INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Transformer-based Speech Recognition Models for Oral History Archives in English, German, and Czech

Jan Lehečka, Jan Švec, Josef V. Psutka, Pavel Ircing

Department of Cybernetics, University of West Bohemia Pilsen, Czech Republic
{jlehecka,honzas,psutka_j,ircing}@kky.zcu.cz

Abstract

This paper is a step forward in our effort to make vast oral history archives more accessible to the public and researchers by breaking down the decoding barriers between the knowledge encoded in the spoken testimonies and users who want to search for the information of their interest. We present new Transformer-based monolingual models suitable for speech recognition of oral history archives in English, German, and Czech. Our experiments show that although the all-purpose speech recognition systems have recently made tremendous progress, the transcription of oral history archives is still a challenging task for them; our tailored models significantly outperformed larger public multilingual models and scored new state-of-the-art results on all tested datasets. Due to the 2-phase fine-tuning process, our models are robust and can be used for oral history archives of various domains. We publicly release our models within a public speech recognition service.

Index Terms: speech recognition, oral history archives

1. Introduction

Oral history archives usually store vast and extremely valuable knowledge from our history recorded within audiovisual interviews. The authentic memories of individual speakers are thus encoded in the spoken utterances, where they are hard to reach by users interested in particular pieces of memories but not having time to listen to the whole interviews. Thus, it is very important to reliably transcribe the speech into text in order to allow efficient searching in the archives. Due to the extreme sizes of oral history archives, such as the one examined in this paper, manual transcription of interviews would be unfeasible. So in the last two decades, researchers around the world have been developing automatic speech recognition (ASR) systems and advanced search engines on top of oral history archives to make the content more accessible.

Several years ago, a new era of the artificial intelligence field started by introducing the Transformer architecture [1]. Three years ago, Transformer-based models established a new paradigm also in the speech recognition domain – the Wav2Vec 2.0 model [2]. Since then, we are witnessing the rapid growth of the family of Wav2Vec-like models along with the necessary rapid growth of large-scale audio datasets for self-supervised pre-training of these models.

In this paper, we are working with data from the Visual History Archive (VHA), which is an audiovisual archive originally collected in the 1990s to preserve the memories of Holocaust survivors. Today, these interviews are stored at the Shoah Foundation Institute (SFI) at the University of Southern California (USC), along with other interviews with witnesses to the history of the entire 20th century (more than 54k interviews). The Holocaust part of the archive contains testimonies in 32 languages of the personal memories of people who survived the World War II Holocaust. Most of them are in English (approximately half of the entire archive). More than 570 testimonies are in Czech (almost 1,000 hours of video), and almost 1,000 testimonies are in German (more than 2,000 hours of video). Interviews (in all languages) collected in the archive contain natural speech, full of disfluencies, emotional excitements, heavy accents, and are often influenced by the high age of speakers (problems with keeping ideas). The average age of all speakers at the time of recording was about 75 years. We denote this archive as SFI-VHA in this paper.

We are contributing to the family of Wav2Vec models with new ASR systems suitable not just for SFI-VHA archives but also for other oral history archives in English, German, and Czech. We are releasing the speech recognition systems as a public service for the research community[1].

Our ultimate goal in this field is to develop a speech recognizer for the oral history archives to a satisfactory level where users will be satisfied with the quality of the transcriptions and, when this is done, to fully concentrate on higher levels of search engines incorporating AI models in order to offer highly-relevant content to user's queries.

2. Related work

The original MALACH (Multilingual Access to Large Spoken Archives) project took place between 2001 and 2006. Its goal was to provide better access to the SFI-VHA archive via ASR and IR techniques. The WER of the ASR systems developed within the project reached 39.40% for English [3] and 38.57% for Czech [4] by the end of the project in 2006. Even after the project finish, the efficiency of the ASR systems was being continuously improved using new approaches, so that in 2011 the WER of 27.11% was achieved for Czech recordings [5]. New training methods based on DNN brought further improvement of WER (21.70% for English [6] and 19.11% for Czech [7]). The best WERs without using end-to-end approaches were published in 2021 [8] and reached 17.85% for English and 14.65% for Czech.

