9th Workshop on Speech and Language Technology in Education (SLaTE)
18-20 August 2023, Dublin, Ireland

Multi-task wav2vec2 Serving as a Pronunciation Training System for Children

Yaroslav Getman, Ragheb Al-Ghezi, Tamás Grósz, Mikko Kurimo

Department of Information and Communications Engineering, Aalto University, Finland
{firstname.lastname}@aalto.fi

Abstract

Computer-assisted learning tools (CAPT) are increasingly reliant on AI tools. Recent studies demonstrated how neural systems pre-trained in a self-supervised fashion, such as wav2vec2, can overcome the data scarcity problem of most CAPT systems, especially if the target users are young children. In most current works, however, the focus lies on fine-tuning these models on a single task, which often leads to catastrophic forgetting and severely limits the capabilities of the fine-tuned model. In this work, we propose the usage of multi-task learning and demonstrate how a single wav2vec2 model can simultaneously generate transcript and assess pronunciation of Swedish children with speech sound disorder and child second language learners of Finnish. We also investigate which layer is the most informative for the rating task. Our multi-task solutions provide higher pronunciation classification performance and competitive ASR accuracy in comparison to the corresponding single-task systems.

Index Terms: ASR, speech assessment, wav2vec2, multi-task, children speech

1. Introduction

Computer-Assisted Pronunciation Training (CAPT) has garnered attention for its potential to enhance self-regulated pronunciation abilities. CAPT systems typically use a method for evaluating pronunciation called Goodness of Pronunciation (GOP) [1, 2]. The GOP approach measures the probability that the expected phone is observed with respect to all the other observable phones, and it typically involves an automatic speech recognition (ASR) system, a forced-aligner, and a scoring module [3]. The ASR system provides acoustic scores, which are used by the forced aligner to determine the phonemes uttered at a given time and calculate the acoustic scores per phoneme. The scoring module then employs the phonetic scores to estimate the GOP for a given speech sample. Recent studies focused on developing GOP models for children have shown that using ASR log posterior probabilities to train classifiers is a more effective approach compared to the conventional GOP pipeline [4].

Developing automatic pronunciation training systems is a challenging task due to the limited amount of training data, especially for special target groups such as child second language learners (L2) or children with speech sound disorder (SSD). However, successful attempts to apply self-supervised deep acoustic models like wav2vec2 [5] to low-resource domains make it possible to develop systems for ASR and various audio classification tasks [6, 7, 8]. The wav2vec2 models can also be fine-tuned directly for speech pronunciation classification. However, the wav2vec2 classifier alone is not aware of the target word the speaker is asked to pronounce. Therefore, a separate ASR system is needed to verify whether the speaker attempted to utter the target word. Unfortunately, combining these two large models results in a computationally intensive system, unsuitable for gamified learning environments such as mobile applications.

Since children speak differently compared to adults [9], an ASR model developed using adult speech can be expected to produce suboptimal transcripts. Earlier works demonstrated that continued pre-training is an efficient tool for domain adaptation [10, 11]. Unfortunately, in our case, this is not a feasible solution due to the lack of in-domain data. Instead, we opted for a continued fine-tuning strategy by first training the pre-trained models for general adult ASR and then adapting them to child speech via our limited target data.

The previously listed constraints motivated us to explore new approaches for fine-tuning CAPT and ASR models using our low-resource corpora. Our main goal is to create a unified end-to-end solution that can act both as an ASR and a CAPT model for children. To achieve this, we employ multi-task learning and show that a single wav2vec2 model can fulfill both tasks simultaneously and, unlike a wav2vec2 fine-tuned purely for the speech rating task, avoid catastrophic forgetting [12].

In addition, we investigate whether the common practice of using the last Transformer layer as input in multi-task learning of wav2vec2 [13, 14, 15, 16, 17] is the best solution. Previous studies suggest that static wav2vec2 representations from some intermediate layers embed more phonetic information valuable for estimating the quality of pronunciation [18, 19], predicting emotions [20] and L2 speakers' proficiency levels [21]. Layers 14-19 provided the best features for phoneme classification in [22]. Apart from these studies, we investigate the performance of hidden representations after jointly optimizing the entire network both for ASR and speech rating tasks, inspired by the fact that a layer's performance cannot be determined solely based on its pre-fine-tuning performance. In other words, we do not freeze the weights of the wav2vec2 during training for these downstream tasks.

