Leveraging supplementary text data to kick-start automatic speech recognition system development with limited transcriptions

Nay San[1], Martijn Bartelds[2], Blaine Billings[3], Ella de Falco[4], Hendi Feriza, Johan Safri, Wawan Sahrozi, Ben Foley[5], Bradley McDonnell[3], Dan Jurafsky[1]
[1]Stanford University; [2]University of Groningen; [3]University of Hawai'i at Mānoa;
[3]University of Washington; [5]University of Queensland
nay.san@stanford.edu

Abstract

Recent research using pre-trained transformer models suggests that just 10 minutes of transcribed speech may be enough to fine-tune such a model for automatic speech recognition (ASR) — at least if we can also leverage vast amounts of text data (803 million tokens). But is that much text data necessary? We study the use of different amounts of text data, both for creating a lexicon that constrains ASR decoding to possible words (e.g. **dogz* vs. *dogs*), and for training larger language models that bias the system toward probable word sequences (e.g. *too dogs* vs. *two dogs*). We perform experiments using 10 minutes of transcribed speech from English (for replicating prior work) and two additional pairs of languages differing in the availability of supplemental text data: Gronings and Frisian (~7.5M token corpora available), and Besemah and Nasal (only small lexica available). For all languages, we found that using only a lexicon did not appreciably improve ASR performance. For Gronings and Frisian, we found that lexica and language models derived from 'novel-length' 80k token subcorpora reduced the word error rate (WER) to 39% on average. Our findings suggest that where a text corpus in the upper tens of thousands of tokens or more is available, fine-tuning a transformer model with just tens of minutes of transcribed speech holds some promise towards obtaining human-correctable transcriptions near the 30% WER rule-of-thumb.

1 Introduction

Automatic speech recognition (ASR) systems can help speed up the labour-intensive process of transcribing speech by providing human-correctable first-pass transcriptions. The Catch-22, however, is that some speech must first be manually transcribed in order to develop a sufficiently performant ASR system. As such, recently developed pre-trained transformer models for speech such as wav2vec 2.0 (Baevski et al., 2020) have spurred on much experimentation to leverage them to accelerate transcription workflows in language documentation contexts (e.g. Coto-Solano et al., 2022; Guillaume et al., 2022; Macaire et al., 2022; San et al., 2022; Zhang et al., 2022), given a much reduced upfront cost in terms of transcribed speech.

Impressively, Baevski et al. (2020) showed that a wav2vec 2.0 model pre-trained on 960 hours of untranscribed English speech required just 10 minutes of transcriptions to yield competitive results on the LibriSpeech ASR benchmark (Panayotov et al., 2015). This result, however, also leveraged the official LibriSpeech lexicon and language model derived from the entire 803 million token text corpus based on 14.5k public domain books.[1] However, an in-domain text corpus of such size is not within immediate reach for many other languages. In this paper, we investigate the real-world ASR performance achievable with 10 minutes of transcribed speech along with more tenable amounts of supplemental text data.

We first replicated the wav2vec 2.0 experiments on LibriSpeech and then extended them by creating lexica and language models using reduced amounts of supplemental text data (8M, 80k, and 8k tokens) from both the in-domain LibriSpeech text corpus and an out-of-domain corpus composed of web-scraped text (Common Crawl: Buck et al., 2014). Our experiments showed that first-pass transcriptions within the 20-30% word error rate (WER) rule-of-thumb (Gaur et al., 2016; Sperber et al., 2016) can indeed be obtained with just 10 minutes of transcribed speech if supplemented with a lexicon and language model derived from a corpus with at least 80k tokens.

We then performed analogous experiments with two pairs of languages differing in the availability of supplemental texts: Gronings and Frisian (two Germanic languages for which a modest amount

[1] https://www.openslr.org/11/

of external text-only data can be sourced), and Besemah and Nasal (two Malayo-Polynesian languages for which the documentation projects' own materials constitute the only available text data). For these experiments, we used 10 minutes of audio from each of the languages to fine-tune the multilingual wav2vec 2.0 XLS-R model (Babu et al., 2021) that is pre-trained on 450k hours of speech from 128 languages.

