INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Self-supervised End-to-End ASR for Low Resource L2 Swedish

Ragheb Al-Ghezi [1], Yaroslav Getman [1*], Aku Rouhe [1], Raili Hildén[2], Mikko Kurimo[1]*

[1] Aalto University, Finland
[2] University of Helsinki, Finland
first.last@aalto.fi, first.last@helsinki.fi

Abstract

Unlike traditional (hybrid) Automatic Speech Recognition (ASR), end-to-end ASR systems simplify the training procedure by directly mapping acoustic features to sequences of graphemes or characters, thereby eliminating the need for specialized acoustic, language, or pronunciation models. However, one drawback of end-to-end ASR systems is that they require more training data than conventional ASR systems to achieve similar word error rate (WER). This makes it difficult to develop ASR systems for tasks where transcribed target data is limited such as developing ASR for Second Language (L2) speakers of Swedish. Nonetheless, recent advancements in self-supervised acoustic learning, manifested in wav2vec models [1, 2, 3], leverage the available untranscribed speech data to provide compact acoustic representation that can achieve low WER when incorporated in end-to-end systems. To this end, we experiment with several monolingual and cross-lingual self-supervised acoustic models to develop end-to-end ASR system for L2 Swedish. Even though our test is very small, it indicates that these systems are competitive in performance with traditional ASR pipeline. Our best model seems to reduce the WER by 7% relative to our traditional ASR baseline trained on the same target data.

Index Terms: Self-supervised, End-to-End L2 ASR, Non-native ASR

1. Introduction

Current state-of-the-art automatic speech recognition systems (ASR) have achieved impeccable performance that matches that of human transcribers in some tasks [4, 5, 6]. However, their performance deteriorates significantly when applied to the speech of non-native speakers and second language (L2) learners [7, 8] due to issues related to word mispronunciation, ungrammaticality, and disfluency [9, 10]. In the context of computer-assisted speaking assessment, developing a high-performant ASR system is crucial, and in order to build highly accurate ASR systems, a large amount of transcribed speech data should be available. While this might not be an issue for languages with high numbers of language learners such as English and Spanish, it is certainly a challenge for languages with fewer learners such as Swedish and Finnish.

Due to data scarcity, low resource L2-ASR systems are usually developed using a traditional ASR pipeline in which customized engineering solutions applied to each stage in the pipeline in order to improvement performance. For example, in pronunciation modeling, each lexical item has multiple pronunciations in order to accommodate L2 speakers' mispronunciation of words. In some cases, customized solutions to pronunciation and language modeling are either difficult to implement, requiring specialized linguistic expertise, or cost-ineffective. Therefore, developing end-to-end L2-ASR systems that do not require separate pronunciation or external language modeling is highly desirable, yet prohibitive because end-to-end systems require large amount of transcribed data.

On the other hand, self-supervised learning has emerged as an effective technique for settings where labeled target data is scarce. The key idea is to learn general representations in a setup where substantial amounts of unlabeled source data are available and thereby leveraging them to improve the performance on a downstream target task for which the amount of labeled data is limited. This is particularly interesting for tasks where substantial effort is required to obtain labeled data, such as speech recognition. In computer vision, representations for ImageNet have proven to be useful to initialize models for tasks such as image captioning or pose estimation. Unsupervised pre-training for computer vision has also shown promise. In NLP, self-supervised pre-trained language models improved many tasks such as text classification, phrase structure parsing, and machine translation.

To this end, this paper examines the generalizability of self-supervised monolingual and cross-lingual pre-trained models to low resource end-to-end L2-ASR for Swedish. Our contribution is two-fold. First, to the best of our knowledge we are the first to develop L2-ASR in a completely end-to-end fashion, without using an external language model nor pronunciation dictionary. L2-ASR is typically developed using traditional pipeline ASR where the custom design of lexicon and language modeling is required due to L2 data scarcity. Second, we demonstrate that the self-supervised pre-trained method for speech can adapt to low resource L2 Swedish using a small amount of fine-tuning data. Furthermore, we describe a new L2 Finland-Swedish speech data set collected and prepared at Aalto University and University of Helsinki.

