INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

A New Benchmark of Aphasia Speech Recognition and Detection Based on E-Branchformer and Multi-task Learning

Jiyang Tang[1], *William Chen*[1], *Xuankai Chang*[1], *Shinji Watanabe*[1], *Brian MacWhinney*[2]

[1]Language Technologies Institute, Carnegie Mellon University, USA
[2]Department of Psychology, Carnegie Mellon University, USA

jiyangta@cs.cmu.edu

Abstract

Aphasia is a language disorder that affects the speaking ability of millions of patients. This paper presents a new benchmark for Aphasia speech recognition and detection tasks using state-of-the-art speech recognition techniques with the Aphsia-Bank dataset. Specifically, we introduce two multi-task learning methods based on the CTC/Attention architecture to perform both tasks simultaneously. Our system achieves state-of-the-art speaker-level detection accuracy (97.3%), and a relative WER reduction of 11% for moderate Aphasia patients. In addition, we demonstrate the generalizability of our approach by applying it to another disordered speech database, the DementiaBank Pitt corpus. We will make our all-in-one recipes and pre-trained model publicly available to facilitate reproducibility. Our standardized data preprocessing pipeline and open-source recipes enable researchers to compare results directly, promoting progress in disordered speech processing.

Index Terms: Disordered Speech Recognition, Assessment of Pathological Speech, Aphasia

1. Introduction

Aphasia is a language disorder that affects patients' abilities to communicate effectively. This condition can manifest in various components of the language, including phonology, grammar, and semantics, among others [1, 2]. Recent studies have developed machine learning methods for Aphasia speech recognition and detection to assist clinicians in the diagnosis and documentation process. The recognition task involves transcribing Aphasia speech into text, while the detection task requires classifying whether a speaker has Aphasia.

For the recognition task, various automatic speech recognition (ASR) architectures have been benchmarked on Aphasia speech data. A recent trend is using a pre-trained Wav2vec2.0 [3] to perform zero-shot or few-shot predictions for low-resource languages [4, 5]. Other benchmarked ASR models include DNN-HMM [6, 7] and RNN [8–10]. While some studies formulate the detection task as a binary classification problem [4, 11], others consider it as an Aphasia Quotient prediction task [8, 9, 12, 13]. Aphasia Quotient (AQ) is a metric used to measure the severity of Aphasia [14]. Linguistic statistics extracted from transcripts or ASR output are commonly used as input features. They include filler words per minute, pauses to words ratio, number of phones per word, and many more [4, 8, 9, 11–13]. Some researchers incorporate acoustic information as well since it also contains signs of Aphasia [9, 10, 12]. The classification or regression models used in these studies vary from classical machine learning models such as SVM [4, 8, 11, 12] to deep neural networks [10, 13, 15, 16].

Although several ASR systems have been tested in these studies, we believe performance can be further improved by leveraging recent state-of-the-art ASR architectures. Furthermore, as most existing Aphasia detectors require text as the input, an ASR system is required if the transcription is not available. Since ASR errors can cascade into the detection system, the detection accuracy might be suboptimal. Therefore, we aim to build an end-to-end system that can perform both tasks simultaneously using the latest ASR technologies. This system should be able to derive linguistic features from acoustic input implicitly and utilize both of them for the tasks.

To the best of our knowledge, we are the first to present an architecture that can detect the presence of Aphasia on both the sentence and the speaker level, while simultaneously transcribing the speech to text. Our system has two variants and achieves state-of-the-art detection performance on the AphasiaBank English subset. This is achieved with the help of the hybrid CTC/Attention ASR architecture [17], E-Branchformer [18], and WavLM [19]. Among existing studies, we found inconsistencies in evaluation metrics, data compositions, and preprocessing procedures. Therefore, we make our code and pre-trained model open-source in the hope of establishing a standardized benchmark environment for both tasks[1]. We demonstrate the effectiveness and generalizability of our approach by applying it to another disordered speech database, the DementiaBank Pitt corpus [20].

2. Method

In this section, we present a system that jointly models Aphasia detection and Aphasia speech recognition. The techniques used in this system have all been proven to be state-of-the-art in various speech processing tasks [17, 18, 21, 22].

