Transfer Learning and Language Model Adaption for Low Resource Speech Recognition

Tasha Baller*, Kaelen Bennett, Howard J. Hamilton
Department of Computer Science, University of Regina, Canada

Abstract

We train an end-to-end recurrent neural network along with an integrated n-gram language model to perform automatic speech recognition on a low resource dataset containing telephone speech. The dataset presents many challenges because it is highly disfluent, contains unique accents and word choices, and is of poor audio quality. Our proposed method uses both transfer learning and language model adaption to obtain a 42.27% Word Error Rate (WER), which improves on existing models (60.18% WER, 74.89% WER) and low resource models (79.82% WER). Nonetheless, the WER is much higher than the current benchmark for high resource languages. Thus, more research is needed to overcome the obstacles low resource speech presents to high quality automatic speech recognition.

Keywords: Automatic Speech Recognition, Transfer Learning, Language Model Adaption, DeepSpeech

1. Introduction

Automatic Speech Recognition (ASR) traditionally relied on sophisticated pipelines composed of multiple models to transform speech to text. In contrast, recent approaches require only one end-to-end model that transforms speech to text via a neural network. End-to-end models not only require less human interaction but also have shown improvements in accuracy over traditional methods [1]. However, they require a significant amount of annotated data to provide good results. Thus, performing accurate ASR is difficult for *low resource speech*, which is any type of speech that is poorly represented in existing corpora and for which only a few hours of transcribed audio are available. The lack of data could be due to costs of transcription or barriers to obtaining many hours of speech.

We currently have a low resource dataset comprising recorded telephone calls of English speaking inmates from Canadian corrections facilities. This speech is spontaneous with a high level of disfluency and degraded quality, due to the nature of telephone audio with much background noise. Many individuals in the dataset are Indigenous and have an accent that is unique to North America. Additionally, some lack linguistic knowledge, mispronounce words, and use unusual phrasing and slang. Since many speech datasets consist of prepared speech in controlled environments, they are not representative of the real world problem of recognizing such challenging speech. Furthermore, these speakers have little to no representation in existing speech corpora. Thus, there is a large mismatch between existing ASR systems' training data and the characteristics of our data, resulting in poor performance.

To address this problem, we develop an end-to-end ASR system that is specifically tuned to recognize inmate-like speech. We propose using transfer learning to adapt a recurrent neural network (RNN) trained to recognize generic English to the environment and speech characteristics in our dataset. For low resource data, end-to-end models perform better with external language constraints, so we incorporate an n-gram language model to augment the adapted RNN. A language model trained on standard text corpora will not contain slang or other phrases present in dataset. Thus, we apply language model adaption to incorporate the unique characteristics of the speech present in our dataset. The proposed approach

*baller@uregina.ca

provides significant improvements over models trained on large corpora of generic English or solely on our low resource data.

The remainder of the paper is organized as follows: Section 2 outlines prior relevant work; Section 3 outlines the proposed method; Section 4 describes our experimental design; Section 5 presents the results; and Section 6 provides conclusions and future work.

2. Relevant Prior Work

Transfer learning is an emerging deep learning technique that promotes knowledge reuse. Transfer learning aims to transfer knowledge from one domain to another by finding commonalities between the existing domain and the target domain. By doing so, the target model is found more quickly, is more robust, and is less dependent on the target training data. These characteristics make transfer learning useful for low resource ASR, including transferring across languages and across speaker characteristics [2].

Transferring across languages is often done to adapt a model trained on a high resource language to a low resource language. Many languages share consonants and vowels and only differ in how they arrange them. Taking advantage of these commonalities can overcome some of the challenges of recognizing low resource languages. This technique has been applied for a variety of low resource languages, including Tunjia [3], Persian [4], and Swiss German [5], to greatly improve the accuracy of the models.

