Towards automatic spoken grammatical error correction of L2 learners of English

Stefano Bannò[1,2,*], Michela Rais[3] and Marco Matassoni[2]

[1]Department of Psychology and Cognitive Science, University of Trento, Corso Bettini 84, Rovereto (TN), 38068, Italy
[2]Fondazione Bruno Kessler, via Sommarive 18, Trento, 38123, Italy
[3]Center for Mind/Brain Sciences, University of Trento, Corso Bettini 31, Rovereto (TN), 38068, Italy

Abstract

The demand for learning English as a second language (L2) has been growing consistently over the past decades, as it has become the lingua franca of culture, entertainment, business, and academia. In this regard, mastering grammar is one of the key elements of L2 proficiency.

In this paper, we illustrate an approach to spoken grammatical error correction (GEC) in a cascaded fashion using only publicly available training data. Specifically, we start from learners' utterances, investigate disfluency detection (DD) and removal, and finally explore GEC. Despite using only publicly available data, we achieve promising results that are aligned with previous studies which leveraged a large proprietary dataset. We discuss these results and reflect on some open issues and challenges of spoken GEC.

Keywords

computer-assisted language learning, spoken grammatical error correction, disfluency detection, L2 assessment and feedback

1. Introduction

With the rise of English as the global language of culture, entertainment, business, and academia, the ability to speak it fluently has become increasingly valued and the demand for learning English as a second language (L2) has been consistently increasing over the past decades [1]. This has resulted in a growing interest in automated approaches to evaluate spoken language proficiency for applications in Computer-Assisted Language Learning (CALL) for both individual practice and classroom settings, as well as to certify proficiency in language exams.

In particular, the assessment of learners' grammar through grammatical error correction (GEC) has attracted considerable attention over the past years. While text-based GEC has become an established area of study [2, 3], spoken GEC is still a relatively new area of research, mainly due to the limited availability of specifically designed and annotated data [4]. Assessing spoken grammar requires several adjustments to standard GEC models as these tend not to generalize to speech. Spoken GEC (see Table 2) is in fact more challenging than written GEC (see Table 1) as spoken grammar tends to be more flexible and less encoded than written grammar [5]. L2 spoken grammar is often characterized by disfluencies, naturally occurring speech events such as pauses, false starts and self-corrections, as well as errors which might differ from the ones made by L2 learners in written texts. As a result, spoken GEC cannot be easily performed with end-to-end systems but is usually implemented in a cascaded fashion consisting of three different modules. First, an automatic speech recognition (ASR) module is used to transcribe the spoken text. This is followed by a disfluency detection (DD) and removal module, which eliminates interruptions and repetitions in the speech. Finally, a spoken GEC system is applied. Recently, we have investigated the use of an end-to-end based on self-supervised learning (SSL) representations to predict the scores related to grammatical correctness of L2 English learners' utterances [6], but, to the best of our knowledge, SSL has not been explored for spoken grammatical error detection or correction.

Following the approach of [4], this paper employs transformer-based models both for DD and spoken GEC and shows that spoken GEC performance can be significantly improved through the application of disfluency detection and that such improvements can be achieved by using publicly available data for the training of the two modules.

2. Data

We exclusively used publicly available data for training our models, which we tested on a the TLT-GEC, a subset of the TLT corpus, a small proprietary corpus of young Italian learners of English presented in [7]. For the DD module training we employed two corpora, the NICT-

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy
*Corresponding author.
sbanno@fbk.eu (S. Bannò); michela.rais@studenti.unitn.it (M. Rais); matasso@fbk.eu (M. Matassoni)
0000-0002-2799-0601 (S. Bannò); 0009-0006-5873-8894 (M. Rais); 0000-0002-9689-1316 (M. Matassoni)

Original	He see the thief is catched by policeman the last night.
Corrected	He saw the thief caught by a policeman last night.

Table 1
Example of written GEC.

Original	uhm he see the the thief is catched by policeman the la- last night
Corrected	he saw the thief caught by a policeman last night

Table 2
Example of spoken GEC.

JLE and KIT Speaking Test Corpus. For the training of the spoken GEC model we used the second release of EFCAMDAT [8, 9, 10] and multiple corpora from the BEA-2019 Shared Task, a task focused on GEC which was organised as part of the the Workshop on Innovative Use of NLP for Building Educational Applications [11].

