GENERATING PHONETIC TRANSCRIPTIONS FOR KOREAN ENGLISH L2 LEARNERS USING MULTIPLE SELF-SUPERVISED-MODEL-BASED ASR SYSTEMS AND ROVER METHOD

Jong In Kim* and Sunhee Kim† Minhwa Chung*‡
* Interdisciplinary Program in Cognitive Science, Seoul National University, Republic of Korea
E-mail: prows12@gmail.com Tel/Fax: +82-028809309
† Department of French Language Education, Seoul National University, Republic of Korea
E-mail: sunhkim@snu.ac.kr Tel/Fax: +82-028807693
‡ Department of Linguistics, Seoul National University, Republic of Korea
E-mail: mchung@snu.ac.kr Tel/Fax: +82-028809195

***Abstract*—In Computer-Assisted Pronunciation Training (CAPT), accurate phonetic transcriptions are essential for identifying mispronunciations in non-native speech corpora. Publicly available corpora for Korean English L2 learners often lack these transcriptions due to their limited size and availability. To address the shortage of accurate phonetic transcriptions, we propose a method that combines multiple Self-Supervised Learning (SSL)-based phone recognition systems with Recognizer Output Voting Error Reduction (ROVER). We trained SSL-based phone recognizers (Data2vec, Hubert, Wav2vec) on the Librispeech and CommonVoice datasets and used them to decode the L2arctic corpus. By applying ROVER, we achieved 85.5% accuracy in phone transcription compared to manual tagging. Additionally, an error analysis of 140 beginner-level sentences from the Korean Spoken English Corpus (NIA144) identified common pronunciation errors among Korean English speakers.**

I. Introduction

In recent years, interest in learning a second language (L2) has significantly increased. This growing enthusiasm for L2 acquisition is driven by its numerous benefits, including improvements in career prospects, educational and academic advancements, and the expansion of social opportunities. In foreign language learning, correct pronunciation is essential as it enhances communication, accurately conveys the speaker's intentions, and improves both listening skills and overall comprehensibility with native speakers.

Computer-Assisted Pronunciation Training (CAPT) plays a crucial role in second language (L2) acquisition. It provides explicit and real-time feedback, which motivates learners to practice pronunciation and supports self-directed learning. Additionally, CAPT allows learners to repeatedly practice specific mispronunciations at their own pace. Typically, a CAPT system includes Mispronunciation Detection and Diagnosis (MDD) and Pronunciation Scoring. MDD identifies and diagnoses specific pronunciation errors, delivering explicit feedback to learners. Pronunciation Scoring assesses the learner's pronunciation against native speaker benchmarks, offering quantitative evaluations that guide improvement.

In MDD, identifying patterns of pronunciation errors is crucial. This process provides personalized feedback to learners and monitors their pronunciation progress, thereby ensuring effective learning outcomes. The error patterns at the phone level can vary based on the learner's proficiency level and are influenced by the specific interactions between the learner's first language (L1) and the target second language (L2).

Accurate pronunciation assessment depends on identifying phone error patterns, yet comprehensive phoneme error datasets are severely lacking. As shown in Table 1, the challenge of limited resources for phonetic annotation in non-native English is evident. For instance, the TIMIT dataset, widely used for native English, includes only 5,232 phone annotations. Datasets for non-native English, such as L2Arctic [1] and Speechocean762 [2], provide even fewer annotations, with just 3,619 and 5,000 phone annotations, respectively. Furthermore, essential datasets for L2 Korean, like NIA144(Topic-Adaptive English Speaking Assessment Data for Korean Speakers), NIA037(English Speech Corpus of Korean Learners for Educational Use), which are large-scale Korean English speech corpora released by AIHub, a part of the National Information Society Agency (NIA) in Korea, lack phonetic transcriptions entirely. This scarcity of annotated data significantly impedes the development of effective Mispronunciation Detection and Diagnosis (MDD) systems

Manual phone transcription is challenging due to pronunciation variations caused by stress, accent, and prosody, as well as the complexity of different phone systems based on the speaker's L1 and L2 background. It requires extensive linguistic knowledge and is often time-consuming and costly, typically necessitating the expertise of phonetics professionals[3], [4]. These factors make accurate phone tagging difficult.