2. Related Work

Several studies propose different approaches to mitigate the lack of training data issue in low-resource speech recognition system such as data augmentation [11, 12] and cross-lingual transfer learning for acoustic modeling [13]. One approach to improve the performance of L2-ASR is to train a DNN acoustic model on a transcribed L1 speech corpus and then adapt the model by freezing the hidden layers and only update the output layer with transcribed L2 speech corpus [14]. Another study, [15], proposes a domain adaption approach involving two native speech corpora (Japanese and English) for developing L2-ASR for Japanese learners. The rationale behind the approach is that features extracted from a model trained jointly on speakers' mother-tongue language (Japanese) and target language (English) provide richer acoustic characteristics of the language learners. On a pronunciation level, some approaches involve creating specialized pronunciation lexicons containing multiple pronunciations for each lexical item in order to accommodate the variations in pronunciation or mispronunciation in L2 speech [16].

[*] equal contribution

http://dx.doi.org/10.21437/Interspeech.2021-1710

One prominent work on unsupervised acoustic representation is wav2vec [1] which aims to represent audio data by solving context-prediction task in a self-supervised manner with an objective function similar to that in word2vec [17]. The model uses an encoder-decoder CNN architecture in which the encoder produces a representation z_i for each time step i and the decoder combines multiple encoded time steps into a combined learning representation c_i for each time step i. Another work is vq-wav2vec [2] which improves the previous work by learning quantized representations of audio data using a future time-step prediction task. The quantization training procedure brings several advantages among which are (1) significant reduction in training time without affecting the performance, and (2) discretization of audio samples. Results of wav2vec works show that applying off-the-shelf acoustic vectors pre-trained using these two methods lead to competing performance in WER in speech recognition tasks.

Discretization also allows a wide range of language model techniques such as token-masking, adopted in BERT [18], to be utilized. For example, implementing a BERT-like pretraining procedure on a dataset like Librispeech not only generates unsupervised contextual acoustic representation for English, but evidence from [19] shows that fine-tuning these vector representations for an out-of-domain task is also possible with small data. In fact, the same study suggests the possibility of having speech recognition systems trained on a near-zero amount of transcribed data. A study [20] suggests that weak supervision is another method that can significantly improve performance. Using a weakly supervised encoder for CTC fine-tuning can reduce WER by 7%. This is later confirmed by wav2vec2.0 [3] which, with the help of a quantizer and a transformer architecture, takes advantage of the token-masking strategy to provide vector representation of speech signal. It has also been shown that such models are even capable of representing cross-lingual knowledge of human speech [21].

Furthermore, the investigation of these pre-trained vector representations, in [22], shows that the pre-trained quantized vector representations of vowel phonemes have locations in latent space similar to those shown in the IPA vowel chart defined by human experts. This means that these latent representations are capable of representing speech data phonetically, and therefore can be used in end-to-end ASR systems as well as other non-ASR applications such as pronunciation and fluency assessment.

3. Data

The data used in this work were collected as a part of DigiTala project [23] which aims to develop automatic tools for the assessment of spoken Finnish and Finland-Swedish languages in order to reduce the work overload of human raters during the Finnish national high school matriculation examination. The Swedish speech data were collected from Finnish high school students responding to free-form and read-aloud speaking tasks. The speech data will become available for research via the Language Bank of Finland[1] after the additional data collection and human assessments of the oral skill levels are finished by summer 2021.

In this paper, only the speech samples from the free-form tasks were used for experiments. This subset (see Table 1) consists of 4777 transcribed recordings with a total duration of 1156 minutes. The samples were collected from 341 students with the speaking skill level varying mostly between A1 and B2 and answering from 1 to 19 free-form tasks. The total amount of different free-form tasks is 26. The responses were recorded in a normal classroom setting with several students speaking simultaneously, so the speech data include some background noise.

To test the ASR performance, a small set of 57 speech samples (roughly 10 minutes) from 7 speakers were chosen covering all tasks except the pair discussion. Due to the small overall size of the dataset, it was not reasonable to remove all overlap between speakers and tasks in the train and test sets.

4. Methodology

In this section, we briefly summarize the architecture design of wav2vec 2.0 model [3] and its pre-training procedure. We also discuss how the pre-trained model can be fine-tuned and utilized in an end-to-end ASR pipeline.