Recent studies have proposed various multi-task learning frameworks for speech processing tasks. For instance, [23] proposed a framework that implements voice activity detection (VAD), speaker clustering (SC), and ASR using wav2vec2 layers in a hierarchical manner. By using different wav2vec2 layers for VAD, SC, and ASR from earlier to later layers, they achieved improved performance and reduced diarisation error rates. Another promising framework was proposed by [13], who demonstrated the effectiveness of wav2vec2 as a unified multi-task system for speaker verification and language identification. Their results indicated that a single wav2vec2 model can achieve competitive performance on both tasks without any task-specific modifications. Similarly, [14] proposed a single

10.21437/SLaTE.2023-8

Dataset	# of Samples					
	1	2	3	4	5	Total
SweSSD	425	662	1148	978	2814	6027
FinL2	68	247	64	579	1166	2124

Table 1: *Distribution of ratings in SweSSD and FinL2.*

multi-task learning framework for end-to-end ASR and accent recognition (AR) simultaneously. They found that sharing only a few encoder layers and a smaller weighting factor with the AR task yields better results. By jointly optimizing network parameters on multiple tasks using a shared backbone, these frameworks offer potential advantages such as reduced computational cost and improved performance.

To summarize, this work presents the following contributions. Firstly, we propose multi-task wav2vec2 solutions with a customized architecture for ASR and pronunciation classification of Swedish children with SSD and children practicing L2 Finnish. We develop multi-task systems that outperform the corresponding single-task ones on the speech rating task and maintain competitive ASR capability, as well as reduce computational costs. Secondly, we analyze the classification performance of our multi-task systems across different architectural choices and showcase a general performance curve that provides insight into which range of layers to narrow the search in similar studies. Lastly, we release our best performing multi-task wav2vec2 models along with the training scripts at `https://github.com/aalto-speech/multitask-wav2vec2`.

2. Data

This study incorporates experiments on two children's speech datasets. The first one, namely SweSSD [24], is a pathological speech corpus consisting of 6027 samples (about 2 hours) recorded from 28 native Swedish speakers aged 6 to 10 years, out of which 16 were diagnosed with an SSD and the rest 12 had typical speech. Our second dataset, named FinL2, is composed of 2124 samples (nearly 1.5 hours) uttered by 24 Ukrainian children of age 7-11 years practicing L2 Finnish. The SweSSD corpus is anonymized, while the L2 Finnish data include speaker IDs, which makes it possible to perform speaker separation when splitting the data.

Both datasets consist of single-word utterances, but the Swedish one has a much more diverse vocabulary. There are 1109 unique words in SweSSD, out of which 422 words occur only once. In contrast, FinL2 includes a set of 90 words repeated by all the speakers. These speech datasets are not transcribed, only the target word for each recording is known.

For each sample, the level of pronunciation is assessed by a human annotator on a scale of 1 to 5. The distribution of ratings in the corpora is reported in Table 1. Both corpora are heavily imbalanced towards the highest level: around half of the samples belong to class 5. The imbalance problem is relevant in particular for the L2 Finnish corpus, in which 2 out of 5 classes have only about 70 samples each.

3. Methods

3.1. wav2vec 2.0

wav2vec 2.0 [5] is a self-supervised learning framework designed for the efficient representation of raw audio data in vector space. The framework consists of three components, including a feature encoder, a context network, and a quantization module. The feature encoder uses a 1D convolutional neural network to convert the raw audio data into feature vectors. The context network is a stack of Transformer [25] encoder blocks which uses relative positional embedding and processes the feature vectors. In this work, we use *wav2vec2 Large* models composed of 317M parameters and 24 Transformer layers. Finally, the quantization module converts the continuous vector output to discrete representations via a Gumbel-Softmax function and maintains multiple codebooks.

The pre-trained model can be fine-tuned for various downstream tasks. For the ASR task, a linear layer, also referred to as the ASR head, is added on top of the network to classify the context representations into character tokens. The model is then fine-tuned on labeled data using a Connectionist Temporal Classification (CTC) loss [26]. After fine-tuning, it can be used directly as an end-to-end system without the need for an external language model.