2.1. NICT-JLE

The National Institute of Information and Communications Technology - Japanese Learner English (NICT-JLE) corpus, originally introduced in [12], is a collection of manual transcriptions of approximately 300 hours of oral interviews of Japanese learners of English which does not include the original audio recordings.[1] A subset of the corpus was manually annotated with disfluencies as well as grammatical errors which were corrected. Furthermore, this subset includes annotations about proficiency scores ranging from A1 to B2 of the Common European Framework of Reference (CEFR) [13].

2.2. KIT Speaking Test Corpus

The Kyoto Institute of Technology (KIT) Speaking Test Corpus, released for public use by [14] consists of manual transcriptions of approximately 4,448 hours of interviews of 574 Japanese undergraduate students.[2] As in the case of NICT-JLE, the corpus does not include the original audio recordings. The manual annotations follow the tagging system employed in the NICT-JLE corpus, however these only include disfluencies, whereas grammatical errors are not annotated. The proficiency level of the students approximately ranges from CEFR level A1 to B2.

2.3. EFCAMDAT

EFCAMDAT is one of the largest publicly available L2 learner corpus and consists of 1,180,310 scripts written by 174,743 L2 learners.[3] The scripts are annotated with POS tags and information on grammatical dependencies, and are partially error-tagged by human experts. After excluding noisy responses and incorrect annotations, we kept 762,475 responses from which we removed punctuation and capitalisation in order to make them more similar to speech transcriptions. We used spaCy[4] to extract pairs of parallel sentences (i.e., original versus correct) from which we removed sentences shorter than 4 words as well as those containing broken XML tags and manual annotations on word limit. Following [15], we further excluded parallel sentences where the token edit distance is higher than 60% of the length of the original sentence in order to guarantee consistency between the original sentences and their corrected counterparts.

2.4. BEA-2019

The corpora from the BEA 2019 shared task are text-based corpora tagged with GEC annotations.[5]
CLC-FCE: the Cambridge Learner Corpus - First Certificate English (CLC-FCE) [16] is a publicly available section of the larger proprietary Cambridge Learner Corpus (CLC) [17] consisting of 1244 FCE exam scripts.[6]
Write & Improve: it is a dataset derived from Write & Improve with Cambridge, an online platform where L2 learners of English can practise their writing skills [18].[7]
LOCNESS: it is a section of the the Louvain Corpus of Native English Essays (LOCNESS), consisting of 100 essays written by L1 English undergraduates from the United Kingdom and the United States [19].
Lang-8: The Lang-8 Corpus of Learner English is a dataset extracted from the Lang-8 website,[8] whose users are encouraged to correct each other's grammar [20, 21].
NUCLE: The National University of Singapore Corpus of Learner English (NUCLE) is a collection of 1,400

[1] alaginrc.nict.go.jp/nict_jle/index_E.html#license
[2] kitstcorpus.jp/
[3] philarion.mml.cam.ac.uk/
[4] spacy.io
[5] cl.cam.ac.uk/research/nl/bea2019st/#data
[6] ilexir.co.uk/datasets/index.html
[7] writeandimprove.com/
[8] lang-8.com/

essays written by Asian undergraduate students at the National University of Singapore [22].

Including EFCAMDAT, the data used for training the spoken GEC system amount to 2,552,825 sentences, which we randomly split into a training set of 2,527,296 and a development set of 25,529 sentences.

As a benchmark for assessing the performance of spoken GEC system we employed the same test set of the CLC-FCE corpus used in previous studies ([23, 4]) with punctuation and capitalisation removed.

2.5. TLT-GEC

The TLT-GEC is a small proprietary dataset of speech utterances of young Italian learners of English which we have manually annotated with disfluencies and two sets of grammatical error corrections performed by two different human annotators. The dataset is derived from the larger TLT-school corpus presented by [7] and contains 1127 sentences for a total of 4.96 hours. The CEFR proficiency levels of the speakers are approximately A2 and B1. The data was split into two sets, a development set of 605 sentences and a test set of 522 sentences with non-overlapping speakers. The ASR transcriptions were obtained through a Conformer model, made available by NVIDIA in the popular NeMo toolkit [9]. The Conformer architecture [24] effectively combines self-attention layers and convolutions blocks to learn simultaneously global and local local correlations; this variant uses a decoder based on CTC loss instead of a standard RNNT/Transducer, substituting the auto-regressive LSTM component with a simpler linear decoder. The word error rate (WER) is 24.72% considering both development and test sets.