4.1. Pre-training

Wav2vec 2.0 uses a convolutional feature encoder $f : \mathcal{X} \mapsto \mathcal{Z}$ to map raw audio sample $\mathcal{X}$ to a sequence of T latent speech representations $\mathbf{z}_1, \ldots, \mathbf{z}_T$. Next, a Transformer network $g : \mathcal{Z} \mapsto \mathcal{C}$ is used to encode these latent representations into contextual representations $\mathbf{c}_1, \ldots, \mathbf{c}_T$ following a masking strategy similar to that of BERT language model [18]. Because feature encoder representations $\mathcal{Z}$ are continuous, they need to be discretized to $\mathbf{q}_1, \ldots, \mathbf{q}_T$ with a quantization module $\mathcal{Z} \mapsto \mathcal{Q}$ before they can be fed to the transformer network.

Wav2vec2 learns latent representation of speech data [3] by optimizing a combined loss $\mathcal{L}$ as a function of a contrastive loss $\mathcal{L}_m$ and a diversity loss $\mathcal{L}_d$ weighted by a constant α:

$$\mathcal{L} = \mathcal{L}_m + \alpha \mathcal{L}_d \tag{1}$$

The contrastive loss, $\mathcal{L}_m$, aims to distinguish the true quantized latent representation $\mathbf{q}$ from a set of negative candidates $\tilde{\mathbf{q}} \in \mathbf{Q}_t$. These negative latent representation are sampled from other masked time steps. In this case of monolingual models, the negative candidates are sampled from the same utterance. However, in a multilingual case, they are sampled from utterances of other languages. The contrastive loss is defined as

$$\mathcal{L}_m = -\log \frac{\exp\left(\operatorname{sim}\left(\mathbf{c}_t, \mathbf{q}_t\right)\right)}{\sum_{\tilde{\mathbf{q}} \sim \mathbf{Q}_t} \exp\left(\operatorname{sim}\left(\mathbf{c}_t, \tilde{\mathbf{q}}\right)\right)} \tag{2}$$

where sim denotes the normalized dot product between the contextual representation $\mathbf{c}$ and $\mathbf{q}$.

Negative and positive sample,V, are stored in a code-book, G, and the goal of the diversity loss is to encourage the model to use all the entries in the code-book. It does so by maximizing the entropy H of the averaged softmax distribution over the code-book entries for each group $\bar{p}_g$ across a batch of utterances

$$\mathcal{L}_d = \frac{1}{GV} \sum_{g=1}^{G} -H\left(\bar{p}_g\right) \tag{3}$$

4.2. Finetuning

After pre-training on a large corpus of unlabelled speech, the transformer model is fine-tuned on the target (labeled) speech data by appending a randomly initialized classification layer to predict the characters or the graphemes. In our experiments, we have 35 output tokens, one for each letter in the Swedish alphabet in addition to a word boundary token and a few special tokens. The feature encoder is not trained during the fine-tuning,

[1] https://www.kielipankki.fi/language-bank/

Table 1: *Statistics (number of questions, number of responses, duration in minutes) for each task in the free-form speech subset, as well as for test and training sets used in this work and in total.*

Task Description	**No. of Subtasks**	**No. of Responses**	**Duration, min.**
General speaking situations	10	1290	216
Image description	6	1513	484
Question answering, short responses	5	1088	131
Question answering, long responses	1	150	40
Question asking	3	735	283
Pair discussion	1	1	2
Smaller training set	25	2287	273.6
Larger training set	25	4442	957.3
Full training set	26	4720	1146.3
Test set	20	57	9.5
Total	26	4777	1156

and a masking strategy similar to SpecAugment [24] is applied to the output of the feature encoder. The model is altogether optimized by minimizing a Connectionist Temporal Classification (CTC) loss using Adam optimizer with adaptive learning rate.

5. Experiments

For the purposes of this paper, we use pre-trained monolingual and multilingual self-supervised speech models in an end-to-end ASR pipeline. We also compare the performance in L2-ASR against traditional ASR trained on the same target data.