Dataset	L1/L2	Hrs	Sentences	Phone Label
TIMIT	L1	4	5,232	5,232
L2arctic	L2 Various Countries	11.2	11,026	3,619
Speechocean762	L2 Mandarin	6	50,00	5,000
Commonvoice v6	L1	2,182	1,821,529	x
Librispeech	L1	1,000	292,367	x
NIA012	L2 Korean	5,000	7,276,761	x
NIA037-SPK	L2 Korean	1016.81	34,340	x
NIA144	L2 Korean	300	66,889	x

TABLE I
A COMPARISON OF BASIC STATISTICS, FOCUSING ON THE DISTINCTIONS BETWEEN NONNATIVE AND NATIVE ENGLISH DATASETS; NIA012: KOREAN CHILDREN'S ENGLISH SPEECH DATA(AIHUB), NIA037-SPK: ENGLISH SPEAKING EVALUATION DATASET FOR KOREANS(AIHUB), NIA144: TOPIC-BASED ENGLISH SPEAKING EVALUATION DATASET FOR KOREANS(AIHUB)

A study indicates that the correlation for phoneme tagging is 0.88[5]. In contrast, automatic phone transcription offers several advantages. It is cost-effective, reducing the financial burden of manual annotation. It ensures consistent annotation across datasets, minimizing human error and variability. Additionally, automatic transcription systems can adapt to various L1-L2 combinations, making them versatile for different linguistic backgrounds. They also handle large corpora efficiently, a task that is impractical with manual tagging due to the time and labor involved. However, Automatic Phonetic Transcription (APT) is still not perfect due to the complexity and accuracy of ASR, human speech variability, limitations in current technology, and the need for high precision in certain applications. Given these challenges in automatic phonetic transcription, our study proposes a methodology to address these issues.

To tackle the challenges of automatic labeling, researchers have developed various techniques. Most studies in this area have focused on automatic tagging for speech recognition. For example, [6] introduced the momentum pseudo-labeling method, which integrates online and offline models. Online models predict pseudo-labels guided by offline models, iteratively improving performance. Similarly, [7] applied pseudo-labeling to 60 languages using multilingual concepts, fine-tuning models with pseudo-labels across different languages. To address noisy data in pseudo-labeling, [8] modified the training objective to detect incorrect labels, reducing errors. In pronunciation assessment, [9] proposed a zero-shot approach using the Hubert model, incorporating a transformer encoder, k-means clustering, and a scoring module to evaluate pronunciation. While sentence-level automatic tagging has been extensively studied, more thorough research is needed for phone-level automatic transcription.

To address this challenge, we propose an efficient method for automatically generating phonetic transcriptions for large non-native English corpora. Our approach utilizes multiple SSL-based phone recognizer systems in conjunction with the Recognizer Output Voting Error Reduction (ROVER) technique. ROVER enhances transcription robustness by reducing errors and improving accuracy. This approach enables the analysis of phone-level error patterns in non-native speech corpora.

The remainder of this paper describes our methods (Section 2), outlines the experimental conditions (Section 3), and discusses the results (Sections 4 and 5)

II. PROPOSED METHOD

A. Step1. Constructing the three types of SSL-based phone recognizer

In initial phase, we trained phone recognizer to obtain actual phone sequences. We assume that pronunciations by native English speakers adhere to standard pronunciation. Using available English corpora such as LibriSpeech and CommonVoice, we use audio transcriptions. We then apply a Grapheme-to-Phoneme (G2P) conversion to map written forms (graphemes) to their phonetic representations. Specifically, we train three type of SSL-based phone recognizers, including Data2vec[10], Hubert[11], and Wav2vec 2.0[12].

B. Step2. Actual Phonetic Transcription via Recognizer Output Voting Error Reduction (ROVER)

In the second phase, we derive the final phone sequence using Recognizer Output Voting Error Reduction (ROVER)[13]. ROVER combines outputs from multiple phone recognizers to produce a composite result, reducing the overall error rate compared to individual systems. In our experiment, ROVER was applied to the outputs of three phone recognizers to obtain high accuracy. The three phone recognizers used are Wav2vec-xlsr53, Hubert-large, Data2vec-large.