After the introduction of end-to-end Transformer-based audio models, [9] reported a significant improvement for the Czech dataset (WER=10.48%). For English, we are not aware of any reported improvements since then. For German, we didn't find any related work reporting ASR results in the literature, so this paper is the first one to report ASR results for this dataset.

This research was supported by the Czech Science Foundation (GA CR), project No. GA22-27800S.

[1] https://uwebasr.zcu.cz

10.21437/Interspeech.2023-872

3. Datasets

For each target language, we converted available transcribed videos from the SFI-VHA archive into a unified audio format (16kHz mono). We sliced long train and development recordings into segments not exceeding 30s, which is a reasonable limit of input examples during training due to GPU memory limits. To allow the trained ASR models to learn long-distance dependencies (up to 30s), we kept together as much context as possible in each segment while favoring segments ending with a full stop and pause. We cleaned all transcripts by removing non-speech events and punctuation and mapping texts into lowercase. The data statistics are shown in Tab. 1.

Table 1: *Fine-tuning datasets. We show the number of hours, words in transcripts (in thousands), and the average length of train/dev/test segments of audio (in seconds).*

	English			German			Czech		
	train	dev	test	train	dev	test	train	dev	test
# hours	245.7	9.2	4.3	1 803	33.0	80.8	87.2	19.2	9.0
# words	1 934	73	36	13 428	252	575	615	137	63
avg-len	27.0	25.2	5.2	20.3	20.2	1 692	24.1	24.1	10.6

3.1. English and Czech datasets

For English and Czech, we used datasets released under the Linguistic Data Consortium (LDC) – English [10] and Czech [11]. We adopted the same train-dev-test splits as in [8] and segmented train and development parts using time labels from the annotations into segments complying with the input limits of Transformers. The test parts for these two languages were already cleaned and contained only selected shorter segments (usually covering the maximum length of a single speaker's utterance without overlaps). As we found in [9], the Czech dataset contains a mix of formal and colloquial Czech in transcripts, causing a mismatch between train and test data, so we converted all Czech training transcripts into formal Czech to keep our results comparable.

3.2. German dataset

In order to make the German video interviews of the SFI-VHA more usable for research and teaching at universities, but also for educational work in schools, Freie Universität Berlin transcribed approximately 900 German-language interviews with a total duration of almost 2,000 hours.[2] The transcription of testimonies was part of the project "Witnesses to the Shoah" [12] funded by the German Lottery Foundation Berlin (Stiftung Deutsche Klassenlotterie Berlin) and took place between 2008 and 2013. The transcripts were made according to a set of rules created especially for the project. Software developed specifically for the transcription project automatically provided the segmentation of the transcripts according to the one-minute segmentation of the video interviews in The Visual History Archive. The transcripts were prepared by a total of over 100 freelance transcribers. The quality management then checked the correct spelling and punctuation, compliance with transcription guidelines, and the consistency of the text and video.

[2]Transcripts are publicly available at `https://transcripts.vha.fu-berlin.de`.

Since there was no train-dev-test split in the data, we split it randomly to keep records from 4% of all speakers in the test part and 2% in the development part. Because the transcriptions contained 1-minute segments without word-level time labels and training of Wav2Vec models requires a 30s limit of input data, we re-segmented the train and development data using force alignment. First, we fine-tuned a German Wav2vec 2.0 model using the CommonVoice dataset [13] to get a base German ASR system. For each recording, we used this ASR system together with a language model trained only from its reference transcript to get the word-level time labels while forcing the ASR system to decode n-grams from the reference. We aligned decoded words with the reference text and adopted the time labels for reference words wherever the transcriptions were in sync. Finally, we segmented long recordings on pauses while favoring long segments (but not exceeding 30s) ending with a full stop. In the test part, we left the long recordings untouched to avoid segmentation errors affecting the quality of the test data. Hence, recordings in the German test dataset are much longer and less clean than test data from other languages.

4. Pre-trained models

Where available, we used public pre-trained models. However, due to a lack of high-quality monolingual pre-trained models, we also pre-trained one new model (German base) from scratch. For the newly pre-trained model, we used Wav2Vec 2.0 architecture [2] (we will use the shorter abbreviation *W2V2* in our experiments and the following text). We adopted the same hyperparameter setting as in the paper, i.e., we trained the base model (12 Transformer blocks, model dimension 768, 8 attention heads, total 95 million parameters) for 400 thousand steps with a batch size of about 1.6 hours. We used `Fairseq` tool[3] for both pre-training and fine-tuning of models.

