When Whisper Listens to Aphasia: Advancing Robust Post-Stroke Speech Recognition

Anonymous submission to Interspeech 2024

Abstract

Despite recent advancements in Automatic Speech Recognition (ASR), its accuracy remains low for pathological speech, thereby limiting AI-based healthcare interventions in such settings. This work address this challenge by fine-tuning Whisper [1], an ASR known for its ability to capture high-dimensional features in healthy speech. Using a comprehensive dataset of Patients with Stroke (PwS), we fine-tuned Whisper and significantly reduced Word Error Rate (WER), surpassing previous work on severe aphasia. To demonstrate its generalisability, we tested the model on a separate database, AphasiaBank [2], and observed a lower WER despite variations in dialect, linguistics, and test protocols. Our result on the Aphasiabank was superior to previous ASRs trained on this database, confirming the generalisability of our approach. These outcomes not only address ASR limitations in impaired speech, but also establish the foundations for standardised and versatile AI solutions for remote speech monitoring for timely diagnosis and intervention.

Index Terms: Speech Recognition, Fine-tuning, Pathological Speech

1. Introduction

Aphasia is a communication disorder resulting from damage to specific areas of the brain responsible for the production and comprehension of language. In the United States, the National Aphasia Association estimates that between 2 and 4 million people live with aphasia [3], with another 350,000 affected individuals in the United Kingdom [4]. Although the disorder can be acquired after any brain injury, stroke remains the primary cause [5, 6], resulting in aphasia in 30% of stroke survivors [7], with a significant impact on recovery [8, 9].

Aphasia manifests as difficulty with different linguistic processes (semantics, fluency, phonology) in each individual with significant heterogeneity in the resulting impairment. Patients may experience word-finding difficulties, sentence comprehension or construction problems, and may also have additional cognitive deficits or motor speech issues, such as dysarthria or apraxia that affect overall language abilities. As a consequence, the heterogeneous nature and severity of aphasic symptoms not only hinder communication, but also have profound social and emotional effects, resulting in self-isolation and exclusion that significantly impact the quality of life of patients [10].

Given the high incidence of aphasia and its severe symptoms, it has extensive social and public health implications. The best evidence for rehabilitation is based on delivering intensive speech therapy, often requiring around 100 hours per month soon after a stroke [11]. Therefore, the process of administering therapy, as well as diagnosing and monitoring, demands substantial time and resources due to its inherent reliance on personalized clinician-patient engagement. This need for regular, extensive therapy sessions poses a significant logistical and financial challenge for healthcare providers [12, 13, 14].

Considering these challenges, there is a pressing need to develop accessible diagnostic and therapeutic tools that patients can use with minimal supervision. The integration of Automatic Speech Recognition (ASR) would enable clinicians to remotely monitor speech patterns more efficiently during the recovery process. By automating elements of language assessment and therapy, ASR would alleviate the burden on healthcare resources, reducing costs while improving access to timely and personalised care. However, progress in this domain has been hindered by the lack of diverse and clinically validated aphasic speech datasets that reflect the heterogeneity of aphasic impairments, significantly undermining the accuracy and so their suitability of ASR in healthcare applications [14, 15, 16].

The AphasiaBank project (see section 2.1) marked a significant milestone in this field. Several research groups [13, 15, 17, 18] developed different ASR algorithms trained on this AphasiaBank. Le et al. (2018) [13] successfully identified medically relevant quantitative measures to predict aphasia and achieved a 39% Word Error Rate (WER) for the ASR in spontaneous aphasic speech, while previous studies reported a 45% WER [15, 18]. However, it is essential to compare these rates with those of healthy speakers, which typically exhibit around 10% WER in studies that used attention-based architectures [19].