Given phone sequences $P_Hubert, P_Data2vec, P_Wav2vec$ from three phone reocgnizer systems:

$$P_Hubert = [p_{Hubert,1}, p_{Hubert,2}, \ldots, p_{Hubert,T_1}]$$

$$P_Data2vec = [p_{Data2vec,1}, p_{Data2vec,2}, \ldots, p_{Data2vec,T_1}]$$

$$P_Wav2vec = [p_{Wav2vec,1}, p_{Wav2vec,2}, \ldots, p_{Wav2vec,T_1}]$$

In alignment, each phone sequence is aligned to group matching phones at each time t. Algorithms such as dynamic programming with a word transition network (WTN) are used to perform this alignment.

$$Hubert = \text{Align}(P_1, P_2, \ldots, P_n) \rightarrow \{A_1, A_2, \ldots, A_T\}$$

$$Data2vec = \text{Align}(P_1, P_2, \ldots, P_n) \rightarrow \{A_1, A_2, \ldots, A_T\}$$

$$Wav2vec = \text{Align}(P_1, P_2, \ldots, P_n) \rightarrow \{A_1, A_2, \ldots, A_T\}$$

where A_t is the set of aligned phones at time t.

Subsequently, the voting module analyzes this data to generate the final transcription.

In Voting Process, at each time t the most frequent phone in the alignment set A_t. is selected. The voting metric follows the majority rule.

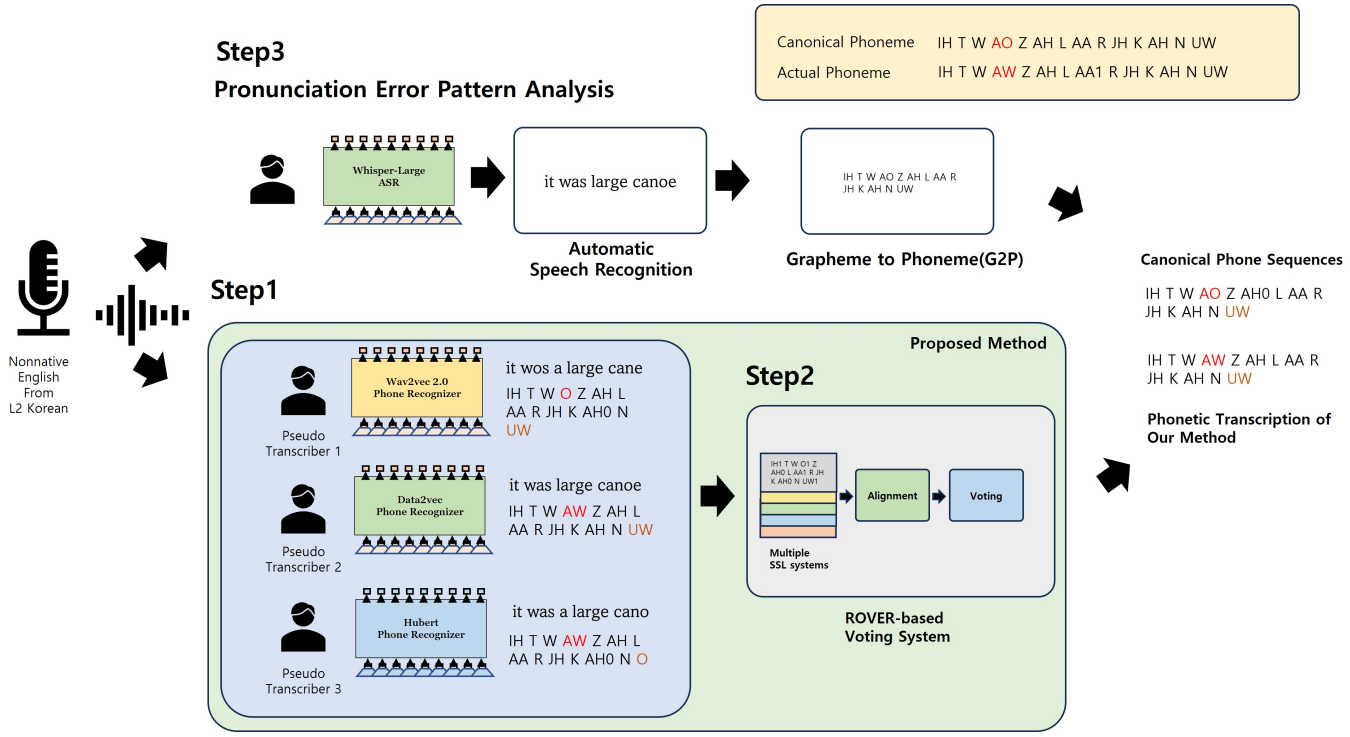

Fig. 1. The procedure of our ROVER-based phonetic transcription methodology. In Steps 1 and 2, we detail our proposed method for generating actual phone transcriptions. In Step 3, we analyze and present the pronunciation error patterns by comparing these actual phone transcriptions with the canonical transcription.