2.1. Hybrid CTC/Attention

Our proposed method is based on the hybrid CTC/Attention ASR architecture [17]. This architecture comprises an encoder, denoted by $\mathrm{Enc}(\cdot)$, and a decoder, denoted by $\mathrm{Dec}(\cdot)$. The encoder captures the acoustic information and can optionally generate a text sequence using Connectionist Temporal Classification (CTC) [23]. The text sequence is primarily predicted by an attention-based decoder in an auto-regressive manner given the encoder's hidden states [17].

The input to the encoder, denoted as $\mathbf{X} = (\mathbf{x}_l \in \mathbb{R}^D | l = 1, \ldots, L)$, is a sequence of L acoustic feature vectors, where each vector has D dimensions. The ground truth text sequence is denoted as $T = (t_k \in V | k = 1, \ldots, K)$, which contains K text tokens from a vocabulary V. Using the CTC algorithm [23], the encoder predicts the likelihood of generating the

[1] https://github.com/espnet/espnet

10.21437/Interspeech.2023-2191

text sequence given the input $P_{\text{Enc}}(T|X)$. The encoder hidden state output is denoted as H:

$$\mathbf{H} = \text{Enc}(\mathbf{X}) \tag{1}$$

$$P(T|X) = \text{CTC}(\mathbf{H}) \tag{2}$$

The decoder models $P(T|X)$ given the encoder hidden states and prior token predictions [24]:

$$P(t_k|X, T_{1:k-1}) = \text{Dec}(\mathbf{H}, T_{1:k-1}) \tag{3}$$

$$P_{\text{Dec}}(T|X) \approx \prod_{k}^{K} P(t_k|X, T_{1:k-1}) \tag{4}$$

During training, the model is optimized using the weighted sum of the CTC loss and the decoder loss [17]:

$$\mathcal{L} = \lambda\mathcal{L}_{\text{CTC}} + (1-\lambda)\mathcal{L}_{\text{Dec}} \tag{5}$$

$$= -\lambda \log P_{\text{Enc}}(T|X) - (1-\lambda)\log P_{\text{Dec}}(T|X) \tag{6}$$

where the CTC weight λ is an hyper-parameter. The output of the encoder and decoder is jointly decoded using beam search to produce the final hypothesis during inference [17]. The system is often evaluated with word error rate (WER).

2.2. Intermediate CTC

Intermediate CTC (InterCTC) was proposed to regularize deep encoder networks and to support multi-task learning [22, 25, 26]. To achieve this, the existing CTC module is applied to the output of an intermediate encoder layer with index e. Then subsequent encoder layers incorporate the intermediate predictions into their input. Equation 1 can be reorganized as:

$$\mathbf{H}_e = \text{Enc}_{1:e}(\mathbf{X}) \tag{7}$$

$$P(Z_{\text{Inter}}|X) = \text{CTC}(\mathbf{H}_e) \tag{8}$$

$$\mathbf{H} = \text{Enc}_{e+1:E}(\text{NRM}(\mathbf{H}_e) + \text{LIN}(Z_{\text{Inter}})) \tag{9}$$

where E is the total number of encoder layers, and Z_{Inter} is the latent sequence of the InterCTC target sequence $T_{\text{Inter}} = (t'_k|k = 1, \ldots, K')$. $\text{NRM}(\cdot)$ and $\text{LIN}(\cdot)$ refer to a normalization layer and a linear layer respectively. The negative log likelihood of generating T_{Inter} is used as the InterCTC loss:

$$\mathcal{L}_{\text{Inter}} = -\log P_{\text{Inter}}(T_{\text{Inter}}|X) \tag{10}$$

The choice of T_{Inter} is dependent on the task. During training, the intermediate layer is optimized to correctly predict T_{Inter} by including $\mathcal{L}_{\text{Inter}}$ in the loss function:

$$\mathcal{L}'_{\text{CTC}} = \alpha\mathcal{L}_{\text{Inter}} + (1-\alpha)\mathcal{L}_{\text{CTC}} \tag{11}$$

where the InterCTC weight α is a hyper-parameter. The updated overall loss function is obtained by inserting Equation 11 into Equation 5:

$$\mathcal{L}' = \lambda\mathcal{L}'_{\text{CTC}} + (1-\lambda)\mathcal{L}_{\text{Dec}} \tag{12}$$

Note that it is possible to apply CTC to multiple encoder layers while having different target sequences for each. In that case, the average of all InterCTC losses is used as $\mathcal{L}_{\text{Inter}}$ [22, 26].