For all four languages we found that using only a lexicon to restrict ASR system output to possible words did not appreciably reduce the WER. For Gronings and Frisian, where there was sufficient text data to derive both lexica and language models from various samples (~7.5M, 80k, and 8k tokens), we found that those derived from 80k tokens of text reduced the WER to 34.9% for Gronings and 43.0% for Frisian (mean: 39.0%).

While these error rates are above the 20-30% rule-of-thumb for first-pass transcriptions (Gaur et al., 2016; Sperber et al., 2017), it is worth noting that the multilingual XLS-R model used for Gronings and Frisian was not pre-trained on the target datasets. This is in contrast to the English wav2vec 2.0 model, which was pre-trained on all 960 hours of the target LibriSpeech dataset. Thus, if combined with a domain adaptation technique (e.g. continued pretraining: Gururangan et al., 2020), our results suggest that where 80k or more of supplementary text is available, fine-tuning a pre-trained model with just tens of minutes of transcribed speech could help kick-start a virtuous cycle of data collection and training for ASR system development.

2 Motivations

2.1 Related work

As mentioned above, there have been several recent studies appraising the utility of pre-trained transformers for ASR in language documentation settings. Coto-Solano (2022) reported that the XLS-R model fine-tuned with 4 hours of transcribed speech from Cook Islands Māori yielded a WER of 22.9% without using a language model (LM). Similarly, Guillaume et al. (2022) found that the XLS-R model fine-tuned with 10 hours of transcriptions from Japhug yielded a WER of 18.5% also without a LM. Notably, both these studies explicitly mentioned that the system outputs (~20% WER) did indeed appear suitable as first-pass transcriptions for their respective projects and that LM integration is a clear next step.

Of the two studies that have examined the use of a LM with a fine-tuned wav2vec 2.0 model for language documentation projects (Macaire et al., 2022; San et al., 2022), a common theme has been to examine the effect of varying amounts of transcribed speech (e.g. 10–70 minutes) with and without the use of a LM trained on the full corpus (e.g. 74.5k tokens of text). Similarly, the original wav2vec 2.0 experiments by Baevski et al. (2020) also only examined the use of different amounts of fine-tuning data (10 minutes to 960 hours), with or without the use of LMs trained on an 803M token corpus.

Given the broader community interest in using fine-tuned wav2vec 2.0 models with LM integration, we undertook a series of experiments holding constant the amount of fine-tuning data (i.e. 10 minutes) while varying the amount of LM training text (e.g. 8k–8M tokens) to complement the aforementioned studies.

2.2 Language projects

For Besemah, Nasal, and Gronings, the motivation for the development of ASR systems is to help derive first-pass transcriptions and hence accelerate the process of indexing a large collection of audiovisual materials. Besemah (ISO 639-3: pse) and Nasal (ISO 639-3: nsy) are two Malayo-Polynesian languages spoken in Sumatra, Indonesia. For both languages, approximately 45 hours of informal conversations have been collected as part of fieldwork by author BM. Part of these collections have been transcribed by author BM and collaborators from the Besemah (author HF) and Nasal communities (authors JS and WS). The collections are managed by authors BB and EF and are accessible at PARADISEC (McDonnell, 2008, 2019).

Gronings (ISO 639-3: gos) is a Low Saxon language variant spoken in the province of Groningen in the Netherlands. An ongoing language documentation project (of which author MB is part) has so far recorded approximately 15 hours of speech, with more being continually gathered. The materials will be used to create ASR and text-to-speech systems to be made available through an online cultural portal.[2] For an additional point of comparison in our experiments, we also included West Frisian (another minority language spoken in the Netherlands; ISO 639-3: fry), using data from the FAME!

[2] https://www.woordwaark.nl

ASR corpus (Yilmaz et al., 2017).

3 Method

For English, we used the '1h/0' ten minute training set defined in Libri-Light (Kahn et al., 2020) to approximate the fine-tuning experiments in Baevski et al. (2020). As supplementary text data, we used the official normalised LibriSpeech as the in-domain corpus and Common Crawl (Buck et al., 2014) as the out-of-domain corpus.