3. Disfluency detection

We performed DD as a sequence tagging task using a BERT-based [25] token classifier:

$$\mathbf{d}_{1:M} = \mathrm{BERT}(w_{1:M}) \quad p(r_m|w_{1:M}) = f_d(\mathbf{d}_m)$$

where r_m is a binary tag which indicates whether word w_m is fluent or disfluent. Subsequently, all words classified as disfluencies are removed from the transcriptions. Table 3 considers the example previously shown in Table 2 and clarifies each passage once again.

Specifically, the BERT-based model consists of a BERT layer in the version provided by the HuggingFace Transformer Library [26] (*bert-base-uncased*), a dropout layer, a dense layer of 768 nodes, a dropout layer, another dense layer of 128 nodes, and finally the output layer. The model is trained on NICT-JLE and KIT Speaking Test Corpus and uses an Adam optimiser [27] with batch size 64, learning rate 1e-06, dropout rate 0.2, and negative log likelihood as loss.

For evaluation, we use precision, recall, and F_1 scores. Table 4 shows the results of the DD model on the test and development sets of TLT-GEC in terms of precision, recall and F_1 score.

4. GEC

For the GEC model, we used a T5 model [28] initialised from the version provided by the HuggingFace Transformer Library [26] (*t5-base*) trained on EFCAMDAT and BEA-2019 with the exclusion of the CLC-FCE test set, that we used to compare the results on TLT-GEC. We set the maximum sequence length to 64 using an AdamW optimiser [29] with learning rate 1e-5, batch size 32.

To evaluate the performance of our model, we use two common metrics for GEC, i.e., MaxMatch (M^2) score [30] and General Language Evaluation Understanding (GLEU) metric [31]. The former computes the F-score of edits over the optimal phrasal alignment between the hypothesis and the reference sentences, whereas the latter is inspired by BLEU [32] and captures grammatical corrections as well as fluency rewrites.

In Table 5, we report the results of the spoken GEC system on the TLT-GEC test set in terms of M^2 and GLEU. For further comparison, we also report the results of our model on the CLC-FCE test and we compare them to the results of the GEC model described in [4]. We also report the agreement between the two human annotators.

Considering the performance on CLC-FCE test set, it can be observed that our proposed model performs moderately better than the model from [4]. These results are quite remarkable, given that we used only publicly available data, whereas [4] employed the entire CLC corpus in addition to the BEA-2019 data.

For completeness, we report the results on TLT-school considering the performance of the GEC model on the manual transcriptions with disfluencies (dsf), with disfluencies manually removed (flt), and with disfluencies automatically removed (autoflt). As expected, there is a remarkable improvement both in terms of GLEU and M^2 when disfluencies are removed from the transcriptions. Finally, we report the performance of our GEC system on ASR transcriptions. It can be observed that also in this case removing disfluencies improves the performance for both metrics. It also noticeable that the performance on the ASR transcriptions (autoflt) is slightly better than the one on manual transcriptions (dsf) in terms of GLEU.

[9]https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_conformer_ctc_large

Disfluent	**uhm** he see **the** the thief is catched by policeman the **la-** last night
Fluent	he see the thief is catched by policeman the last night
Corrected	he saw the thief caught by a policeman last night

Table 3
DD + spoken GEC. The disfluencies are indicated in bold.

	Precision ↑	Recall ↑	F_1 ↑
TLT-GEC dev	83.27	87.05	85.12
TLT-GEC test	80.94	83.93	82.41

Table 4
Results of DD on the TLT-GEC development and test sets in terms of Precision, Recall, and F_1 Score.

		GLEU ↑	M^2 ↑
CLC-FCE test	**Our model**	70.05	57.86
	[4]	-	56.60
TLT-GEC test (manual transcriptions)	**Agreement**	80.32	79.86
	dsf	35.73	49.11
	flt	66.44	65.81
	autoflt	58.89	57.65
TLT-GEC test (ASR transcriptions)	**dsf**	33.85	39.23
	autoflt	38.35	40.45

Table 5
Results of GEC on CLC-FCE test set and TLT-GEC test set (manual and ASR transcriptions) in terms of M^2 and GLEU (**dsf** = transcriptions with disfluencies; **flt** = transcriptions with disfluencies manually removed; **autoflt** = transcriptions with disfluencies automatically removed).

5. Conclusions and future works

In this paper, we explored an approach to automatic spoken grammatical error correction of Italian learners of English using only publicly available training data.

First, we investigated DD. Our DD module achieved a good performance in terms of Precision, Recall and F_1 score on both the development and test sets of the TLT-GEC.