2.3. Speech Recognizer

Our speech recognizer follows the design of a hybrid CTC/Attention architecture described in Section 2.1. It transcribes the acoustic feature sequence $\mathbf{X}_{ij}$ belonging to speaker s_j to the corresponding text token sequence T_{ij}.

We experiment with recently proposed encoder architectures that enhance acoustic modeling ability over the original Transformer. One of these architectures, called Conformer, sequentially combines convolution and self-attention. This allows for capturing both the global and the local context of input sequences [27]. On the other hand, Branchformer models these contexts using parallel branches and merges them together. Both architectures demonstrate competitive performance in speech processing tasks [28]. In subsequent studies, E-Branchformer is proposed to enhance Branchformer further. It comprises a better method for merging the branches, and it achieves the new state-of-the-art ASR performance [18].

Meanwhile, self-supervised learning representation (SSLR) has been developed to improve the generalizability of acoustic feature extraction. SSLR leverages a large amount of unlabeled speech data to learn a universal representation from speech signals. Studies show significant performance improvement in ASR and other downstream tasks by using SSLR as the input of the encoder [19, 21, 29, 30].

2.4. Aphasia Detectors

We present two types of Aphasia detectors based on the speech recognizer, the tag-based detector and the InterCTC-based detector. Inspired by the use of language identifiers in multilingual speech recognition [31–33], we form an extended vocabulary V' by adding two Aphasia tag tokens to V:

$$V' = V \cup \{\texttt{[APH]}, \texttt{[NONAPH]}\} \tag{13}$$

We then train the ASR model using $T^{\text{tag}}_{ij} = (t_k \in V'|k = 1, \ldots, K)$ as the ground truth, where T^{tag}_{ij} is formed by inserting one or more Aphasia tags to T_{ij}. Specifically, `[APH]` is inserted if the speaker has Aphasia while `[NONAPH]` is inserted if the speaker is healthy. This method effectively trains the model to perform both tasks jointly. Moreover, the model leverages both linguistic and acoustic information to detect Aphasia, as the encoder first generates an initial tag prediction based on the acoustic features, and the decoder then refines the prediction based on prior textual context. During inference, the sentence-level prediction is obtained by taking out the tag token from the predicted sequence. Three tag insertion strategies will be tested in Section 3: prepending, appending, and using both. We note that all tag tokens are excluded from WER computation.

InterCTC is proven to be effective at identifying language identity in multilingual ASR as part of a multi-task objective. By conditioning on its language identity predictions, the ASR model achieves state-of-the-art performance on FLEURS [22]. Inspired by this, the second type of Aphasia detector uses InterCTC to classify input speech as either Aphasia or healthy speech. During training, the ground truth sequence T^{inter}_{ij} for InterCTC contains an Aphasia tag token. During inference, the prediction $\hat{y}_{ij}$ is generated by checking the tag produced by InterCTC greedy search. This approach allows us to select which encoder layer to use for the best speaker-level accuracy.

For both the tag-based and InterCTC-based detectors, the speaker-level Aphasia prediction y_j is obtained via majority voting of y_{ij} for all i.

3. Experiments

In this section, we first explore the impact of state-of-the-art encoder architectures and SSLR on Aphasia speech recognition. We then analyze the performance of the proposed method for recognition and detection tasks. All of our experiments were conducted using ESPnet2 [34].

3.1. Datasets

3.1.1. CHAT Transcripts

CHAT [35] is a standardized transcription format for describing conversational interactions, used by both AphasiaBank and DementiaBank. Besides the textual representations of spoken words, it includes a set of notations that describes non-speech sounds, paraphasias, phonology, morphology, and more. The transcript cleaning procedures differ between prior works, making it difficult to fairly compare their machine learning systems. Therefore, we derive a pipeline based on previous work [5] in the hope of standardizing this process for future research.

