gujar, E. Izmailova *et al.*, "Verification, analytical validation, and clinical validation (v3): the foundation of determining fit-for-purpose for biometric monitoring technologies (biomets)," *npj digital Medicine*, vol. 3, no. 1, p. 55, 2020.
 - [23] M. Bernard and H. Titeux, "Phonemizer: Text to phones transcription for multiple languages in python," *Journal of Open Source Software*, vol. 6, no. 68, p. 3958, 2021.
 - [24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.
 - [25] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.
 - [26] J. Lorah, "Effect size measures for multilevel models: Definition, interpretation, and timss example," *Large-scale assessments in education*, vol. 6, no. 1, pp. 1–11, 2018.
 - [27] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: appropriate use and interpretation," *Anesthesia & analgesia*, vol. 126, no. 5, pp. 1763–1768, 2018.
 - [28] S. Sudhakara, M. K. Ramanathi, C. Yarra, and P. K. Ghosh, "An improved goodness of pronunciation (gop) measure for pronunciation evaluation with dnn-hmm system considering hmm transition probabilities," in *INTER_SPEECH*, vol. 2, 2019, pp. 954–958.
 - [29] S. Cheng, Z. Liu, L. Li, Z. Tang, D. Wang, and T. F. Zheng, "Asr-free pronunciation assessment," *arXiv preprint arXiv:2005.11902*, 2020.
 - [30] M. Amengual, "Acoustic correlates of the spanish tap-trill contrast: Heritage and l2 spanish speakers," *Heritage Language Journal*, vol. 13, no. 2, pp. 88–112, 2016.
 - [31] M. Niikura and U. Hirschfeld, "Perception of assimilated and non-assimilated coda nasal by japanese learners of german," in *ICPhS*, 2015.
 - [32] G. M. Stegmann, S. Hahn, J. Liss, J. Shefner, S. Rutkove, K. Shelton, C. J. Duncan, and V. Berisha, "Early detection and tracking of bulbar changes in als via frequent and remote speech analysis," *NPJ digital medicine*, vol. 3, no. 1, p. 132, 2020.
 - [33] A. Wisler, K. Teplansky, J. R. Green, Y. Yunusova, T. Campbell, D. Heitzman, and J. Wang, "Speech-based estimation of bulbar regression in amyotrophic lateral sclerosis," in *Proceedings of the Eighth Workshop on Speech and Language Processing for Assistive Technologies*, 2019, pp. 24–31.
 - [34] L. Leite Neto, M. C. França Júnior, and R. Y. S. Chun, "Speech intelligibility in people with amyotrophic lateral sclerosis (als)," in *CoDAS*, vol. 33. SciELO Brasil, 2021, p. e20190214.