Previous investigations into aphasic speech recognition on aphasia employed conventional ASR architectures, characterised by distinct acoustic, linguistic, and pronunciation components. These studies were primarily oriented towards a hybrid system of Hidden Markov Model Deep Neural Networks (HMM-DNN) [13, 15, 20], or variations of Long Short-Term Memory (LSTM) models [18, 21, 22]. A Cantonese version of AphasiaBank has been implemented by Liu et al. (2018) [23]. In this case Multilayer Time Delay Neural Network (MT-DNN) with a Bidirectional Long Short-Term Memory (BLSTM) model structure was adopted. Such study obtained 18.5% of WER for unimpaired speech and 42.4% for impaired speech [24]. Tang et al. (2023) [25] achieved significant improvements in AphasiaBank WER using attention mechanisms like E-Branchformer and Conformers (combination of Convolutional Neural Networks and Transformers), reaching an average WER of 26% across different levels of aphasia severity.

By using attention mechanisms, in this paper we introduce a new contribution to the field by fine-tuning OpenAI's Whisper ASR [19], minimising the WER to a level more acceptable for clinical application compared to earlier attempts. To achieve this, we have created a dataset specifically designed to capture the diverse speech errors encountered in Patients with

Stroke (PwS), constituting a comprehensive quality-labeled corpus. This work represents a significant step towards the development of more efficient and accessible therapeutic interventions for PwS. In the following paragraphs, we discuss the implications and generalisability of our ASR model, highlighting its impact on patient care and outlining directions for future research in the field.

2. Methods

2.1. Dataset

A comprehensive aphasic speech dataset for application in clinical and scientific settings to train ASR systems for aphasia is being collected by our group. The corpus contains speech recordings of PwS who have participated in two longitudinal studies of stroke[1]. The database is composed of speech recordings from ≈1000 patients that describe the picture scene from the Comprehensive Aphasia Test (CAT) [26] and aims to be a substantial and useful resource for any other future research in these domains. It is complemented with quantitative assessments of speech data in English where speech therapists have transcribed audio recordings orthographically and phonetically using International Phonetic Alphabet (IPA).

The labeled dataset used for this study comprises 425 audio recordings of speech from 353 individuals with PwS, some of whom underwent repeated testing to capture recovery and intra-individual variability in speech. The total duration of the recordings was 10 hours. Additionally, we included an age-matched control data set ($\mu = 60.63$ years, $\sigma = 9.50$ years) with a total duration of one hour, which was used to establish the benchmark performance of ASR and enable a fair comparison between healthy and pathological speech.

We used the open-access AphasiaBank dataset [2] for an additional test set. The corpus contains data from 466 speakers, including narrative and procedural types of discourse. Content categorization is based on aphasia severity, evaluated using the WAB[2] scale which generates an Aphasia Quotient indicating mild, moderate, severe, or very severe aphasia [2, 28]. This diverse dataset offers a benchmark to assess the generalisability of our models across different contexts in terms of dialects, test protocols, and linguistic characteristics. It demonstrates the adaptability and effectiveness of our fine-tuned model in real-world scenarios beyond its original training domain.

2.2. ASR Architecture

Whisper uses an encoder-decoder Transformer model; the encoder processes 80-channel log-Mel spectrograms utilising two convolutional layers and sinusoidal positional encoding for efficient and context-aware audio representation. A stack of transformer blocks extracts long-range dependencies within these features. The decoder replicates this design, using learned positional embeddings and the same number of Transformer blocks, enabling a multi-task speech processing pipeline (such as transcription, translation or voice activity detection). Whisper's key strength lies in its enormous healthy speech training dataset of 680 000 hours [19]. This variegated training dataset entails various environments, recording setups, speakers, and languages (with approximately 20% being non-English). The language models available differs in sizes in terms of parameter count (from 39M to 1.55B) and Transformers layer depth (4 to 32). The encoder produces fixed-dimensional vectors, whose size (384 to 1280) increases with model capacity, while the temporal dimension (1500) remains constant. Whisper uses the same byte-level BPE text tokenizer as GPT-2 [29, 30], supporting both the English-only and multilingual models. The diversity of the dataset, along with weakly supervised labels, contributes to Whisper's near-human-level accuracy on healthy speech [19]. Its robust and generalisable potential might align with the atypical acoustic patterns and heterogeneity found in pathological speech.