$$W_t = \arg\max_{p \in A_t} \sum_{i=1}^{n} \delta(p, P_i[t])$$

where $\delta(p, P_i[t])$ is a function that returns 1 if phone p is present at time t in the output of recognizer P_i, and 0 otherwise.

Combining the outputs of the three phone recognizers at each time t to generate the final phone sequence.

$$\text{Output} = \{P_1, P_2, \ldots, P_T\}$$

By combining the outputs of these modules, we unify shared characteristics among three phone recognizers, leading to a representation of phone annotations at each time t and generating the final phone transcription sequence

C. Step3. Phone-level Error Pattern Analysis

We adapt our automatic phonetic transcription method for phone-level error analysis. we use officially available Whisper-based ASR systems to convert raw audio into text, focusing on creating canonical phone sequences. Next, we apply a grapheme-to-phone (G2P) model to convert the decoded text into canonical phone sequences for comparing with the actual phone sequences generated by our method. We compare two sets of results in our analysis. By comparing the standard sequences with those generated by our method, we can identify error patterns. Our method assumes the availability of only raw audio files from L2 English speakers, with no textual information.

III. Experiments

A. Data

To train the phone recognizer, we use native English corpora, specifically CommonVoice[14] and LibriSpeech[15]. These corpora are chosen for their diverse range of spoken English styles, which enhances recognizer accuracy. The training set includes 1,951,560 utterances from CommonVoice and 562,480 utterances from LibriSpeech, totaling 2,514,042 utterances. CommonVoice v13 provides 1,366 hours of speech, while LibriSpeech offers 961 hours, amounting to a combined total of 2,327 hours of read speech.

We convert the transcribed text into ARPA-style phone representations using the Montreal Forced Aligner with the English US ARPA v2.00 G2P models. We use 70 ARPA phone types, which are categorized into two groups: one group excludes stress markers (e.g., AA0, AA1, AA2 grouped as AA), and the other group includes stress markers with two-group (specifically, (AA0, AA1) into AA, and AA2).

B. SSL-Based Phone Recognizers

We use the fairseq toolkits to build SSL-based phone recognizers: Data2vec[10], Hubert[11], and Wav2vec 2.0[12]. We apply the default training settings provided by fairseq. We train three type of SSL based phone recognizers: Wav2vec-xlsr53, Hubert-large, Data2vec-large. For decoding, we use the flashlight toolkit to perform Viterbi decoding without a language model.

phoneme recognizer	Validation set (PER)	Testset (PER)
Data2vec-large	5.395	7.373
Hubert-large	6.968	9.915
Wav2vec xlsr-53	5.846	7.83

TABLE II
ASSESSING THE PERFORMANCE OF PHONE RECOGNIZER SYSTEM WITH SSL (PER : PHONE ERROR RATE)

C. ROVER-based phone transcription

To use ROVER, segmental boundaries (time alignment) are required. We obtain these boundaries by performing forced alignment with the Wav2Vec acoustic model. We then use the SCTK toolkit to integrate phone recognizers through the ROVER method. we use three recognizers to obtain final phone sequences in the ROVER system.

D. Evaluation

We verify the performance of Automatic Phonetic Transcription. We compare the manual transcription of an existing publicly available L2 Arctic dataset with our proposed automatic transcription, calculating alignment accuracy.

E. Analyzing Error Patterns

The L2Arctic corpus [1] is a collection of nonnative English speech data designed for mispronunciation detection. It includes recordings from 24 nonnative English speakers, with a balanced gender distribution and 150 utterances per participant. We use the L2Arctic dataset to assess the performance of our phonetic transcription method by comparing the phone transcriptions generated by our system with the manual annotations for Korean speakers in the dataset.

To further investigate our methodology, we used the NIA144 dataset to analyze and compare phone-based error patterns with our methodology. The NIA144 corpus includes 450 hours of recordings from L2 Korean speakers, categorized into five proficiency levels. We selected 140 cases from the lowest proficiency level (Level 1) to analyze the resulting error patterns. If the error patterns identified by our method align closely with common errors in Korean English, it indicates the reliability of our method.