For Besemah, Nasal, Gronings and Frisian, we constructed comparable 4-hour datasets for our broader set of ASR experiments. Each dataset is composed of an 80/10/10 train/dev/test split, approximately 24 minutes for each of the dev and test sets and 3.2 hours for training set, from which we sampled 10 minutes for the experiments reported here. As supplementary text data, we sourced two 9.5k word lists for Nasal and Besemah from project materials and remaining transcriptions not included in the ASR corpora. For Gronings and Frisian, we sourced two ~7.5M corpora used in de Vries et al. (2021).

For English, we used the monolingual English wav2vec 2.0 model (Baevski et al., 2020). For the other four languages, we used the multilingual XLS-R model (Babu et al., 2021). For fine-tuning these models, we used the HuggingFace Transformers library (Wolf et al., 2019). For beam search decoding with a lexicon and language model, we used the torchaudio implementation of the Flashlight decoder (formerly wav2letter: Kahn et al., 2022) used in the original paper. We fine-tuned each model for 12k steps and then selected the best checkpoint based on the dev set WER using greedy decoding (for Besemah and Nasal) or beam search decoding with fixed parameters (language model weight: 2, word insertion penalty: -1). We then performed a parameter search with the best checkpoint to further optimise the dev set WER. We repeated the decoding experiments with 5 different random samples of each size of sub-corpora. All our experiment code is available on GitHub.[3]

4 Results and discussion

The results from our experiments are collated in Table 1. For English, we found that fine-tuning with 10 minutes of transcribed speech yielded a test set WER of 40.2 (Row E1) without the use of supplementary texts and a WER of 14.2 (Row E3) using a lexicon and language model derived from the full 803M LibriSpeech text corpus. These WERs are consistent with those reported by Baevski et al. (2020, Table 9), respectively: 45.3 and 13.1, allowing for some error likely from differences in the 10 minute samples.[4]

For all languages, we found that using only a lexicon did not appreciably reduce the WER. For example, for English, the test set WER remained practically the same with or without a lexicon (Row E1 vs. E2: 40.2 vs. 40.5). For Nasal, the test set WER increased from 70.7 without a lexicon (Row N1) to 75.1 when using a small 9.5k token lexicon (Row N2), which led to many erroneous substitutions from the combination of an already high WER and high out-of-vocabulary rate.

For English, Gronings, and Frisian, we found that using lexica and language models derived from 80k of out-of-domain supplementary texts appreciably reduced the mean test set WER. For English, the test set WER was reduced from 40.2 (Row E1) to 27.7 (Row E8). For Gronings, the test set WER was reduced from 44.0 (Row G1) to 34.9 (Row G4). For Frisian, the test set WER was reduced from 53.1 (Row F1) to 43.0 (Row F4). These results suggest that where supplemental texts in the upper tens of thousands or more are available, fine-tuning a wav2vec 2.0 model with just tens of minutes of speech holds promise for deriving first-pass transcriptions near the 20-30% rule-of-thumb.

For English, the acceptably low sub-30% WER likely reflect an idealised set of circumstances in that the wav2vec 2.0 model used was pre-trained on the target dataset and that the genre of the audio is read speech. Regarding genre, the higher proportion of read speech in the Gronings dataset likely reflects its lower WER compared to Frisian, which has mainly news and radio broadcasts. Regarding pre-training, the multilingual XLS-R model used for Gronings and Frisian was not pre-trained on these datasets. This disadvantage could be partially overcome by a domain adaptation technique such as continued pre-training (Gururangan et al., 2020) or by using a wav2vec 2.0 model pre-trained on a similar language such as Dutch (Bartelds and Wieling, 2022). Macaire et al. (2022) report a 5–9% reduction in WER when using a wav2vec 2.0

[3] https://anonymous.4open.science/r/w2v2-10min-exps-computel6

[4] There was no indication as to which of the six 10-minute Libri-Light training sets were used in the original experiments.

model pre-trained on French over a multilingual model when fine-tuning for ASR on Gwadloupéyen and Morisien, two French-based creole languages.

For languages where tens of thousands of tokens in supplementary texts are not available, we suspect it remains unavoidable for the time being that more than 10 minutes of transcribed speech is required to begin ASR system development. For reducing the out-of-vocabulary rate with lexica and LMs derived from small text corpora, decomposing words into sub-word units (e.g. syllables) may be worth investigating. In trialling an interactive transcription app for Kunwinjku, Le Ferrand et al. (2022) report that incorporating syllable unigram frequencies derived from a word list improved the word retrieval performance by 4% (F-score).