2.3. Data Pre-processing

A team of two speech therapists and three trained postgraduate students conducted verbatim transcriptions of the audio recordings, resulting in a high level of agreement (73% inter-rater reliability at the word level). The transcriptions followed the formatting guidelines provided by Codes for the Human Analysis of Transcripts (CHAT) [31] and were processed using Computerised Language ANalysis software (CLAN) [32]. To handle special symbols used for error coding, such as those denoting semantic inconsistencies or speech fragments, the text was pre-processed to remove these symbols and punctuation. In instances of neologisms or vocalisations, phonetic alphabet representations were provided by the transcribers. These representations were then heuristically mapped to a sequence of phones in the Latin alphabet without altering their sequence [20].

Furthermore, human transcriptions included false starts and unique symbols for filler words like "er", "erm", and other isolated sounds or interjections specific to aphasic speech patterns. Previous experiments identified spelling variations in filler words between American and British English leading to an increase in WER due to differences in written linguistic practices. Specifically, British usage includes "er" and "erm", whereas American usage includes "uh" And "um". To address this issue, we standardised these filler words according to predominant American English conventions within our Whisper training dataset.

Subsequently, each conversation line was segmented to extract only the participants' dialogue from both human transcriptions and the audio file, excluding the assessor speech. This was possible with the manually inserted timestamps in the transcriptions and was later additionally verified with speech identification techniques through Speech Brain [33]. After converting all the files into *.wav* format and resampling at 16 kHz, audio files longer than 30 seconds were discarded to comply with Whisper's constraints and to avoid memory problems. Similarly, audio files shorter than 3 seconds were also excluded to prevent potential issues during the computation of Fourier transform for spectrograms generation or CTC layer alignment in the neural network [28]. Our approach to segment data by sentences, rather than fixed lengths, aimed to accommodate variable sentence lengths and mitigate overfitting in the training. In addition to Speech Brain, this processing utilised Python packages from FFmpeg [34], Pydub [1] and SoX [35].

2.4. Fine-Tuning

2.4.1. Training Data Allocation

After cleaning the database and pre-processing the interviewers' speech, we obtained a corpus of around 7 hours of audio data. The dataset was partitioned based on individual audio files with varying durations, resulting in a split of 78% for training

[1]To ensure author anonymity, the link and database references will be added after the review process

[2]Western Aphasia Battery [27]

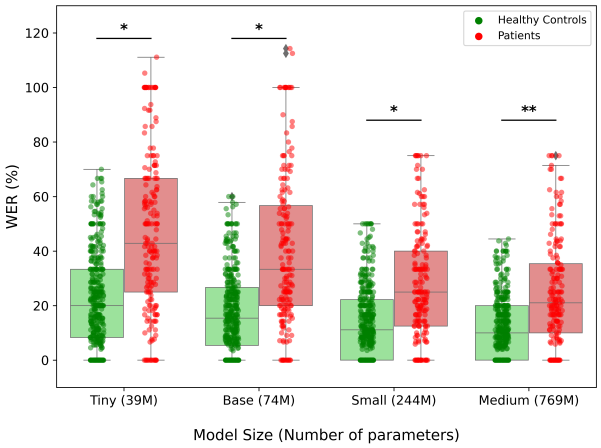

Figure 1: *Comparison of Word Error Rate (WER) of non-fine-tuned Whisper for different model sizes between Healthy Controls (in green) and Aphasia Patients (in red).* $^{***}: p < 0.001$; $^{**}: p < 0.01$; $^{*}: p < 0.05$.

(equivalent to 341 minutes), 13% for validation (equivalent to 57 minutes), and 9% designated as the unseen test set (equivalent to 40 minutes). Each division took into account a stratified splitting based on the severity of aphasia cases, ensuring a balanced representation of aphasia severity.

The held-out unseen test contained audio segments of patients that were not present in the training at all (i.e. even though every patient has multiple audio segments, none of these have been mixed in the training and validation). This selection ensured that the testing set was independent of the training and validation data in any possible way, thereby mitigating bias and allowing for a reliable assessment of model generalisation performance. The choice to treat each patient's data separately was crucial in our effort to develop a tool designed for zero-shot accurate transcription of individual patient speech.