IV. RESULTS

A. Performance of an SSL-Based Phone Recognizer System via phone error rate(PER)

In Table 2, Wav2vec-xlsr-53 and Data2vec-large exhibit lower PER values for both the validation and test sets, indicating more accurate phone recognition compared to other models. Notably, Data2vec-large achieves the lowest validation PER (5.395), demonstrating good performance in self-supervised learning on the English corpus.

	ROVER(Recognizer Output Voting Error Reduction)
Data2vec-large	85.4%/85.2%
Hubert-large	85.0%/84.7%
Wav2vec-xlsr53	84.9%/84.6%
ROVER Accuracy	**85.5%/85.3%**

TABLE III
COMPARISON OF OUR METHOD AND L2 ARCTIC HUMAN ANNOTATION ON KOREAN SPOKEN ENGLISH, BOTH WITH AND WITHOUT STRESS MARKER(ACCURACY)

	ROVER(D2vlarge+Hubert-large+W2vxlsr53)
Arabic	84.2%
Hindi	83.8%
Mandarin	81.0%
Spanish	80.3%
Vietnamese	78.1%
Korean	85.5%

TABLE IV
PERFORMING AN EXTENSIVE ANALYSIS OF ROVER(DATAVEC-LARGE + HUBERT-LARGE + WAV2VEC-XLSR53) TO EVALUATE THE ALIGNMENT ACCURACY OF OUR PHONETIC TRANSCRIPTION METHOD IN COMPARISON TO THE MANUAL ANNOTATION IN L2ARCTIC, ENCOMPASSING MULTIPLE COUNTRIES(ACCURACY)

B. Comparison of our method and L2 Arctic human annotation on L2 English from various country, both with and without stress

In Table 4, the model demonstrates the highest accuracy (85.5%) for Korean, highlighting its effectiveness in recognizing phones for Korean English L2 learners. Arabic and Hindi also show high accuracy, indicating strong performance with the phonetics of these languages. Mandarin and Spanish follow with slightly lower accuracy, likely due to their tonal nature and unique phonological features. Vietnamese has the lowest accuracy among the languages tested, suggesting challenges in capturing its phonetic nuances.

C. Analyzing Error Patterns

As shown in Table 3, the results highlight the performance of the ROVER models. The ROVER method achieves an accuracy of 85.5% without stress markers and 85.3% with stress markers

In Table 5, comparing our phonetic transcription results with the manual annotations from L2Arctic for L2 Korean reveals variations in pronunciation, particularly with consonants. The findings include changes such as converting the open 'AH' sound to the fricative 'TH', altering the sibilant 'S' to its voiced counterpart 'Z', replacing 'D' with the voiced dental fricative 'DH', and interchanging the tense 'IY' with the lax 'IH'

We adapt our transcription method to the NIA144 dataset. We examined error patterns by analyzing 140 instances of Korean spoken English in NIA144 using our phone labeling method. In Table 6, we evaluate the efficacy of our system by comparing the error patterns identified through our phone annotation method with common consonant errors in non-native English speech by L2 Korean learners. Our findings align with the patterns documented in [5], including errors such as (/z/, /s/), (/dh/, /d/), (/v/, /h/), and (/f/, /p/)

	Substitutions	Deletion	Addition
1	'AH','TH'	'D'	'D'
2	'S', 'Z'	'DH'	'T'
3	'D', 'DH'	'IY'	'R'
4	'IY', 'IH'	'AH'	'HH'
5	'AO', 'OW'	'T'	'AO'
6	'AE', 'EH'	'AX'	'AH'
7	'AE', 'AH'	'S'	'K'
8	'IY', 'TH'	'Y'	'DH'
9	'Z', 'S'	'IH'	'TH'
10	'UW','OW'	'UW'	'W'

TABLE V
EXAMINING THE PATTERNS OF ERRORS PRESENT IN OUR PHONETIC TRANSCRIPTION WITH THE MANUAL ANNOTATIONS ON KOREAN SPOKEN ENGLISH