3.2. Pronunciation Rating

For the purposes of automatic pronunciation rating, we used several wav2vec2-based approaches. First, we built single-task wav2vec2 models with a classification head on top of the network and trained them with the cross-entropy (CE) objective function. This head is composed of a linear projection layer followed by average pooling and a classification layer. Second, we implemented multi-task wav2vec2 solutions fine-tuned for ASR and speech pronunciation classification tasks simultaneously by feeding the corresponding heads with certain Transformer blocks, see Figure 1. Since our models were originally fine-tuned for adult ASR, we kept the ASR head connected to the last Transformer layer and put the classification head after a layer X in the range from 1 to 24. This way, the weights of the Transformer layers after the layer X were optimized purely by propagating the CTC loss, while the rest of the Transformer layers as well as the CNN network, were trained with a combined gradient of the CTC and the CE loss.

In addition, the predictions of the fine-tuned wav2vec2 classification head can be further adjusted by an external decision tree (*CER DT* in Figure 1) trained on character error rates (CERs) extracted from the wav2vec2 ASR component. The rationale for using CERs as features is that they can function as a crude measure of pronunciation error. In this case, the output class probabilities are merged between the DT and the classification head. The DT can also be used as a separate system on top of a single- or a multi-task wav2vec2 ASR head.

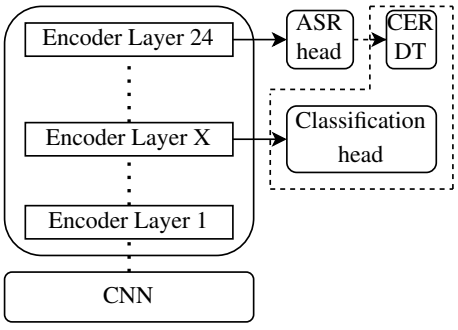

Figure 1: *Multi-task wav2vec2 system overview.*

Lastly, we fine-tuned separate wav2vec2 models for chil-

dren ASR. First, these models were needed to create ensemble systems of single-task classification systems, which cannot generate transcripts on their own, and CER decision trees. Second, they served as base models for training single-task speech rating systems by replacing the ASR head with a classification head. This provides a more fair comparison to multi-task solutions which utilized both the text transcripts and the rating labels during training.

4. Experiments

Due to the difficulties in collecting a sufficient amount of children's speech, we had to test our systems using cross-validation (CV) and split our data into 6 folds where only one fold was used for testing at a time. The CV setup enabled us to evaluate our models on the entire dataset at the cost of increased training time (6 models instead of one). Furthermore, training 6 models helped us investigate the stability of different approaches.

The second issue we had to face was data imbalance. As can be seen from Table 1, the high-rated classes (4 and 5) are extremely overrepresented in both corpora. Therefore, standard metrics like accuracy would not reflect the actual performance of our systems faithfully. To properly compare different rating models, we opted to use the Unweighted Average Recall (UAR), which averages the recalls of each category, thus giving us an estimate of the per-class performance. Additionally, we must note that not all errors are equal in our case, and the difference between the predicted rating and the annotated one should be considered too. To account for this, we report the Mean Absolute Error (MAE), which provides an estimate of the average difference between the automatically generated labels and the human-annotated ones. Finally, we assess system-human and human-human agreement using the weighted quadratic kappa and Spearman's correlation (κ and ρ in Table 2). For SweSSD, about 20% of the corpus was randomly sampled and re-evaluated by the original human rater 6 months after the original annotation. For FinL2, 10% of the data was rated by 3 raters, and the one with the highest agreement with the original ratings was chosen.

Since the children's speech datasets are rated, but not transcribed, we trained and evaluated our ASR models only by speech samples rated as 4 or 5 (containing only minor mispronunciations) with target words as reference transcripts. During multi-task training, all recordings are fed into the wav2vec2 network, but the CTC loss is not calculated for samples belonging to a level lower than 4, and they are masked during the word and the character error rate (WER and CER) calculation.

As our base models for the experiments, we used publicly available *wav2vec2 Large* models. The Swedish model was originally pre-trained on audiobooks and other speech from collections of the National Library of Sweden [27] and fine-tuned on NST [28] and CommonVoice [29], while the Finnish model was pre-trained on the Uralic (Finnish, Estonian and Hungarian) subset of the European parliamentary speech collection called VoxPopuli [30] and fine-tuned on 100 hours of colloquial Finnish from the Lahjoita Puhetta (Donate Speech) corpus [31].