2.4.2. *Training Configuration*

Using the off-the-shelf version of Whisper, the training procedure utilised a batch size of 16 per device, implementing gradient accumulation for efficient GPU resource utilisation. A cosine learning rate schedule was adopted, starting with an initial learning rate of 1×10^{-5} and incorporating 500 warmup steps. Training was optimised through AdamW, updating the entire model by the Cross-Entropy loss [36], defined as

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(\hat{y}_{ij}), \quad (1)$$

where N is the total number of samples or data points, C is the total number of classes or categories, y_{ij} is the true probability or label of class j for sample i and $\hat{y}_{ij}$ is the predicted probability of class j for sample i.

Model evaluation took place every 1000 steps, with the goal of minimizing the WER on the validation set. A final model was saved at each 1000-step interval, and its retention depended on demonstrating the best performance (i.e. lowest WER). Training continued until reaching a maximum of 6000 steps, utilizing mixed-precision (fp16) for efficiency and gradient checkpointing to manage memory overheads. The training was implemented using PyTorch [37] and Hugging Face Transformers repository [38], leveraging a NVIDIA RTX 6000 GPU. The *Tiny* model required less than three hours to train, while both the *Small* and *Base* models needed just over 3 hours; in contrast, training the *Medium* model took approximately 7 hours.

The WER metric was based on the string edit distance. This calculates the minimum number of steps required to convert the output from Whisper to the actual human transcription string. The WER between human and automated transcription was first calculated from the off-the-shelf Whisper version, generating a baseline performance. Such baseline was then compared with the fine-tuned WER. To assess the statistical significance, we performed non-parametric Mann-Whitney tests due to the robustness against non-normally distributed data. We also checked the statistical significance between the baseline performances of healthy controls and the unseen test set, and the results are detailed in the next section.

3. Results

Three main findings are presented in this section. Firstly, we show how the baseline performance of Whisper is significantly different between healthy controls and patients (Fig.1). Across all model sizes, Whisper performance resulted in higher WER in patients ($p < 0.05$) with the highest difference found using the *Tiny* model (+24.32%). These findings confirm that the worse baseline performances observed in our PwS dataset are intrinsically related to speech impairment characteristics and not due to audio quality concerns. It is important to note that some audio segments scored a WER larger than 100%. This is given by the calculation of this measure, which is the addition of insertions, deletions, and substitutions required to transform the reference transcription (the ground truth) into the hypothesis transcription (the model's output), divided by the number of total words. Therefore, if the number of operation to change the string exceeds the total number of words in the sentence, the resulting WER can indeed be greater than 100%.

Secondly, we focused on the impact of fine-tuning with the PwS training set on the held-out unseen set. When compared with their respective baseline, our analysis demonstrated robust statistically significant improvements for every model size (2^{nd} row of Table 1). The *Tiny* and *Small* model experienced a similar improvement in terms of WER (-4.3% and -3.5% respectively). However, the most substantial improvements were evident in the larger model. For the *Small* model, fine-tuning led to a 11.3 percentage reduction in WER, highlighting the model's enhanced capability to accurately transcribe PwS speech. Likewise, in the *Medium* dataset, the WER decreased by 10.9%, further affirming the efficacy of fine-tuning in adapting the model to the variability of post-stroke speech patterns.

Thirdly, in the assessment of the AphasiaBank dataset, aimed at the evaluation of the fine-tuned Whisper generalisability properties, promising results emerged across most model sizes. Particularly noteworthy were the improvements in the *Base*, *Small*, and *Medium* datasets, where the fine-tuned model obtained significant improvements when compared to their baseline version (3^{nd} row of Table 1). Specifically, the *Medium* model demonstrated exceptional robust generalization capabilities, achieving a substantial reduction of 14.3 points in WER, reducing from 35.8% to 21.5%. However, a notable exception was observed with the *Tiny* model on the AphasiaBank dataset. Here fine-tuning resulted in a slight degradation of WER performance by 3.6%. Such result highlights the challenge that smaller models face in effectively generalising across diverse test sets compared to their training data.