The second module of our cascaded framework is a spoken GEC system which achieves results aligned with previous studies. As we expected, we found that disfluency removal has a positive impact on GEC on both manual and ASR transcriptions of the TLT-GEC. Furthermore, we observed that the fully automated system (i.e., ASR+DD+GEC) achieves higher results than the system including manual transcriptions with disfluencies in terms of GLEU.

Although we identified disfluencies as problematic elements for spoken GEC and we investigated an efficient way to detect and remove them, we acknowledge that there are still several open problems which are particularly evident in the TLT-GEC data. Specifically, the presence of code-switched words is a challenging issue, as can be seen in the following example drawn from the data (manual transcriptions):[10]

hello my name is giovanni uhm and i'm from trento and i live in rovereto uhm rovereto is in nord italien uhm uhm and uhm hobby uhm f- f- my favourite hobby uhm is uhm football and and koch

As can be observed, not only does the answer feature Italian names and toponyms, but it also contains German code-switched words. The output of the GEC system after automatically removing the disfluencies is the following:

hello my name is giovanni and i'm from trento and i live in trento it is in north italien my favourite hobby is football and cooking

It appears to handle the code-switched words *nord* and *koch* quite efficiently, but it fails to correct *italien.* [11]

Therefore, future works will attempt to address the problem of named entities recognition and code-switching in the framework of spoken GEC.

Another interesting problem concerns the relevance of learners' answers to the question prompts. For example, one of the question prompts is:

What country would you like to visit in the future? Why?

A sample answer drawn from the data is the following:

i like to visit turkey because i like speaking the language [...]

Although the answer is grammatically correct if considered individually, it does, in fact, contain a verbal error in relation to the question prompt. We also plan to address this issue starting from concatenating the question

[10] We only changed the first name and one toponym due to privacy reasons, but the example is still valid.

[11] In fact, it also does not correct the agreement error *hobby is football and cooking*, which should feature *hobbies are* instead of *hobby is.*

prompt with the learner's answer.

Finally, we plan to investigate an SSL-based approach (e.g., using wav2vec 2.0 [33] or more recent models such as HuBERT [34] or WavLM [35]) to spoken GEC. Specifically, it would be interesting to generate synthetic audio data using a text-to-speech system on the written learner corpora we used in this paper for training our models.

References

[1] P. Howson, The English effect, British Council, London, 2013.

[2] Y. Wang, Y. Wang, K. Dang, J. Liu, Z. Liu, A comprehensive survey of grammatical error correction, ACM Transactions on Intelligent Systems and Technology (TIST) 12 (2021) 1–51. doi:10.1145/3474840.

[3] C. Bryant, Z. Yuan, M. R. Qorib, H. Cao, H. T. Ng, T. Briscoe, Grammatical error correction: A survey of the state of the art, arXiv preprint arXiv:2211.05166 (2022). doi:10.48550/arXiv.2211.05166.

[4] Y. Lu, S. Bannò, M. J. F. Gales, On assessing and developing spoken 'grammatical error correction' systems, in: Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), Association for Computational Linguistics, Seattle, Washington, 2022, pp. 51–60. doi:10.18653/v1/2022.bea-1.9.

[5] M. McCarthy, R. Carter, Ten criteria for a spoken grammar, in: Explorations in corpus linguistics, Cambridge University Press, 2006, pp. 27–52.

[6] S. Bannò, M. Matassoni, Proficiency assessment of L2 spoken English using wav2vec 2.0, in: 2022 IEEE Spoken Language Technology Workshop (SLT), 2023, pp. 1088–1095. doi:10.1109/SLT54892.2023.10023019.

[7] R. Gretter, M. Matassoni, S. Bannò, D. Falavigna, TLT-school: a corpus of non native children speech, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 378–385. URL: https://aclanthology.org/2020.lrec-1.47.

[8] J. Geertzen, T. Alexopoulou, A. Korhonen, Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT), in: Proceedings of the 31st Second Language Research Forum, Cascadilla Proceedings Project, Somerville, 2013, pp. 240–254. URL: http://www.lingref.com/cpp/slrf/2012/paper3100.pdf.

[9] Y. Huang, J. Geertzen, R. Baker, A. Korhonen, T. Alexopoulou, The EF Cambridge Open Language Database (EFCAMDAT): Information for users, 2017.

[10] Y. Huang, A. Murakami, T. Alexopoulou, A. Korhonen, Dependency parsing of learner English, International Journal of Corpus Linguistics 23 (2018) 28–54. doi:10.1075/ijcl.16080.hua.