The baseline is a chain Kaldi [25] model adapted from a Tedlium recipe[2] to our data. It consists of an acoustic model (AM) trained on the speech data described in section 3 and a language model (LM) built from the transcriptions of the training set. The AM is a Time-delayed Neural Network (TDNN) acoustic model with Long Short-term Memory (LSTM). There are 6 TDNN layers of size 512 and 3 LSTM layers of size 512 after every second TDNN layer. The input features are 40-d high-resolution Mel-Frequency Cepstral Coefficients (MFCCs) and 100-dim i-vectors. Volume perturbation and 3-way speed perturbation are applied during training for augmenting the AM training data. The LM is a four-gram language model with Witten-Bell discounting applied.

For the pre-trained models, we use the publicly available [3,4] Wav2Vec2 models of two model sizes: *Base* (about 95 million parameters) and *Large* (up to 317 million parameters). Two of them are monolingual and pre-trained on 4.5K hours of unlabelled native Swedish speech data of VoxPopuli dataset (sourced from European Parliament (EP) plenary session recordings [26]): a smaller *Wav2Vec2 Base* model which we denote as *Wav2vec2.0-Swedish-Base*, and a larger *Wav2Vec2 Large* model which we name as *Wav2vec2.0-Swedish-Large*. Another two models are multilingual Wav2Vec2 Large models. The one denoted as *Wav2vec2.0-XLSR-Multilingual-Large* is pre-trained on about 56K hours of unlabelled speech from Common Voice [27], BABEL [28] and Multilingual Librispeech (MLS) datasets [29], including 3 hours of native Swedish speech. The latter one, *Wav2vec2.0-100K-Multilingual-Large*, is pre-trained on the full 100K-hour VoxPopuli dataset.

To reduce GPU memory consumption in the fine-tuning phase, speech samples with a duration of more than 11 seconds were removed from the training set, resulting in a smaller training set (named as *Smaller training set* in Table 1) with a total duration reduced from 1146.3 to 273.6 minutes. All Wav2Vec2 models were first fine-tuned on this set on a single Tesla V100 GPU for 30-50 epochs. The fine-tuning of each model takes from 3 to 5 hours, depending on the number of epochs. Table 2 presents the word error rates (WER) after fine-tuning the different pre-trained models. With this setup, none of the Wav2Vec2 models outperforms the conventional TDNN-LSTM baseline. However, some preliminary results can be observed. First, the relative improvement of the *Large* architecture over the *Base* architecture is 31.6% for the monolingual model. Second, Large models pre-trained on a bigger amount of native Swedish speech data outperform the *XLSR* model where there are only 3 hours of Swedish speech data included during pre-training. Third, *Wav2vec2.0-100K-Multilingual Large* outperforms *Wav2vec2.0-Swedish-Large* with a relative improvement of 3.77% and is thus chosen for further experiments.

Table 2: *Results of experiments conducted in this work. Columns represent our developed models as well as the amount of fine-tuning data used and corresponding WER.*

Model	**No. of Samples**	**Size, min.**	**WER, %**
Baseline: TDNN-LSTM+LM	4720	1146.3	17.66
Wav2vec2.0-Swedish-Base	2287	273.6	32.98
Wav2vec2.0-Swedish-Large	2287	273.6	22.55
Wav2vec2.0-XLSR Multilingual Large	2287	273.6	28.30
Wav2vec2.0-100K	2287	273.6	21.70
Multilingual Large	4442	957.3	**16.38**

Next, the threshold for the duration of the training samples was increased up to 30 seconds, resulting in a dataset (named as *Larger training set* in Table 1) with a total duration of 957.3 minutes which is 83.5% of the original training set used for training the baseline. The model *Wav2vec2.0-100K-Multilingual Large* was then fine-tuned on this training set on 5 Tesla V100 GPUs for about 12 hours for 40 epochs. The resulting fine-tuned model achieves 16.38% WER, outperforming the conventional ASR system by 7.25% relative (see Table 2). In terms of character error rate (CER), the relative improvement to the baseline is 35.28%: *Wav2vec2.0-100K-Multilingual Large* reduced CER from 10.64% to 6.89%. This seems to suggest that the output produced by this end-to-end ASR is acoustically much closer to the reference than the output of a traditional ASR system, which utilizes a language model and (native) pronunci-