Dataset	Tiny *(39M)*		Base *(74M)*		Small *(244M)*		Medium *(769M)*	
	Baseline	*Fine-Tuned*	*Baseline*	*Fine-Tuned*	*Baseline*	*Fine-Tuned*	*Baseline*	*Fine-Tuned*
Validation (PwS)	62.8%	38.2%	42.8%	28.1%	55.4%	32.1%	37.1%	**25.9%**
Unseen (PwS)	36.6%	32.3%	38.5%	35.0%	26.8%	15.5%	25.6%	**14.7%**
AphasiaBank	40.3%	42.9%	38.1%	29.1%	33.8%	24.4%	35.8%	**21.5%**

Table 1: *Average Word Error Rate for different Whisper model sizes on various datasets, comparing fine-tuned and baseline performances. All fine-tuned models showed statistically significant improvements with respect to their baseline* ($p < 0.01$).

Overall, despite the exception of the *Tiny* model with the AphasiaBank, we experienced a success in the fine-tuning across multiple model sizes and dataset sources, highlighting the models' robustness in understanding the specific nuances of impaired speech while adapting to diverse linguistic contexts.

4. Discussion

By fine-tuning the state-of-the-art ASR with our dataset of aphasic speech, we were able to achieve the best performance of such models reported in the field thus far. More importantly, the analysis of the model performances on our PwS unseen dataset and AphasiaBank demonstrates generalization capabilities of the fine-tuned models of Whisper, a feature useful when applied in healthcare settings. The best WER achieved in this study surpasses outcomes reported in comparable investigations tested on AphasiaBank, as elaborated in Section 1. However, our results are particularly significant since our model was *not* trained on AphasiaBank but on a different dataset. This work also emphasises the trade-off between model complexity and generalisation performance, since smaller models revealed limitations in performances on contexts different from those used for training.

It is plausible that fine-tuning accurately enhances the model's ability to learn fluency and phonological deficits of PwS, thereby enabling the recognition of these patterns even in disparate datasets. In the fine-tuned version, improvements were observed in dysfluencies such as whole-word and whole-phrase repetition, syntactic errors, or filler-word usage, which were previously overlooked or transcribed incorrectly by the ASR. However, challenges persist, such as the skipping of intelligible words and patterns like false starts (e.g. "The k- kitum... the kitty"), as well as the frequent usage of filler words in patients which were, in the automated transcription, not always transcribed. Indeed, while the performance improvements with fine-tuning are significant, WER remains relatively high compared to the levels typically achieved in healthy speech recognition (around 10-13% WER [19]). Furthermore, the current performance, although improved compared to earlier published attempts, may not yet meet stringent requirements of clinical settings.

The baseline performance on the American AphasiaBank dataset is lower compared to our PwS validation dataset, likely due to Whisper's predominant training on American English speakers. This discrepancy highlights the importance of considering dataset-specific features when evaluating model performance across different datasets. Additionally, although the unseen set was chosen based on a balanced stratification of aphasia severity, the baseline WER is lower than that of the validation set. Therefore, interpreting a 14.7% WER (our best result) should be approached judiciously, considering the significant role played by variability in speakers' impairments and their diagnoses. This observation underscores the need for developing speech pathology tools that prioritize tailored assessments rather than standardized diagnoses - a shift crucial for achieving accurate recovery outcomes for each individual patient.

5. Limitation and Future Work

The results demonstrate how the improved WER represents a promising solution for efficiently using ASR tools in pathological speech recognition. However, it is important to acknowledge limitations that require consideration. Specifically, the *Large* model of Whisper was not tested in this study. With 1.5 billion parameters, such model could have offered further insights into the interplay between model complexity and performance. Further optimization of training parameters and the use of speech enhancement techniques such as noise reduction could improve the accuracy of ASR systems. Lastly, Although gathering 7 hours of data is a significant achievement for a pathological speech dataset, it remains relatively small for comprehensive fine-tuning purposes given its intended clinical application.