[11] C. Bryant, M. Felice, Ø. E. Andersen, T. Briscoe, The BEA-2019 shared task on grammatical error correction, in: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Florence, Italy, 2019, pp. 52–75. doi:10.18653/v1/W19-4406.

[12] E. Izumi, K. Uchimoto, H. Isahara, The NICT JLE corpus: Exploiting the language learners' speech database for research and education, International journal of the computer, the internet and management 12 (2004) 119–125.

[13] Council of Europe, Common European Framework of Reference for Languages: Learning, Teaching, Assessment, Cambridge University Press, Cambridge, 2001. URL: https://rm.coe.int/1680459f97.

[14] K. Kanzawa, H. Mitsunaga, G. Edmonds, Y. Hato, Y. Tsubota, M. Mori, Y. Shimizu, Development and administration of a Skype-based English speaking test in a Japanese high school, Bulletin of Kyoto Institute of Technology 14 (2022) 27–47.

[15] Y.-C. Lo, J.-J. Chen, C. Yang, J. Chang, Cool English: a grammatical error correction system based on large learner corpora, in: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Santa Fe, New Mexico, 2018, pp. 82–85. URL: https://aclanthology.org/C18-2018.

[16] H. Yannakoudakis, T. Briscoe, B. Medlock, A new dataset and method for automatically grading ESOL texts, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 180–189. URL: https://aclanthology.org/P11-1019.

[17] D. Nicholls, The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT, in: Proceedings of the Corpus Linguistics 2003 Conference, 2003, pp. 572–581.

[18] H. Yannakoudakis, Ø. E. Andersen, A. Geranpayeh, T. Briscoe, D. Nicholls, Developing an automated writing placement system for ESL learners, Applied Measurement in Education 31 (2018) 251–267. doi:10.1080/08957347.2018.1464447.

[19] S. Granger, The computer learner corpus: a versatile new source of data for SLA research, in: S. Granger (Ed.), Learner English on computer, Routledge, London, 1998, pp. 3–18. doi:10.4324/

9781315841342.

[20] T. Mizumoto, Y. Hayashibe, M. Komachi, M. Nagata, Y. Matsumoto, The effect of learner corpus size in grammatical error correction of ESL writings, in: Proceedings of COLING 2012: Posters, The COLING 2012 Organizing Committee, Mumbai, India, 2012, pp. 863–872. URL: https://aclanthology.org/C12-2084.

[21] T. Tajiri, M. Komachi, Y. Matsumoto, Tense and aspect error correction for ESL learners using global context, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Jeju Island, Korea, 2012, pp. 198–202. URL: https://aclanthology.org/P12-2039.

[22] D. Dahlmeier, H. T. Ng, S. M. Wu, Building a large annotated corpus of learner English: The NUS corpus of learner English, in: Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, 2013, pp. 22–31.

[23] Y. Fathullah, M. Gales, A. Malinin, Ensemble distillation approaches for grammatical error correction, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 2745–2749. doi:10.1109/ICASSP39728.2021.9413385.

[24] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, et al., Conformer: Convolution-augmented transformer for speech recognition, arXiv preprint arXiv:2005.08100 (2020).

[25] J. Devlin, M. Chang, L. Kenton, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv e-prints (2018) arXiv:1810.04805. doi:10.48550/arXiv.1810.04805.

[26] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. doi:10.18653/v1/2020.emnlp-demos.6.

[27] D. Kingma, J. Ba, Adam: a method for stochastic optimization, in: International Conference on Learning Representations, 2014.

[28] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, et al., Exploring the limits of transfer learning with a unified text-to-text transformer., Journal of Machine Learning Research 21 (2020) 1–67.

[29] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations 2019, 2019.

[30] D. Dahlmeier, H. T. Ng, Better evaluation for grammatical error correction, in: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Montréal, Canada, 2012, pp. 568–572. URL: https://aclanthology.org/N12-1067.

[31] C. Napoles, K. Sakaguchi, M. Post, J. Tetreault, Ground truth for grammatical error correction metrics, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Beijing, China, 2015, pp. 588–593. doi:10.3115/v1/P15-2097.

[32] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[33] A. Baevski, H. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, in: NeurIPS 2020, 2020, pp. 1–12.

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

[35] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al., WavLM: Large-scale self-supervised pre-training for full stack speech processing, IEEE Journal of Selected Topics in Signal Processing (2022). doi:10.1109/JSTSP.2022.3188113.