The specific steps of our pipeline are as follows. (1) Keep the textual representations of retracing, repetitions, filler words, phonological fragments, and IPA annotations while removing their markers. (2) Replace laughter markers with a special token <LAU>. (3) Remove pre-codes, postcodes, punctuations, comments, explanations, special utterance terminators, and special form markers (4) Remove markers of word errors, interruption, paralinguistics, pauses, overlap precedes, local events, gestures, and unrecognized words. (5) Remove all empty sentences after the above steps.

3.1.2. AphasiaBank and DementiaBank

AphasiaBank [36] is a popular speech corpus among the existing work. The dataset contains spontaneous conversations between investigators and Aphasia patients. It also includes conversations with healthy individuals as the control group. All experiments in this paper are performed using the English subset. Similar to [5], we obtain the training, validation, and test set by drawing 56%, 19%, and 25% percent of Aphasic speakers from each severity. There are four severity levels, each corresponding to a range of AQ scores: mild (AQ $>$ 75), moderate (50 $<$ AQ $\leq$ 75), severe (25 $<$ AQ $\leq$ 50), and very severe (0 $\leq$ AQ $\leq$ 25) [36]. The control group is split using the same ratio and merged with patients' data. Doing so ensures our data splits are representative across all severity levels. We then slice the recordings into sentences using the timestamps provided in the CHAT transcripts while cleaning them as described in Section 3.1.1. After that, sentences shorter than 0.3 seconds or longer than 30 seconds are removed. Before data augmentation, the training set contains 42.7 hours of patient data and 22.7 hours of control group data while the test data contains 20.1 and 10.1 hours. Details can be found in our code release.

Dementia speech recognition and detection have been a popular research topic as well [37–42]. We use the DementiaBank Pitt corpus [20] to test the generalizability of our design. Similar to recent studies [37, 38], we use the ADReSS challenge [43] test set, which is a subset of the DementiaBank Pitt corpus, for evaluation and the remaining data in the corpus for training and validation. We note that audio from the challenge test set has been enhanced with noise removal and volume normalization, while the transcripts have been preprocessed. To preserve a consistent data pipeline, we instead use the original recordings and transcripts from the Pitt corpus as our test data. Details can be found in our code base.

Model	Patient WER	Control WER	Overall WER
Baselines			
Conformer	40.3	35.3	38.1
E-Branchformer	36.2	31.2	34.0
Proposed Methods			
E-Branchformer+WavLM	26.4	17.0	22.2
+Tag-prepend	26.3	16.9	22.2
+Tag-append	**26.2**	16.9	**22.1**
+Tag-prepend/append	26.3	**16.8**	**22.1**
+InterCTC-6	26.3	16.9	**22.1**
+InterCTC-9	26.3	16.9	22.2
+InterCTC-6/Tag-prepend	26.3	16.9	**22.1**

Table 1: *Word error rate (WER) of proposed methods evaluated on AphasiaBank.*

3.2. Experimental Setups

Baseline: We first build two ASR systems using Conformer [27] and E-Branchformer [18], as described in Section 2.3. The Conformer encoder has 12 blocks, each having 2048 hidden units and 4 attention heads. The E-Branchformer encoder has 12 blocks, each with 1024 hidden units, and 4 attention heads. The cgMLP module has 3072 units and the convolution kernel size is 31. Both systems use a Transformer decoderwith 6 blocks, each having 2048 hidden units and 4 attention heads. The Conformer and E-Branchformer models have 44.2 and 45.7 million trainable parameters respectively. For the detection task, we reproduce the Aphasia detection experiment from a previous study. The detector is a support vector machine (SVM) that takes in linguistic features extracted from the oracle transcript to predict a binary classification label [4].
Proposed Method: We first build a system with learned acoustic representations extracted from WavLM [19] as the input to the E-Branchformer encoder. Using it as a foundation, we build tag-based and InterCTC-based detectors as described in Section 2.4. We also investigate the impact of tag insertion positions: prepending, appending, and both. Meanwhile, we apply InterCTC to the 6th and the 9th encoder layer respectively, and analyze their performance difference. We set both the CTC and InterCTC weight to 0.3 and the inference beam size to 10.