4.1. W2V2-base models

For English and Czech, we used already published pre-trained base-sized models – `wav2vec2-base`[4] [2] for English, and `ClTRUS`[5] [14] for Czech.

We didn't find any suitable pre-trained monolingual model for German, so we pre-trained a new base-sized model from scratch. Since W2V2 models are known to scale well with the size of pre-training data, we tried to gather as much public unlabeled speech data as possible. We collected over 65 thousand hours of German speech from various sources. The collection includes recordings from the German portion of the VoxPopuli dataset [15] (28k hours), a mix of self-crawled publicly available German podcasts (25k hours), audiobooks from the LibriVox project[6] (4.2k hours), recordings from several oral history archives (4k hours), selected speech data from BAS CLARIN Repository [16] (2.3k hours), German portion of CommonVoice corpus 11.0 [13] (1.2k hours), and a smaller amount of data from several other domains. The pre-training took about two weeks on a machine with four NVIDIA A100 GPUs.

[3]`https://github.com/pytorch/fairseq`

[4]`https://dl.fbaipublicfiles.com/fairseq/wav2vec/wav2vec_small.pt`

[5]`https://huggingface.co/fav-kky/wav2vec2-base-cs-80k-ClTRUS`

[6]`https://librivox.org`

4.2. Multilingual models

To compare W2V2 models with other public models, we selected `W2V2-XLS-R-300M` [17], a popular model pre-trained on 128 languages and approximately 436 thousand hours of unlabeled speech data. We experimented with the 300M variant, which has more than 300 million parameters, i.e., more than 3× more than the W2V2-base model.

Finally, we compared W2V2 models with `Whisper` [18], another popular model trained on 99 languages from 680,000 hours of multilingual and multitask labeled data. This model differs from W2V2 models in two main aspects: (1) it has an encoder-decoder architecture, where the decoder serves as an audio-conditioned language model, (2) the input is not raw audio signal, but Mel spectrogram. We experimented with two sizes: `Whisper-small` (12+12 Transformer layers, model dimension 768, 12 attention heads, total 244 million parameters) and `Whisper-large` (32+32 Transformer layers, model dimension 1280, 20 attention heads, total 1.55 billion parameters). Moreover, we experimented with the language identification ability of the model. We compared the decoding results when we input the correct language along each recording (`lang-spec`) and when the model identifies the language automatically from the input signal (`lang-auto`).

5. Fine-tuned models

We prepared training and development ASR data for fine-tuning as described in Sec. 3. If not stated otherwise, we used the same hyperparameters as in [2] and optimized models with the Connectionist Temporal Classification (CTC) loss [19] on labeled data. We also used a 2-phase fine-tuning setup as recommended in [14] to scale the fine-tuning up (see Fig. 1). Specifically, we were experimenting with three fine-tuning settings.

In-domain fine-tuning (FT_{ID}) is a standard single-phase fine-tuning on labeled data from only one target language. We fine-tuned the models for 80k updates with a batch size of 27 minutes. We used the learning rate 2×10^{-5} for English and Czech datasets, and 8×10^{-5} for the larger German dataset, as these learning rates gave us consistently the best results on development datasets across all models.

General-domain fine-tuning (FT_{GD}) is a single-phase fine-tuning on all labeled ASR data we were able to collect for the target language. The aim is to train a universal domain-independent ASR model for each language which can be used as a starting checkpoint for the second fine-tuning phase or as a general ASR model for unknown domains. Datasets for FT_{GD} vary between languages in both diversity and amount. For English, we used 12.5 thousand hours of data from the CommonVoice, SFI-VHA, and GigaSpeech [20]. For German, we used a mix of CommonVoice, VoxPopuli, SFI-VHA, LibriSpeech [21], and BAS Repositories [16], which was together over 6 thousand hours of data. For Czech, we used a mix of the CommonVoice, VoxPopuli, SFI-VHA, radio and TV shows, and telephone data, summing up to almost 6 thousand hours of transcribed speech. Since these datasets are large, we increased the number of fine-tuning updates to 160k and the batch size by a factor of 4 (i.e. 108 minutes).