Another use of transfer learning is to tune a model to the characteristics of different speakers. Speakers vary in their vocal tract lengths, how they pronounce words, and how they arrange words to form sentences. These variations result from differences in sex, age, accent, and overall linguistic knowledge. All possible variations cannot be fully integrated into a training set and thus performance degradation can occur when dealing with under-represented speakers. Shivakumar and Georgio [6] achieved a 9.1% relative improvement in word error rate (WER) by adapting models originally trained for adults to children.

Although existing research has considered how transfer learning can improve ASR models, there is limited research on the compounding effects of adding an adapted language model. Adapting by interpolating between two or more language models has been shown to provide better performance than using a model trained on a single corpora [7]. Thus, due to the unique word selection and phrases of our dataset, a language model that interpolates a generic English model with an inmate-specific one is used to improve accuracy.

3. Proposed Method

Our proposed method uses a combination of transfer learning and language model adaptation to achieve ASR for our low resource speech. This method is demonstrated using Mozilla’s Project DeepSpeech (DeepSpeech) [8], which is based on Baidu’s research [9]. DeepSpeech is an end-to-end ASR system that incorporates an optional language model to transform audio composed of spoken words to the corresponding transcript. Mozilla provides a pre-trained RNN as the end-to-end model and a scorer, which consists of two sub-components, a 5-gram language model and its corresponding trie data structure. The scorer is used to integrate a language model into the system to guide the decoder to more accurate results. DeepSpeech, with pre-trained models, obtained a 5.97% WER on the LibriSpeech clean test [10].

RNN Training: The training is done via transfer learning using the pre-trained RNN provided by Mozilla. The pre-trained RNN was trained for easily understood generic English using a combination of telephone [11, 12] and read speech [13, 14]. These datasets were created for the use of research and although there are quality problems with the telephone data, the speakers enunciate well and the English is easily understood when compared with the speech in our low resource dataset. To adapt to the speech characteristics of our dataset, the model weights of the pre-trained RNN are used as a starting point for another phase of

training using our dataset. Low resource datasets can easily fall victim to overfitting, thus, we implemented early stopping [15] and dropout [16] during training.

Language Model Adaption: The language model adaption is performed using a combination of two 4-gram language models. For our low resource dataset, there are not enough inmate-like sentences to justify using anything larger than a 4-gram model. The first language model is trained using the same speech corpus as the Mozilla language model. This corpus is the normalized training text of the LibriSpeech dataset, which is composed of sentences from books. The LibriSpeech text corpus seemingly has little in common with inmate-like speech, but it contains basic information on English grammar and common words. Since our dataset contains many specialized words and phrases, a second language model is created by training exclusively on the transcripts from inmate calls.

Language model adaption is used to combine the two language models to create a mixed model that incorporates the foundational English from the first and the domain specific language from the second. The two models are combined using a mixture coefficient, λ , which is determined by how well each model predicts the word sequences of a held out validation set of inmate transcripts. For a word sequence W , the probability of the word sequence in the mixed model $p_m(W)$ is computed by Eq. 3.1, where $p_g(W)$ and $p_s(W)$ are the probabilities of word sequence W in the generic English and domain specific language models, respectively. The mixed language model and a generated trie structure are combined to form a scorer used for inference in the system.

$$p_m(W) = (1 - \lambda) \cdot p_g(W) + \lambda \cdot p_s(W) \quad (3.1)$$

4. Experimental Design

We ran all experiments on a PC with one GeForce GTX 1080 GPU running Ubuntu 18.04.2. We used Mozilla’s Project DeepSpeech version 0.7.4 [8]. For language model creation, we used SRILM version 1.7.3 [17] and KenLM [18].

Experiments: We perform two experiments, one for speech recognition and another for speaker adaptation. With a limited number of speakers, an end-to-end model can overfit to speaker characteristics by simply adapting to the speakers in the training set instead of generalizing to their shared characteristics. For speech recognition, we evaluate our proposed system on inmate-like speech from a set of speakers disjoint from those in the training set. For speaker adaptation, we evaluate on samples of inmate-like speech that are separate from the training set but are from the same speakers of the training set.