4.1. Selecting the Optimal Hidden Layer for Speech Rating

As discussed in Section 1, the performance of static acoustic embeddings of wav2vec2 on various target tasks varies across hidden layers, and the ones from the top layer are not always the best. However, the best-performing hidden representations extracted from a certain Transformer layer do not guarantee this layer is the most optimal after the wav2vec2 model is jointly optimized for ASR and speech classification. Therefore, our preliminary experiments were aimed to determine how the choice of the Transformer layer to precede the classification head affects the classification performance after multi-task fine-tuning.

For each language, we kept the CTC head after the last wav2vec2 layer and connected the classification head with a projection layer (256 output units) on top of one Transformer block at a time, then fine-tuned the system with a learning rate of 7e-5 for 20 epochs and measured the development recall. Both losses (CTC and CE) had the same weight of 1, based on our preliminary experiments. We repeated training by keeping each of the 6 folds in turn as a development set and excluded the corresponding test folds from training and evaluation to avoid unintentional optimization toward the test data.

The results of this experiment are shown in Figure 2. Apart from SweSSD, the development recalls on FinL2 vary considerably between folds, which confirms the necessity of having a cross-validation setup in the later experiments. Otherwise, our preliminary experiments revealed that both Swedish and Finnish models consistently provide low classification performance when the classification head is connected to any layer in the first half of the wav2vec2 network. Another interesting observation is that when the classification head is connected to the last hidden layer, the classification performance (UAR) often degrades. The performance drop is particularly noticeable in L2 Finnish results: the development recall is always lower when the classification head is connected to the last layer than the second last one, and for 4 out of 6 folds this setup is the worst. Based on these results, we chose the best-performing layers in terms of development recalls (#19 for SweSSD and #17 for FinL2) in addition to the last layer for training and analyzing the final systems.

4.2. Multi-task Experiments

In our main experiments (Table 2), we trained multi-task systems (MT_W2V2) and compared them to single-task wav2vec2 models trained solely for ASR (W2V2_ASR) or speech rating (W2V2_RATING). We used the same training hyperparameters as reported in Section 4.1. In addition, we built simple decision tree (DT) classifiers by following a 6-fold CV using the CERs of each wav2vec2 system as input and aggregated the label probabilities to analyze if wav2vec2 and DT can complement each other in an ensemble ([system]+CER_DT). It should be noted that each ensemble system has a different DT trained on CERs from the corresponding ASR model, however, we report separate DT results only for the best-performing one.

The single-task ASR systems achieved 17.01%/6.34% and 4.47%/1.36% (WER/CER) for SweSSD and FinL2, respectively. It should be noted that such low error rates for L2 Finnish can be a consequence of over-learning the same set of 90 words uttered by all 24 speakers even though the ASR does not rely on any lexicon or language model. In contrast, multi-task training of wav2vec2 always provided slightly worse ASR performance compared to that of the single-task models. Furthermore, CER and WER increased when the classification head was not connected to the last hidden layer of the multi-task wav2vec2.

Our single-task wav2vec2 classifiers outperformed the multi-task solutions in terms of UAR if both the ASR and the classification head of the multi-task system were put after the last Transformer layer. However, the situation changed when the Transformer layer for pronunciation rating was chosen properly. For SweSSD, choosing the most appropriate encoder block

Figure 2: *Development recalls across cross-validation folds after fine-tuning in a multi-task setup with a classification head connected to a certain wav2vec2 Transformer layer.*

to precede the classification head from the preliminary experiments (layer 19) resulted in 0.74% absolute UAR improvement over the single-task wav2vec2. For Finnish the corresponding improvements are even bigger: The UAR of MT_W2V2 (L17) is 14.75% higher than the single-task classification system.

The DTs trained on the CERs extracted from the wav2vec2 models have the lowest UAR and the highest MAE, which indicates that the CER alone does not contain all the needed information for pronunciation rating. However, adding them to an ensemble with the wav2vec2 systems proved to be beneficial. Combining the class probabilities of the best multi-task system and the corresponding decision tree provided an additional UAR improvement of 1.45% and 0.9% for Swedish and Finnish, respectively. Moreover, these DTs can act as a verification system and complement the neural classification component, which is not aware of the target word. For example, they would provide a penalty term for uttering a word other than the target word even if the child's speech is otherwise intelligible and pronunciation is good.