In all experiments, we use speed perturbation with ratios of 0.9 and 1.1, as well as SpecAugment [44], to augment the data. We choose the Adam optimizer with a learning rate of 10^{-3} and a weight decay of 10^{-6}. We employ `warmuplr` learning rate scheduler with 2500 warm-up steps and a gradient clipping of 1. Each final model is selected by averaging the 10 checkpoints with the highest validation accuracy out of 40 epochs. More details can be found in our code base.

3.3. Results and Discussion

Overall, the proposed systems achieve both accurate Aphasia speech recognition and detection at the same time. As shown in Table 1, switching from Conformer to E-Branchformer leads to a significant ASR performance improvement by 4.1 WER absolute. Adding WavLM reduces the WER further by 11.8. This proves the effectiveness of using a state-of-the-art ASR encoder and SSLR for Aphasia speech recognition. Surprisingly, both types of detectors lead to a slightly better ASR performance than the vanilla ASR model (0.1 WER reduction). This implies that the ASR predictions can be refined based on Aphasia detection results. We compare the ASR performance of our systems with previous work in detail in Table 2. Our systems obtained significant lower WER for mild, moderate, and severe patients,

Model	Metric	Patient					Control	Overall
		Overall	Mild	Moderate	Severe	Very severe		
DNN-HMM [6]	PER	-	47.4	52.8	61.0	75.8	-	-
DNN-HMM + MOE [45]	PER	36.8	33.1	41.6	62.9		-	-
Wav2vec2 (zero-shot) [4]	WER	56.0	-	-	-	-	37.5	47.1
BLSTM-RNN+i-Vector+LM [8]	WER	-	33.7	41.1	49.2	63.2	-	-
Wav2vec2 [5]	WER	-	23.6	36.8	36.4	**59.1**	-	-
E-Branchformer+WavLM								
+Tag-prepend	WER	**26.3**	22.3	32.8	**34.5**	72.5	**16.9**	22.2
+InterCTC-6	WER	**26.3**	22.3	**32.6**	34.7	71.7	**16.9**	**22.1**
+InterCTC-6/Tag-prepend	WER	**26.3**	**22.1**	32.9	34.8	73.3	**16.9**	**22.1**

Table 2: *The recognition word error rate of proposed methods and existing work on the AphasiaBank English subset. The metrics are phoneme error rate (PER) and word error rate (WER). Note that existing studies use different data splits than ours.*

Model	Accuracy	
	Sent	Spk
SVM [4]	-	96.2
E-Branchformer+WavLM		
+Tag-prepend	89.3	95.1
+Tag-append	89.2	95.1
+Tag-prepend/append	**90.8**	95.7
+InterCTC-6	85.2	**97.3**
+InterCTC-9	84.5	**97.3**
+InterCTC-6/Tag-prepend	89.7	96.7

Table 3: *Sentence-level (Sent) and speaker-level (Spk) detection accuracy of proposed methods on AphasiaBank. [4] is reproduced using the official code with oracle transcripts as the input. For +Tag-prepend/append and +InterCTC-6/Tag-prepend experiments, only the Tag-prepend output is reported since the difference is negligible.*

even against systems using an external language model. Despite this, they have a much higher WER for very severe Aphasia patients. We believe this is because hybrid CTC/Attention architectures are data-hungry, but the number of utterances and their average duration is much smaller for very severe patients.

From Table 3, we can see that the tag-based Aphasia detectors have the best sentence-level Aphasia detection accuracy. Interestingly, although the performance difference between prepending and appending Aphasia tags is insignificant, inserting at both positions leads to slightly better sentence-level and speaker-level accuracy. Meanwhile, the InterCTC-based detector at layer 6 achieves state-of-the-art speaker-level accuracy (97.3%), surpassing the SVM baseline. However, its sentence-level accuracy is lower than those of tag-based detectors. This corresponds to previous studies showing that middle encoder layers are more important to speaker-related tasks while the bottom layers are more relevant to ASR and related tasks [19, 30]. We also find that tag-based detectors produce significantly more false positives for speakers who do not have Aphasia but are less fluent than others, thus having a lower speaker-level accuracy. This implies that tag-based detectors are sometimes too sensitive to dysfluency.