5 Conclusion

We investigated the real-world performance obtainable when leveraging various amounts of supplemental text data to help kick-start automatic speech recognition (ASR) system development with only a limited amount of transcribed speech. Our results suggest that fine-tuning a pre-trained transformer model with just 10 minutes of transcribed speech may hold some promise for deriving human-correctable first-pass transcriptions if the ASR system can incorporate a lexicon and language model derived from a 'novel-length' corpus with at least 80,000 tokens or more of text.

Acknowledgements

We would like to thank people from the Nasal and Besemah communities who took part in these documentation projects, especially Anton Supriyadi, Sarkani, Asfan Fikri Sanaf, and Kencana Dewi as well as collaborators on the Nasal project, including Yanti and Jacob Hakim. We are also grateful to the Ministry of Research and Technology in Indonesia for providing permissions for research on Besemah and Nasal as well as the Center for Culture and Language Studies at Atma Jaya Catholic University of Indonesia for sponsoring this research. This material is based upon work supported by the National Science Foundation under Grant No. (1911641). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Martijn Bartelds and Martijn Wieling. 2022. Quantifying language variation acoustically with few resources. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3735–3741, Seattle, United States. Association for Computational Linguistics.

Christian Buck, Kenneth Heafield, and Bas van Ooyen. 2014. N-gram counts and language models from the common crawl. In *Proceedings of the Language Resources and Evaluation Conference*, Reykjavik, Iceland.

Rolando Coto-Solano. 2022. Evaluating word embeddings in extremely under-resourced languages: A case study in Bribri. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4455–4467, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Rolando Coto-Solano, Sally Akevai Nicholas, Samiha Datta, Victoria Quint, Piripi Wills, Emma Ngakuravaru Powell, Liam Koka'ua, Syed Tanveer, and Isaac Feldman. 2022. Development of automatic speech recognition for the documentation of Cook Islands Māori. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3872–3882, Marseille, France. European Language Resources Association.

Wietse de Vries, Martijn Bartelds, Malvina Nissim, and Martijn Wieling. 2021. Adapting monolingual models: Data can be scarce when language similarity is high. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4901–4907, Online. Association for Computational Linguistics.

Yashesh Gaur, Walter S Lasecki, Florian Metze, and Jeffrey P Bigham. 2016. The effects of automatic speech recognition quality on human transcription latency. In *Proceedings of the 13th International Web for All Conference*, pages 1–8.

Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato,

wav2vec 2.0 model (Fine-tuning data, 10 mins)	Language model (Text corpus size, no. of tokens)	Mean WER (SD)		Mean CER (SD)	
		dev	test	dev	test
LS-960 (English)	E1. None	40.4	40.2	13.0	12.8
	E2. None, lexicon only (973k)	40.1	40.5	12.3	12.2
	LIBRISPEECH (in-domain)				
	E3. 4-gram, full corpus (803M)	13.4	14.2	6.51	6.75
	E4. 4-gram, subset (8M)	15.9 (0.09)	16.6 (0.07)	7.41 (0.05)	7.58 (0.07)
	E5. 3-gram, subset (80k)	23.5 (0.13)	24.1 (0.12)	10.5 (0.07)	10.6 (0.10)
	E6. 3-gram, subset (**8k**)	**36.5** (0.20)	**36.8** (0.34)	17.1 (0.09)	17.0 (0.15)
	COMMONCRAWL-EN (out-of-domain)				
	E8. 3-gram, subset (**80k**)	**27.7** (0.13)	**27.7** (0.10)	12.1 (0.13)	12.0 (0.13)
	E9. 3-gram, subset (8k)	39.8 (0.12)	39.8 (0.24)	18.5 (0.19)	18.3 (0.22)
XLS-R (Gronings)	G1. None	43.8	44.0	11.2	11.3
	G2. None, lexicon only (216k)	41.4	40.7	10.5	10.3
	G3. 4-gram, full corpus (7.6M)	23.4 (0.22)	22.3 (0.22)	7.61 (0.62)	7.44 (0.71)
	G4. 3-gram, subset (**80k**)	**35.6** (0.30)	**34.9** (0.51)	11.8 (0.28)	11.4 (0.34)
	G5. 3-gram, subset (8k)	45.6 (0.41)	46.3 (0.24)	16.9 (0.67)	16.7 (0.43)
XLS-R (Frisian)	F1. None	55.1	53.1	18.6	18.3
	F2. None, lexicon only (251k)	53.8	51.7	17.9	17.6
	F3. 4-gram, full corpus (7.4M)	36.7	35.2	16.3	15.9
	F4. 3-gram, subset (**80k**)	**44.4** (0.37)	**43.0** (0.22)	20.3 (0.16)	19.9 (0.22)
	F5. 3-gram, subset (8k)	54.2 (0.39)	52.2 (0.29)	26.2 (0.30)	26.0 (0.25)
XLS-R (Besemah)	B1. None	62.3	62.1	21.4	21.2
	B2. None, lexicon only (9.5k)	59.9	62.2	20.1	21.2
XLS-R (Nasal)	N1. None	67.2	70.7	23.1	25.5
	N2. None, lexicon only (9.5k)	70.1	75.1	24.1	26.1