As our dataset continues to expand, we aim to enhance our training efforts by integrating AphasiaBank data and employing additional data augmentation techniques. Given that our dataset uniquely includes phonetically labeled data with IPA transcriptions, we plan to use this information to gain valuable insights into automatic detection of phonetic errors. This area has been relatively unexplored due to the scarcity of phonetic datasets and the challenges in adapting models trained on standard English text to IPA. Consequently, one of our future directions involves leveraging both phonetic and standard text transcriptions to develop a multimodal system capable of accurately assessing post-stroke speech.

6. Conclusion

The superior performance of our model in recognizing speech from diverse dialects and test-protocols in post-stroke pathological speech emphasizes its potential for healthcare settings. These speech recognition systems can be customized to provide therapeutic recommendations, reduce the workload on healthcare providers and enhance professionals' care delivery. This work represents a significant advancement in pathological speech recognition, laying the groundwork for transformative developments in AI-powered health interventions and the broader expansion of remote patient care.

7. References

[1] J. Robert, M. Webbie *et al.*, "Pydub," 2018. [Online]. Available: http://pydub.com/

[2] B. MacWhinney, D. Fromm, M. Forbes, and A. Holland, "Aphasiabank: Methods for studying discourse," *Aphasiology*, vol. 25, no. 11, pp. 1286–1307, 2011.

[3] "Aphasia statistics - national survey on aphasia awareness, national aphasia association website," accessed: 2024-01-2. [Online]. Available: http://www.aphasia.org/aphasia-resources/aphasia-therapy-guide/

[4] "Aphasia and its effects, stroke association website," accessed: 2024-01-2. [Online]. Available: https://www.stroke.org.uk/what-is-aphasia/aphasia-awareness

[5] A. Azhar, S. Maqbool, G. A. Butt, S. Iftikhar, and G. Iftikhar, "Frequency of aphasia and its symptoms in stroke patients," *J Speech Pathol Ther*, vol. 2, no. 2, p. 121, 2017.

[6] A. C. Laska, A. Hellblom, V. Murray, T. Kahan, and M. Von Arbin, "Aphasia in acute stroke and relation to outcome," *Journal of internal medicine*, vol. 249, no. 5, pp. 413–422, 2001.

[7] A. Grönberg, I. Henriksson, M. Stenman, and A. G. Lindgren, "Incidence of aphasia in ischemic stroke," *Neuroepidemiology*, vol. 56, no. 3, pp. 174–182, 2022.

[8] R. M. Lazar and A. K. Boehme, "Aphasia as a predictor of stroke outcome," *Current neurology and neuroscience reports*, vol. 17, no. 11, pp. 1–5, 2017.

[9] F. Geranmayeh, R. Leech, and R. J. Wise, "Network dysfunction predicts speech production after left hemisphere stroke," *Neurology*, vol. 86, no. 14, pp. 1296–1305, 2016.

[10] S. Spaccavento, A. Craca, M. Del Prete, R. Falcone, A. Colucci, A. Di Palma, and A. Loverre, "Quality of life measurement and outcome in aphasia," *Neuropsychiatric disease and treatment*, vol. 10, p. 27, 2014.

[11] M. C. Brady, H. Kelly, J. Godwin, P. Enderby, and P. Campbell, "Speech and language therapy for aphasia following stroke," *Cochrane database of systematic reviews*, no. 6, 2016.

[12] R. Palmer, P. Enderby, C. Cooper, N. Latimer, S. Julious, G. Paterson, M. Dimairo, S. Dixon, J. Mortley, R. Hilton *et al.*, "Computer therapy compared with usual care for people with long-standing aphasia poststroke: a pilot randomized controlled trial," *Stroke*, vol. 43, no. 7, pp. 1904–1911, 2012.

[13] D. Le, K. Licata, and E. M. Provost, "Automatic quantitative analysis of spontaneous aphasic speech," *Speech Communication*, vol. 100, pp. 1–12, 2018.