[2] https://github.com/kaldi-asr/kaldi/tree/master/egs/tedlium/s5_r2

[3] https://github.com/facebookresearch/voxpopuli#pre-trained-models

[4] https://github.com/pytorch/fairseq/tree/master/examples/wav2vec#pre-trained-models

Table 3: *A comparison between the decoded output of traditional ASR and our best model (Wav2vec2.0-100K-Multilingual-Large) along with the ground transcripts.*

(1)	Ref	öö	vad	är	din	ansigt	av	kungligen	i	sverige
	Hyp (Baseline)	öö	vad	är	din	åsikt	av			sverige
	Hyp (Our system)	öö	vad	är	din	ansikt	av	kunglingen	i	sverige
(2)	Ref	jag	tänker att	...	på	armen först	kan	kanske i dragsvik	...	studera
	Hyp (Baseline)	jag	tycker att	...	på	armen först		kanske idag	...	studera
	Hyp (Our system)	jag	tänker att	...	på	armenförst	kan	kanskeragvi	...	och studera
(3)	Ref	ööm	joensuu	är jätte	mysig	och det	är jätte bra			
	Hyp (Baseline)	öö	jag är så	är jätte	mysig	och det	är jätte bra			
	Hyp (Our system)	öm	jag en so	är jätte	mysik	och det	är jätte bra			

ation dictionary, even though the WER is only slightly lower. Because this would be interesting for the L2-ASR task, we will take a closer look at the actual ASR output in the next section.

6. Analysis of Results

In this section, we investigate some of the outputs of *Wav2vec2.0-100K-Multilingual Large* (our best model), and compare them to the output of the conventional ASR system (Kaldi). As shown in Table 3, in the first example, it seems that Kaldi does not recognize *"ansigt"* (*face*) and *"kungligen"* (*royal*) because they both do not appear in the training dataset. However, our model, interestingly, manages to extrapolate words that are not very phonemically far from the ground truth. For the first word *"ansigt"*, our model outputs *"ansikt"* which is one letter or phoneme away from the reference. Similarly, the second word *"kungligen"* is recognized as *"kunglingen"* which again not very far phonemically from the true label. In example number 2, Kaldi outputs a word *"tycker"* different from the reference while our model correctly outputs the correct word *"tänker"*. While both outputs have similar meaning (*think*), the language model in the Kaldi model favors *"tycker"* while our model adheres to the acoustic or phonetic representation. In this particular example, the speaker did not succeed to pronounce the word *"kanske"* so they repeat the first segment *"kan"* and Kaldi filtered out the disfluency error, while our model outputs the disfluency error as-is but it failed to output the word "kanske" because the latter is followed by a proper noun *"Dragsvik"* (a village in Finland).

Another example of where our model decodes partially correct phonetic outputs is the word *"Joensuu"* recognized as *"jag en so"* in sentence no. 3. On the other hand, Kaldi's hypothesis *"jag är så"* (*I am so...*) is grammatically correct because of the LM but the middle segment *"är"* is still phonetically different from that of the reference word.

Based on these observations and the WER and CER results in the previous section, we believe that the Wav2vec2.0-100K Multilingual model has learned to encode a useful universal phonemic representation that allows it to be used for cross-lingual transfer learning even for settings like ours where the data is very limited. This is suggested by the studies we surveyed and confirmed by the results of our experiments. Therefore, we find that this research direction is promising in the context of low resource L2 ASR as it reduces the engineering efforts required to design specialized pronunciation lexicons or language models.

7. Conclusions

In this paper, we discussed the use of monolingual and cross-lingual pre-trained acoustic models in an end-to-end ASR system for second language learners of Swedish. Our experiments show that models pre-trained on large size of untranscribed L1 Swedish speech data give a competitive performance to that of traditional ASR system without the need for customized modeling of language or pronunciation. Furthermore, models pre-trained on a large amount of multilingual untranscribed speech data outperform the traditional system using a reasonable amount of transcribed data. In our analysis of decoded outputs, our best model managed to correctly decode words that do not appear in the training dataset whereas the traditional ASR system failed to decode out-of-training words. We also noticed that our E2E-L2-ASR decodes morphemes or segments of words resultant from speaker's disfluency or mispronunciation. While this can be easily remedied by an external language model, it proves that these self-supervised models provide strong representation of phonemes. Regardless of the results we achieve, it is premature to indicate how effective or robust self-supervised E2E-L2-ASR systems are. Thus, our future work would include experiments on L2 Finnish ASR in addition to conducting thorough analysis of the hidden, latent representations that Wav2vec2 models encode during the learning.