Finally, more accurate tag-based predictions can be obtained by combining InterCTC and tag-prepending. This suggests that tag predictions are refined based on prior InterCTC predictions. A similar result is discovered in a previous study where the language identity predictions are more accurate by incorporating an InterCTC auxiliary task [22]. In addition, the combined model has higher sentence-level accuracy and lower speaker-level accuracy compared to its InterCTC counterpart, which demands future investigation.

Model	Patient	Control	Overall	Accuracy	
				Sent	Spk
Conformer [38]	-	-	29.7	-	-
Conformer [37]	-	-	25.5	-	91.7
E-Branchformer+WavLM					
+Tag-prepend	39.1	15.0	**24.8**	65.6	83.3
+InterCTC-6	39.6	15.0	25.1	61.3	77.1

Table 4: *Test result of proposed methods on DementiaBank. The metric for speech recognition is the word error rate (WER). The metrics for Dementia detection are sentence-level (Sent) and speaker-level (Spk) accuracy. Other studies [39–42] are not listed as their models are trained and tested on different data. Note that [37, 38] use a larger and cleaner training set.*

Table 4 shows evaluation results for DementiaBank. Although the overall WER is much lower than those in previous studies, Dementia detection accuracy is suboptimal. As we drew original recordings from the DementiaBank Pitt corpus, the audio is often noisy and has variable speaking volume. Consequently, the model is less effective at acoustic modeling, as seen by the decreased InterCTC detection accuracy. The results also suggest that linguistic features are more important for Dementia detection than Aphasia. Furthermore, majority voting for speaker-level predictions is less effective in this case as the number of sentences per speaker is typically between 5 to 20. Despite this, we believe our method has the potential to be adapted to other disordered speech in future studies.

4. Conclusion

In this paper, we build an all-in-one Aphasia speech recognition and detection system and test its performance using AphasiaBank and DementiaBank. We also standardize the data processing and model evaluation process to establish a public benchmark. Future studies are required to improve the recognition performance for severe Aphasia patients and the detection performance on DementiaBank. We can also further investigate the impact of joint learning and combining detector methods, and explore the potential benefits of fine-tuning a pre-trained healthy ASR system using disordered speech.

5. Acknowledgements

This work used the Bridges2 system at PSC and Delta system at NCSA through allocation CIS210014 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

6. References

[1] M. Danly and B. Shapiro, "Speech prosody in broca's aphasia," *Brain and Language*, vol. 16, no. 2, pp. 171–190, 1982.

[2] S. Ash *et al.*, "Speech errors in progressive non-fluent aphasia," en, *Brain and Language*, vol. 113, no. 1, pp. 13–20, 2010.

[3] A. Baevski *et al.*, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020.

[4] G. Chatzoudis *et al.*, "Zero-shot cross-lingual aphasia detection using automatic speech recognition," in *Proc. Interspeech*, 2022.

[5] 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, 2021.

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

[7] M. Perez, Z. Aldeneh, and E. Provost, "Aphasic speech recognition using a mixture of speech intelligibility experts," in *Proc. Interspeech*, 2020, pp. 4986–4990.

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

[9] Y. Qin, T. Lee, and A. Kong, "Automatic assessment of speech impairment in cantonese-speaking people with aphasia," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 331–345, 2020.

[10] Y. Qin *et al.*, "An end-to-end approach to automatic speech assessment for cantonese-speaking people with aphasia," *Journal of Signal Processing Systems*, vol. 92, pp. 819–830, 2019.

[11] A. Balagopalan *et al.*, "Cross-language aphasia detection using optimal transport domain adaptation," in *NeurIPS*, 2019.

[12] Y. Qin, T. Lee, and A. Kong, "Automatic speech assessment for aphasic patients based on syllable-level embedding and supra-segmental duration features," in *Proc. ICASSP*, 2018, pp. 5994–5998.

[13] Y. Qin *et al.*, "Automatic speech assessment for people with aphasia using TDNN-BLSTM with multi-task learning," in *Proc. Interspeech*, 2018, pp. 3418–3422.

[14] A. Kertesz, "Western aphasia battery–revised," 2007.

[15] Y. Qin *et al.*, "Aphasia detection for cantonese-speaking and mandarin-speaking patients using pre-trained language models," in *Proc. ISCSLP*, 2022, pp. 359–363.