The results of the human agreement indicate that the pronunciation rating of short speech samples is difficult even for human experts. While the human result outperformed our best system on SweSSD, the annotator still did not completely agree with herself when re-annotating the data. The human raters' disagreement is even higher in FinL2, where they outperformed our best model only in terms of Spearman's correlation (ρ). Note that the human performance level is estimated only on 10% of the data.

System	WER, %	CER, %	UAR, %	MAE	ρ	κ
SweSSD						
CER_DT	*N/A*	*N/A*	43.39	.69	.681	.334
W2V2_ASR	**17.01**	**6.34**	*N/A*	*N/A*	*N/A*	*N/A*
W2V2_RATING	*N/A*	*N/A*	48.26	.54	.734	.435
↳ + CER_DT			49.28	.53	.737	.440
MT_W2V2	17.17	6.42	47.64	.53	.752	.430
↳ + CER_DT			49.48	**.52**	**.763**	.439
MT_W2V2 (L19)	18.62	7.05	49.00	.56	.718	.437
↳ + CER_DT			**50.45**	.55	.734	**.445**
Human (20% data)	*N/A*	*N/A*	**65.75**	**.39**	**.877**	**.572**
FinL2						
CER_DT	*N/A*	*N/A*	39.05	.57	.573	.256
W2V2_ASR	**4.47**	**1.36**	*N/A*	*N/A*	*N/A*	*N/A*
W2V2_RATING	*N/A*	*N/A*	46.06	.39	.720	.532
↳ + CER_DT			49.57	.38	.722	.538
MT_W2V2	6.30	2.13	43.07	.37	.729	.526
↳ + CER_DT			45.29	.37	**.730**	.529
MT_W2V2 (L17)	6.76	2.22	60.81	**.36**	.720	.565
↳ + CER_DT			**61.71**	**.36**	.723	**.571**
Human (10% data)	*N/A*	*N/A*	44.40	.50	**.757**	.406

Table 2: *Results of speech rating experiments. The multi-task systems (MT_W2V2) can do both ASR and speech rating. The Transformer layer number preceding the classification head is added in the parenthesis if other than the last layer (L24).*

5. Conclusions

In this work, we developed automatic pronunciation training systems for Swedish children with SSD and L2 children learning Finnish. We demonstrated that computationally efficient multi-task solutions can provide higher pronunciation rating performance and competitive ASR accuracy compared to separately optimized wav2vec2 systems for ASR and rating. Moreover, we conducted a thorough analysis of the multi-task model classification performance for 2 different wav2vec2 models and 2 target datasets. We showed that for pronunciation rating by the multi-task wav2vec2, the last hidden layer commonly chosen in previous works for audio classification tasks is not the best one. Although the choice of the best layer seems to be model- and dataset-specific, we showcased the general shape of the performance curve that offers insight into which layers others should try. To train a similar model for the speech rating task on a new dataset, the range of search can be limited to a certain region of layers. Next, we will integrate the multi-task system into our mobile CAPT game application for children to see the practical effect in feedback speed and performance.

6. Acknowledgments

The computational resources were provided by Aalto ScienceIT. This work was supported by NordForsk through funding for Technology-enhanced foreign and second-language learning of Nordic languages, project number 103893.

7. References

[1] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[2] C. Tejedor García *et al.*, "Design and evaluation of mobile computer-assisted pronunciation training tools for second language learning," 2020.

[3] R. Karhila, A. Smolander, S. Ylinen, and M. Kurimo, "Transparent pronunciation scoring using articulatorily weighted phoneme edit distance," in *Proceedings of Interspeech*. International Speech Communication Association, 2019, pp. 1866–1870.

[4] J. Wang, Y. Qin, Z. Peng, and T. Lee, "Child Speech Disorder Detection with Siamese Recurrent Network Using Speech Attribute Features," in *Proc. Interspeech 2019*, 2019, pp. 3885–3889.

[5] 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.

[6] X. Xu, Y. Kang, S. Cao, B. Lin, and L. Ma, "Explore wav2vec 2.0 for Mispronunciation Detection," in *Proc. Interspeech 2021*, 2021, pp. 4428–4432.