2-phase fine-tuning (FT_{GD} + FT_{ID}) is a sequence of two previously described fine-tunings of the model, i.e., the FT_{GD} followed by the FT_{ID} as depicted in Fig. 1. Models trained with this 2-phase fine-tuning are robust speech recognizers that can profit from large-scale out-of-domain ASR data while preferring the in-domain predictions due to the second phase [14].

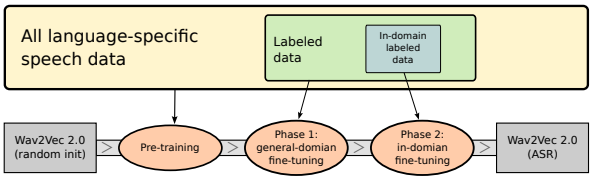

Figure 1: *The scheme of 2-phase fine-tuning.*

6. Language models

When evaluating fine-tuned models, we experimented with two decoding strategies: (1) CTC grapheme-based lexicon-free decoder and (2) CTC beam search decoder with a language model (LM). For strategy (2), we used `pyctcdecode`[7] tool and selected the LM of the correct target language when evaluating the models. The decoding with LM usually improves speech recognition performance by bringing useful language information into the decoding process while penalizing improbable outputs in the target language.

For each language, we trained large-scale 4-gram LM using `KenLM` [22] toolkit. We used web pages from the Common Crawl project[8] as a text data source. For each language, we collected more than 100GB of cleaned and deduplicated text. To keep the LMs of practical sizes, we pruned all unigrams with counts lower than ten and higher-order n-grams with counts lower than 100. We trained all LMs in lowercase as all fine-tuning transcripts were converted into lowercase. The sizes of LM vocabulary were 3.7 million words (English), 9.9 million words (German), and 4.8 million words (Czech).

Decoding with LMs trained from the Common Crawl project has a great advantage for oral history archives. These large-scale LMs could also cover very rare words (e.g., names of small villages or publicly unknown people) mentioned only several times somewhere on the Internet. Correct recognition of these content-bearing rare words is critical for users when searching for some very specific piece of information.

7. Pipeline of processing the archives

We are using fine-tuned modes as a core part of the archive processing pipeline. First, we run the ASR inference on overlapping signal chunks to give the model sufficient context for each chunk center. Then, we put together the output chunk-center logits of the whole sequence and apply CTC decoding with the language model. Since our models generate only lowercase characters without punctuation, we then process the output with a text Transformer fine-tuned to restore casing and punctuation marks. Throughout the whole pipeline, we keep the time alignment of words with the audio to finally generate subtitles of optimal duration and amount of text that are easy to read for users watching the video.

8. Experiments

In our experiments, we pre-trained and fine-tuned all described models and evaluated them on the test part of a relevant language dataset. The test parts were held out during the whole fine-tuning process and had no speaker overlaps with train or development parts. We compared models in terms of word er-

[7] `https://github.com/kensho-technologies/pyctcdecode`

[8] `https://commoncrawl.org`

ror rate (WER). Our results are tabulated in Tab. 2. It is worth noting that test parts have different qualities for each language; thus, WER values cannot be compared across languages. E.g., Czech models have lower WER than German models of the same size, but that does not mean they are better because the Czech test dataset was pre-processed to be very clean, whereas the German test part contains the original signal with all speech overlaps, long pauses with background noise, acoustic non-speech events, etc. We can only compare models evaluated on the same test dataset (i.e., results from the same column in Tab. 2).

Table 2: *Evaluation results in terms of WER* [%].