8. Acknowledgements

This work is part of Digitala project which is funded by the Academy of Finland (grant numbers 322619, 322625, 322965). The computational resources were provided by Aalto ScienceIT.

9. References

[1] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[2] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," *arXiv preprint arXiv:1910.05453*, 2019.

[3] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[4] G. Saon, H.-K. J. Kuo, S. Rennie, and M. Picheny, "The ibm 2015 english conversational telephone speech recognition system," *arXiv preprint arXiv:1505.05899*, 2015.

[5] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The microsoft 2017 conversational speech recognition system," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5934–5938.

[6] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4774–4778.

[7] S. Park and J. Culnan, "A comparison between native and non-native speech for automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 145, no. 3, pp. 1827–1827, 2019.

[8] A. Rajpal, A. Rao, C. Yarra, R. Aggarwal, and P. K. Ghosh, "Pseudo likelihood correction technique for low resource accented asr," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7434–7438.

[9] Y. Gao, B. M. L. Srivastava, and J. Salsman, "Spoken english intelligibility remediation with pocketsphinx alignment and feature extraction improves substantially over the state of the art," in *2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*. IEEE, 2018, pp. 924–927.

[10] N. Jose, B. R. Chakravarthi, S. Suryawanshi, E. Sherly, and J. P. McCrae, "A survey of current datasets for code-switching research," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*. IEEE, 2020, pp. 136–141.

[11] A. Ragni, K. M. Knill, S. P. Rath, and M. J. Gales, "Data augmentation for low resource languages," in *INTERSPEECH 2014: 15th Annual Conference of the International Speech Communication Association*. International Speech Communication Association (ISCA), 2014, pp. 810–814.

[12] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1469–1477, 2015.

[13] J. Cui, B. Kingsbury, B. Ramabhadran, A. Sethy, K. Audhkhasi, X. Cui, E. Kislal, L. Mangu, M. Nussbaum-Thom, M. Picheny *et al.*, "Multilingual representations for low resource speech recognition and keyword search," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 259–266.

[14] M. Matassoni, R. Gretter, D. Falavigna, and D. Giuliani, "Non-native children speech recognition through transfer learning," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6229–6233.

[15] R. Duan, T. Kawahara, M. Dantsuji, and H. Nanjo, "Cross-lingual transfer learning of non-native acoustic modeling for pronunciation error detection and diagnosis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 391–401, 2019.

[16] S. Schaden, "Generating non-native pronunciation lexicons by phonological rule," in *Proc. ICSLP*. Citeseer, 2004, pp. 2545–2548.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[19] A. Baevski, M. Auli, and A. Mohamed, "Effectiveness of self-supervised pre-training for speech recognition," *arXiv preprint arXiv:1911.03912*, 2019.

[20] K. Singh, D. Okhonko, J. Liu, Y. Wang, F. Zhang, R. Girshick, S. Edunov, F. Peng, Y. Saraf, G. Zweig *et al.*, "Training asr models by generation of contextual information," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7864–7868.

[21] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[22] A. H. Liu, T. Tu, H.-y. Lee, and L.-s. Lee, "Towards unsupervised speech recognition and synthesis with quantized speech representation learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7259–7263.

[23] H. Kallio, R. Hilden, M. Kurimo, M. Vainio, R. Karhila, and E. Lindroos, "Developing a high-stake digital spoken language proficiency assessment: Results from pilot tests," 2016, pp. 1214–1214, international Technology, Education and Development Conference, INTED ; Conference date: 07-03-2016 Through 09-03-2016.

[24] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[26] C. Wang, M. Rivière, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," *arXiv preprint arXiv:2101.00390*, 2021.

[27] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[28] M. J. Gales, K. M. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low-resource languages: Babel project research at cued," in *Fourth International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU-2014)*. International Speech Communication Association (ISCA), 2014, pp. 16–23.

[29] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," *arXiv preprint arXiv:2012.03411*, 2020.