[16] K. Dunfield and G. Neumann, "Automatic quantitative prediction of severity in fluent aphasia using sentence representation similarity," in *Proceedings of the RaPID Workshop.*, 2020.

[17] S. Watanabe *et al.*, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[18] K. Kim *et al.*, "E-branchformer: Branchformer with enhanced merging for speech recognition," in *Proc. SLT*, 2023, pp. 84–91.

[19] S. Chen *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2021.

[20] J. Becker *et al.*, "The Natural History of Alzheimer's Disease: Description of Study Cohort and Accuracy of Diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[21] S. Yang *et al.*, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

[22] W. Chen *et al.*, "Improving massively multilingual asr with auxiliary ctc objectives," in *Proc. ICASSP (in press)*, 2023.

[23] A. Graves *et al.*, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006, pp. 369–376.

[24] D. Povey *et al.*, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. Interspeech*, 2016, pp. 2751–2755.

[25] J. Lee and S. Watanabe, "Intermediate loss regularization for ctc-based speech recognition," in *Proc. ICASSP*, 2021, pp. 6224–6228.

[26] J. Nozaki and T. Komatsu, "Relaxing the conditional independence assumption of ctc-based ASR by conditioning on intermediate predictions," in *Proc. Interspeech*, 2021, pp. 3735–3739.

[27] A. Gulati *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[28] Y. Peng *et al.*, "Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding," in *Proc. ICML*, 2022, pp. 17 627–17 643.

[29] W. Hsu *et al.*, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, pp. 3451–3460, 2021.

[30] Y. Masuyama *et al.*, "End-to-end integration of speech recognition, dereverberation, beamforming, and self-supervised learning representation," in *Proc. SLT*, 2023, pp. 260–265.

[31] S. Toshniwal *et al.*, "Multilingual speech recognition with a single end-to-end model," in *Proc. ICASSP*, 2018, pp. 4904–4908.

[32] S. Watanabe, T. Hori, and J. R. Hershey, "Language independent end-to-end architecture for joint language identification and speech recognition," in *Proc. ASRU*, 2017, pp. 265–271.

[33] S. Zhou, S. Xu, and B. Xu, *Multilingual end-to-end speech recognition with a single transformer on low-resource languages*, 2018.

[34] S. Watanabe *et al.*, "Espnet: End-to-end speech processing toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.

[35] B. MacWhinney, "The childes project: Tools for analyzing talk," *Child Language Teaching and Therapy*, vol. 8, no. 2, pp. 217–218, 1992.

[36] B. MacWhinney *et al.*, "Aphasiabank: Methods for studying discourse," *Aphasiology*, vol. 25, no. 11, pp. 1286–1307, 2011, PMID: 22923879.

[37] T. Wang *et al.*, "Conformer based elderly speech recognition system for alzheimer's disease detection," in *Proc. Interspeech*, 2022.

[38] S. Hu *et al.*, "Exploiting cross-domain and cross-lingual ultrasound tongue imaging features for elderly and dysarthric speech recognition," *ArXiv*, vol. abs/2206.07327, 2022.

[39] R. B. Ammar and Y. Ayed, "Evaluation of acoustic features for early diagnosis of alzheimer disease," in *International Conference on Intelligent Systems Design and Applications*, 2019.

[40] F. Bertini *et al.*, "An automatic alzheimer's disease classifier based on spontaneous spoken english," *Computer Speech & Language*, vol. 72, p. 101 298, 2022.

[41] C. Bhat and S. Kopparapu, "Identification of alzheimer's disease using non-linguistic audio descriptors," in *Proc. EUSIPCO*, 2019, pp. 1–5.

[42] J. Chen, J. Zhu, and J. Ye, "An attention-based hybrid network for automatic detection of alzheimer's disease from narrative speech," in *Proc. Interspeech*, 2019, pp. 4085–4089.

[43] S. Luz *et al.*, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," in *Proc. Interspeech*, 2020, pp. 2172–2176.

[44] D. Park *et al.*, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[45] M. Perez, Z. Aldeneh, and E. Provost, "Aphasic speech recognition using a mixture of speech intelligibility experts," in *Proc. Interspeech*, 2020, pp. 4986–4990.