	Substitutions	reference [3]	Deletion	Addition
1	**'F', 'P'**	**/z/, /s/**	'T'	'T'
2	**'Z', 'S'**	**/dh/,/d/**	'N'	'N'
3	'L', 'R'	**/v/,/b/**	'D'	'D'
4	'D', 'T'	**/f/,/p/**	'S'	'S'
5	**'DH', 'D'**		'L'	'M'
6	**'S', 'Z'**		'R'	'L'
7	**'V', 'B'**		'W'	'R'
8	'B', 'P'		'Z'	'K'
9	'G', 'K'		'K'	'TH'
10	'T', 'D'		'Y'	'W'

TABLE VI
EXAMINING ERROR PATTERNS IN OUR PROPOSED PHONE ANNOTATION TO ENGLISH NIA144 DATASETS SPOKEN BY KOREAN: A COMPREHENSIVE INVESTIGATION OF 140 CASE; DATA2VEC-LARGE + HUBERT-LARGE + WAV2VEC-XLSR53

V. DISCUSSION

Our research focuses on developing automated phonetic transcription for non-native English speakers, particularly L2 Korean learners. While our system did not achieve high accuracy, the experiments have provided valuable insights into handling the issue of sparse phone annotations.

The accuracy of our phonetic transcription is reliable, especially when compared to previous research on non-native L2 Korean speakers. Earlier studies show an 88% agreement among human transcribers [5]. Despite a 3% margin of error, our transcription accuracy of 85.5% demonstrates a strong consensus and reliability in comparison with human transcriber agreement

Analyzing error patterns in phonetic transcription using real, unlabeled data from 140 cases in the NIA144 dataset reveals similarities with the error patterns found in previous research on Korean English learners [5]. These results suggest that our method is effective for transcribing the speech of non-native English speakers, especially Korean learners. Additionally, our method shows promise for analyzing pronunciation errors in other language pairs, particularly for minority languages where manual phonetic transcription is challenging.

VI. CONCLUSION

In summary, we have addressed the challenges by combining a self-supervised learning-based phone recognizer with ROVER. Our experiments using the L2arctic corpus for Korean learners achieved an 85.5% accuracy in phone recognition. This consistent accuracy and the analysis of error patterns on unlabeled corpora demonstrate our method's effectiveness for transcribing extensive non-native English speech. Future work will focus on adapting our approach for different non-native English corpora and conducting thorough error analyses to better understand phonetic transcription challenges.

REFERENCES

[1] G. Zhao, S. Sonsaat, A. Silpachai, *et al.*, "L2-arctic: A non-native english speech corpus.," in *Interspeech*, 2018, pp. 2783–2787.

[2] J. Zhang, Z. Zhang, Y. Wang, *et al.*, "Speechocean762: An open-source non-native english speech corpus for pronunciation assessment," *arXiv preprint arXiv:2104.01378*, 2021.

[3] L. Yang, J. Zhang, and T. Shinozaki, "Self-supervised learning with multi-target contrastive coding for non-native acoustic modeling of mispronunciation verification," *Proc. Interspeech 2022*, pp. 4312–4316, 2022.

[4] D. Zhang, A. Ganesan, S. Campbell, and D. Korzekwa, "L2-gen: A neural phoneme paraphrasing approach to l2 speech synthesis for mispronunciation diagnosis," 2022.

[5] H. Hong, S. Kim, and M. Chung, "A corpus-based analysis of english segments produced by korean learners," *Journal of Phonetics*, vol. 46, pp. 52–67, 2014.

[6] Y. Higuchi, N. Moritz, J. L. Roux, and T. Hori, "Momentum pseudo-labeling for semi-supervised speech recognition," *arXiv preprint arXiv:2106.08922*, 2021.

[7] L. Lugosch, T. Likhomanenko, G. Synnaeve, and R. Collobert, "Pseudo-labeling for massively multilingual speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7687–7691.

[8] H. Zhu, D. Gao, G. Cheng, D. Povey, P. Zhang, and Y. Yan, "Alternative pseudo-labeling for semi-supervised automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[9] Y. Chen, W. Wang, and C. Wang, "Semi-supervised asr by end-to-end self-training," *arXiv preprint arXiv:2001.09128*, 2020.

[10] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*, PMLR, 2022, pp. 1298–1312.

[11] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[12] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[13] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, IEEE, 1997, pp. 347–354.

[14] R. Ardila, M. Branson, K. Davis, *et al.*, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[15] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.