	English	German	Czech
CNN-TDNN_LF-MMI [8]	17.85	–	14.65
W2V2-base + FT_{ID}	16.91	18.61	11.52 [9]
+ LM	14.70	17.32	9.32
W2V2-base + FT_{GD}	23.68	18.56	19.27
+ LM	21.15	18.71	14.07
W2V2-base + FT_{GD} + FT_{ID}	14.21	17.77	9.97
+ LM	**12.88**	**17.08**	**8.43**
W2V2-XLS-R-300M + FT_{ID}	16.69	24.49	13.36
+ LM	**14.31**	**22.52**	**9.50**
Whisper-small (lang-auto)	28.34	41.81	41.00
Whisper-small (lang-spec)	20.88	25.81	38.17
Whisper-large (lang-auto)	23.79	35.13	28.15
Whisper-large (lang-spec)	17.34	22.99	25.95

For each fine-tuned W2V2 model, we report results with two decoding strategies: (1) CTC grapheme-based lexicon-free decoding, and (2) decoding with CTC beam search decoder with a language model (denoted as the "+ LM" rows in the Tab. 2) – see Sec. 6 for details.

To see how big a step forward ASR systems have made when switching the paradigm to Transformer-based architecture, we also add two years old results scored by the state-of-the-art model of that time [8] for comparison. We also evaluated fine-tuned XLS-R and Whisper to compare W2V2-base models with other relevant models. The Whisper models are already fine-tuned by authors, so we used them as they were.

However, when comparing the models in Tab. 2, it is important to keep in mind the sizes of individual models and fairly compare only models of similar sizes. The W2V2 models of the base size have 95M parameters, Whisper-small is more than $2\times$ the larger (244M parameters), XLS-R model is more than $3\times$ larger (315M parameters) and Whisper-large is about $16\times$ as large as the W2V2-base models.

9. Discussion

Based on our experimental results in Tab. 2, we confirm that including LM from Common Crawl into the CTC decoder improves the ASR results (with only one exception for the German W2V2-base + FT_{GD} model). We can observe larger improvements for models fine-tuned only in a single phase than for models trained in the 2-phase fine-tuning, which means that with longer fine-tuning, the W2V2 models are learning also the in-domain n-gram frequencies from the training transcripts.

For the Czech dataset, we replicated the results of in-domain fine-tuning of the W2V2-base model published in [9] and improved it by a significant margin by adding the large-scale LM (WER decreased from 11.52% to 9.32%) and applied the 2-phase fine-tuning (further improvement to WER=8.43%), which is – to our best knowledge – a new state-of-the-art result on this dataset.

We took a closer look at the German model as it scored higher WERs and behaved somehow differently from the English and Czech models. We found out that the main portion of errors consists of incorrect transcriptions of declined German articles, which sound very alike (e.g., "den" vs. "dem", "eine" vs. "einer" vs. "einem" vs. "einen"), especially when spoken by an old person in a fast, heavily accented, and emotionally stressed utterance. For these words, we observed a large number of disagreements between annotations and predictions. This phenomenon, together with the test data quality, is the main reason for the higher WER values of the German model.

Our results also confirmed the positive impact of 2-phase fine-tuning when compared with in-domain fine-tuning. Moreover, we can observe that when the in-domain dataset is large enough (the German dataset), the positive impact of the 2-phase fine-tuning is rather minor (WER decreased from 17.32% to 17.08%). We hypothesize that as the model has a sufficient amount of in-domain data available, it does not profit so much from out-of-domain datasets anymore.

From the comparison of W2V2 models with the popular multilingual models (XLS-R and Whisper), we can clearly see the superiority of having high-quality monolingual models pre-trained exclusively for one target language over multi-lingual models sharing the weights for many languages at once. Although both XLS-R and Whisper models are much larger and were (pre-)trained from many times more speech data than the monolingual W2V2-base models, the results from W2V2-base FT_{ID} models are very close (English and Czech) or even significantly better (German) than the results from the XLS-R model.

The results from the Whisper models are often far from the results of other models. We hypothesize that it is because the oral history archives are out of domains the model was trained on, and thus the model is at a disadvantage when compared with other models that were fine-tuned directly on the oral history archives domain before evaluation. An interesting result is that it is always beneficial to specify the correct language of the input speech, otherwise, the Whisper models often identify the language incorrectly and the whole transcript of the recording is then unintelligible causing many ASR errors.

10. Conclusions

In this paper, we presented new high-quality monolingual Transformer-based models suitable for speech recognition of oral history archives in English, German, and Czech. These models significantly outperform existing multilingual models while scoring new state-of-the-art results on all tested datasets. Thanks to the 2-phase fine-tuning, the proposed models are robust speech recognizers that can be used directly for oral history archives of various domains. We are providing fine-tuned speech recognizers for the research community and thus hopefully contributing to making the extremely valuable knowledge hidden in the vast oral history archives more accessible to the public and researchers.