[14] S. S. Mahmoud, R. F. Pallaud, A. Kumar, S. Faisal, Y. Wang, and Q. Fang, "A comparative investigation of automatic speech recognition platforms for aphasia assessment batteries," *Sensors*, vol. 23, no. 2, p. 857, 2023.

[15] D. Le and E. M. Provost, "Improving automatic recognition of aphasic speech with aphasiabank." in *Interspeech*, 2016, pp. 2681–2685.

[16] K. C. Fraser, F. Rudzicz, N. Graham, and E. Rochon, "Automatic speech recognition in the diagnosis of primary progressive aphasia," in *Proceedings of the fourth workshop on speech and language processing for assistive technologies*, 2013, pp. 47–54.

[17] D. Le, K. Licata, C. Persad, and E. M. Provost, "Automatic assessment of speech intelligibility for individuals with aphasia," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 11, pp. 2187–2199, 2016.

[18] D. Le, K. Licata, and E. M. Provost, "Automatic paraphasia detection from aphasic speech: A preliminary study." in *Interspeech*, 2017, pp. 294–298.

[19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[20] M. Perez, Z. Aldeneh, and E. M. Provost, "Aphasic speech recognition using a mixture of speech intelligibility experts," *arXiv preprint arXiv:2008.10788*, 2020.

[21] D. S. Barbera, M. Huckvale, V. Fleming, E. Upton, H. Coley-Fisher, C. Doogan, I. Shaw, W. Latham, A. P. Leff, and J. Crinion, "Nuva: a naming utterance verifier for aphasia treatment," *Computer Speech & Language*, vol. 69, p. 101221, 2021.

[22] Y. Qin, T. Lee, S. Feng, and A. P.-H. Kong, "Automatic speech assessment for people with aphasia using tdnn-blstm with multi-task learning." in *Interspeech*, 2018, pp. 3418–3422.

[23] Y. Liu, Y. Qin, S. Feng, T. Lee, and P. Ching, "Disordered speech assessment using kullback-leibler divergence features with multi-task acoustic modeling," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018, pp. 61–65.

[24] G. Sanguedolce, P. A. Naylor, and F. Geranmayeh, "Uncovering the potential for a weakly supervised end-to-end model in recognising speech from patient with post-stroke aphasia," in *Proceedings of the 5th Clinical Natural Language Processing Workshop*, 2023, pp. 182–190.

[25] J. Tang, W. Chen, X. Chang, S. Watanabe, and B. MacWhinney, "A new benchmark of aphasia speech recognition and detection based on e-branchformer and multi-task learning," *arXiv preprint arXiv:2305.13331*, 2023.

[26] K. Swinburn, G. Porter, and D. Howard, "Comprehensive aphasia test," *APA PsycTests*, 2004.

[27] A. H. Risser and O. Spreen, "The western aphasia battery," *Journal of clinical and experimental neuropsychology*, vol. 7, no. 4, pp. 463–470, 1985.

[28] I. G. Torre, M. Romero, and A. Álvarez, "Improving aphasic speech recognition by using novel semi-supervised learning methods on Aphasiabank for English and Spanish," *Applied Sciences*, vol. 11, no. 19, p. 8872, 2021.

[29] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *arXiv preprint arXiv:1508.07909*, 2015.

[30] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[31] B. MacWhinney, *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press, 2014.

[32] G. Conti-Ramsden, "CLAN (Computerized Language Analysis)," *Child Language Teaching and Therapy*, vol. 12, no. 3, pp. 345–349, 1996.

[33] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong *et al.*, "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[34] S. Tomar, "Converting video formats with ffmpeg," *Linux Journal*, vol. 2006, no. 146, p. 10, 2006.

[35] "Sox sound exchange," accessed: 2023-10-20. [Online]. Available: http://sox.sourceforge.net

[36] A. Mao, M. Mohri, and Y. Zhong, "Cross-entropy loss functions: Theoretical analysis and applications," *arXiv preprint arXiv:2304.07288*, 2023.

[37] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.

[38] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.