[7] L. Peng, K. Fu, B. Lin, D. Ke, and J. Zhan, "A Study on Fine-Tuning wav2vec2.0 Model for the Task of Mispronunciation Detection and Diagnosis," in *Proc. Interspeech 2021*, 2021, pp. 4448–4452.

[8] T. Grósz, D. Porjazovski, Y. Getman, S. Kadiri, and M. Kurimo, "Wav2vec2-based paralinguistic systems to recognise vocalised emotions and stuttering," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 7026–7029.

[9] S. W. Lee, A. Potamianos, and S. S. Narayanan, "Acoustics of children's speech: developmental changes of temporal and spectral parameters." *The Journal of the Acoustical Society of America*, vol. 105 3, pp. 1455–68, 1999.

[10] M. DeHaven and J. Billa, "Improving low-resource speech recognition with pretrained speech models: Continued pretraining vs. semi-supervised training," 2022.

[11] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training," in *Proc. Interspeech 2021*, 2021, pp. 721–725.

[12] S. P. Sawant, "Understanding catastrophic forgetting for adaptive deep learning," in *Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD)*, ser. CODS-COMAD '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 282–283.

[13] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on Speaker Verification and Language Identification," in *Proc. Interspeech 2021*, 2021, pp. 1509–1513.

[14] J. Zhang, Y. Peng, V. T. Pham, H. Xu, H. Huang, and E. S. Chng, "E2E-Based Multi-Task Learning Approach to Joint Speech and Accent Recognition," in *Proc. Interspeech 2021*, 2021, pp. 1519–1523.

[15] M. Kunešová and Z. Zajíc, "Multitask detection of speaker changes, overlapping speech and voice activity using wav2vec 2.0," *arXiv preprint. arXiv:2210.14755*, 2022.

[16] S. Hussain, V. Nguyen, S. Zhang, and E. Visser, "Multi-task voice activated framework using self-supervised learning," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6137–6141.

[17] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech Emotion Recognition with Multi-Task Learning," in *Proc. Interspeech 2021*, 2021, pp. 4508–4512.

[18] P. Cormac English, J. D. Kelleher, and J. Carson-Berndsen, "Domain-informed probing of wav2vec 2.0 embeddings for phonetic features," in *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Seattle, Washington: Association for Computational Linguistics, Jul. 2022, pp. 83–91.

[19] Y. K. Singla, J. Shah, C. Chen, and R. R. Shah, "What do audio transformers hear? probing their representations for language delivery & structure," in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2022, pp. 910–925.

[20] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech 2021*, 2021, pp. 3400–3404.

[21] R. Al-Ghezi, Y. Getman, E. Voskoboinik, M. Singh, and M. Kurimo, "Automatic rating of spontaneous speech for low-resource languages," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 339–345.

[22] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 27 826–27 839.

[23] X. Zheng, C. Zhang, and P. Woodland, "Tandem Multitask Training of Speaker Diarisation and Speech Recognition for Meeting Transcription," in *Proc. Interspeech 2022*, 2022, pp. 3844–3848.

[24] S. Strömbergsson, K. Holm, J. Edlund, T. Lagerberg, and A. McAllister, "Audience response system-based evaluation of intelligibility of children's connected speech – validity, reliability and listener differences," *Journal of Communication Disorders*, vol. 87, p. 106037, 2020.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[26] A. Graves and A. Graves, "Connectionist temporal classification," *Supervised sequence labelling with recurrent neural networks*, pp. 61–93, 2012.

[27] M. Malmsten, C. Haffenden, and L. Börjeson, "Hearing voices at the national library – a speech corpus and acoustic model for the swedish language," *arXiv preprint. arXiv:2205.03026*, 2022.

[28] M. B. Birkenes, "NST Swedish Dictation (22 kHz)," https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-17/, 2020.

[29] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222.

[30] C. Wang, M. Riviere, 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," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, vol. 1. Association for Computational Linguistics, Aug. 2021, pp. 993–1003.

[31] A. Moisio, D. Porjazovski, A. Rouhe, Y. Getman, A. Virkkunen, R. AlGhezi, M. Lennes, T. Grósz, K. Lindén, and M. Kurimo, "Lahjoita puhetta: a large-scale corpus of spoken finnish with some benchmarks," *Language Resources and Evaluation*, 2022.