11. Acknowledgements

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

12. References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[2] 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.

[3] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajič, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and Wei-Jing Zhu, "Automatic recognition of spontaneous speech for access to multilingual oral history archives," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 4, pp. 420–435, 2004.

[4] J. Psutka, P. Ircing, J. V. Psutka, J. Hajič, W. Byrne, and J. Mírovský, "Automatic transcription of Czech, Russian and Slovak spontaneous speech in the MALACH project," in *Eurospeech 2005*. ISCA, 2005, pp. 1349–1352.

[5] J. Psutka, J. Švec, J. V. Psutka, J. Vaněk, A. Pražák, L. Šmídl, and P. Ircing, "System for fast lexical and phonetic spoken term detection in a Czech cultural heritage archive," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2011, no. 1, pp. 1–10, 2011.

[6] M. Picheny, Z. Tüske, B. Kingsbury, K. Audhkhasi, X. Cui, and G. Saon, "Challenging the boundaries of speech recognition: The MALACH corpus," in *Interspeech 2019*, 2019, pp. 326–330.

[7] J. Švec, J. Psutka, L. Šmídl, and J. Trmal, "A relevance score estimation for spoken term detection based on RNN-generated pronunciation embeddings," in *Interspeech 2017*, 2017, pp. 2934–2938.

[8] J. V. Psutka, A. Pražák, and J. Vaněk, "Recognition of heavily accented and emotional speech of English and Czech Holocaust survivors using various DNN architectures," in *Speech and Computer*, A. Karpov and R. Potapova, Eds. Cham: Springer International Publishing, 2021, pp. 553–564.

[9] J. Lehečka, J. V. Psutka, and J. Psutka, "Transformer-based automatic speech recognition of formal and colloquial Czech in MALACH project," in *Text, Speech, and Dialogue*, P. Sojka, A. Horák, I. Kopeček, and K. Pala, Eds. Cham: Springer International Publishing, 2022, pp. 301–312.

[10] B. Ramabhadran, S. Gustman, W. Byrne, J. Hajič, D. Oard, J. S. Olsson, M. Picheny, and J. Psutka, "USC-SFI MALACH Interviews and Transcripts English LDC2012S05," Philadelphia: Linguistic Data Consortium, https://catalog.ldc.upenn.edu/LDC2012s05, 2012.

[11] J. Psutka, V. Radová, P. Ircing, J. Matoušek, and L. Müller, "USC-SFI MALACH Interviews and Transcripts Czech LDC2014S04," Philadelphia: Linguistic Data Consortium, https://catalog.ldc.upenn.edu/LDC2014S04, 2014.

[12] V. Nägel and D. Wein, "Witnesses of the shoah: The visual history archive of the shoah foundation in school education," *From Testimony to Story. Video Interviews about Nazi Crimes. Perpectives in Four Countries (Education with Testimonies, Bd. 2)*, pp. 173–179, 2015.

[13] 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," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[14] J. Lehečka, J. Švec, A. Pražák, and J. V. Psutka, "Exploring Capabilities of Monolingual Audio Transformers using Large Datasets in Automatic Speech Recognition of Czech," in *Proc. Interspeech 2022*, 2022, pp. 1831–1835.

[15] 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 (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 993–1003. [Online]. Available: https://aclanthology.org/2021.acl-long.80

[16] U. Reichel, F. Schiel, T. Kisler, C. Draxler, and N. Pörner, "The BAS speech data repository," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 786–791. [Online]. Available: https://aclanthology.org/L16-1126

[17] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Proc. Interspeech 2022*, 2022, pp. 2278–2282.

[18] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[19] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural nets," in *ICML '06: Proceedings of the International Conference on Machine Learning*, 2006.

[20] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Y. Wang, Z. You, and Z. Yan, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," in *Proc. Interspeech 2021*, 2021.

[21] 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)*, 2015, pp. 5206–5210.

[22] K. Heafield, "KenLM: Faster and smaller language model queries," in *Proceedings of the sixth workshop on statistical machine translation*, 2011, pp. 187–197.