Table 1: Results of fine-tuning wav2vec 2.0 models (LS960: pre-trained on 960 hours of English; XLS-R pre-trained on 400k hours of speech from 128 languages) for automatic speech recognition (ASR) using only 10 minutes of transcribed speech from a target language and with varying amounts of reliance on supplementary text data. ASR performance measured by word error rate (WER) and character error rate (CER). For subset experiments, reported means and standard deviations (in parentheses) were derived across 5 runs with different samples of the text corpus. Alphanumeric labels (E1–N1) used to assist with in-text references to results. Highlighted numbers indicate the smallest supplementary text corpus size which yielded an appreciably lower WER.

Minh-Châu Nguyên, and Maxime Fily. 2022. Fine-tuning pre-trained models for automatic speech recognition, experiments on a fieldwork corpus of Japhug (Trans-Himalayan family). In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–178.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. IEEE.

Jacob D Kahn, Vineel Pratap, Tatiana Likhomanenko, Qiantong Xu, Awni Hannun, Jeff Cai, Paden Tomasello, Ann Lee, Edouard Grave, Gilad Avidov, et al. 2022. Flashlight: Enabling innovation in tools for machine learning. In *International Conference on Machine Learning*, pages 10557–10574. PMLR.

Éric Le Ferrand, Steven Bird, and Laurent Besacier. 2022. Fashioning local designs from generic speech

technologies in an Australian aboriginal community. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4274–4285, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Cécile Macaire, Didier Schwab, Benjamin Lecouteux, and Emmanuel Schang. 2022. Automatic speech recognition and query by example for creole languages documentation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2512–2520, Dublin, Ireland. Association for Computational Linguistics.

Bradley McDonnell. 2008. Besemah. PARADISEC Collection BJM01.

Bradley McDonnell. 2019. Documentation of the multilingual linguistic practices of the Nasal speech community. PARADISEC Collection BJM02.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: an ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Nay San, Martijn Bartelds, Tolulope Ogunremi, Alison Mount, Ruben Thompson, Michael Higgins, Roy Barker, Jane Simpson, and Dan Jurafsky. 2022. Automated speech tools for helping communities process restricted-access corpora for language revival efforts. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 41–51, Dublin, Ireland. Association for Computational Linguistics.

Matthias Sperber, Graham Neubig, Satoshi Nakamura, and Alex Waibel. 2016. Optimizing computer-assisted transcription quality with iterative user interfaces. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1986–1992, Portorož, Slovenia. European Language Resources Association (ELRA).

Matthias Sperber, Graham Neubig, Jan Niehues, Satoshi Nakamura, and Alex Waibel. 2017. Transcribing Against Time. *Speech Communication*, 93C:20–30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Emre Yilmaz, Jelske Dijkstra, H Velde, H Heuvel, and DA van Leeuwen. 2017. Longitudinal speaker clustering and verification corpus with code-switching Frisian-Dutch speech. *INTERSPEECH*.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP help revitalize endangered languages? a case study and roadmap for the Cherokee language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1529–1541, Dublin, Ireland. Association for Computational Linguistics.