Dataset: The raw data consists of confidential telephone calls between inmates callers and non-inmate receivers. The majority of these non-inmate speakers share demographics with the inmate speakers and have similar speech characteristics. The audio was provided in Stereo 8kHz MP3 and was converted to Mono 16kHz WAV for compatibility. The audio files were segmented into samples of maximum length 12 seconds to create a total of 15,575 samples. Each sample was meticulously transcribed in lowercase letters without punctuation except for apostrophes and spaces. The samples make up approximately 18 hours of audio containing 18 male and 14 female speakers. Although there is a roughly equal split between male and female speakers, approximately 70% of the data is of male voices.

From the 32 speakers, three speakers containing 1,246 samples and approximately 1.5 hours of audio were withheld from the dataset to provide an assessment on unseen speakers. Of these three speakers, two are male and one is female. The remaining data were arranged into 10 random sets of training (70%), validation (20%), and test (10%) data to allow 10-fold cross validation. This division of samples provides approximately 11.5 hours of inmate-like speech for training, 3.5 hours for validation, and 1.5 hours for testing.

Systems: We compare the performance of four ASR systems, including three which use the DeepSpeech architecture, on inmate-like speech. The Mozilla system uses the pre-trained models provided by Mozilla. The Low Resource (LR) system uses an RNN and a 4-gram language model trained solely on our low resource data. Mozilla-LR is the proposed system, where the RNN and language model training is described in Section 3. Additionally, we include the performance of a state of the art system, known as wav2vec 2.0 [19], which achieved a 1.8% WER on the LibriSpeech clean test.

Parameters: Table 1 outlines the parameters of the ASR systems. The Mozilla RNN was trained in three phases, where the best weights were restored after each phase. To train the LR and Mozilla-LR RNNs, all parameters were set to DeepSpeech’s default values except for the dropout (0.05) and learning rate (0.001). Due to the small size of the dataset, a larger dropout rate of 0.4 for the input and hidden layers, along with a small learning rate of 0.00001, produced good results. Early stopping was used for both the LR and Mozilla-LR RNNs. The LR RNN took 17 epochs to converge, while the Mozilla-LR RNN took 14.

To integrate the language model, the language model weight (α) and word insertion weight (β) are required. These values are determined by running the provided language model optimizer on the trained RNN using the validation set. Due to time and resource constraints, the values for the LR and Mozilla-LR models were determined using the first fold’s RNN model with corresponding validation set and used for all 10 folds. The mixture coefficient, (λ), used for the language model adaption, was calculated separately using SRILM’s compute-best-mix script for the 10 folds; the mean value was 0.1430.

Table 1. Parameters for the ASR Systems

Parameter	Mozilla	LR	Mozilla-LR	Parameter	Mozilla	LR	Mozilla-LR
Train batch size	128	1	1	Epochs (1)	125	17	14
Val batch size	128	1	1	Epochs (2)	100	-	-
Test batch size	128	1	1	Epochs (3)	100	-	-
Hidden units	2048	2048	2048	Dropout rate	0.4	0.4	0.4
Learning rate (1)	1e-4	1e-5	1e-5	Beam width	1024	1024	1024
Learning rate (2)	1e-5	-	-	α	0.9313	0.5376	0.9612
Learning rate (3)	5e-6	-	-	β	1.1834	0.8539	2.7631

5. Results

Results for the ASR systems are shown in WER over the 10 folds. Table 2 provides a comparison on three unseen speakers, including two males and one female. These results reflect the performance on the speech recognition task. Table 3 provides the results on the test set. The instances in this test set are disjoint from those in the training set, but the speakers are not. Thus, the results reflects the performance of the LR and Mozilla-LR systems on the speaker adaptation task. Since the language model is optional for DeepSpeech systems, we report the results with the language model (RNN + LM) and without (RNN).

First, consider the results using only the RNNs. With a 51.33% WER the Mozilla-LR RNN, which was trained on both generic English and inmate-like speech, outperformed the other two RNNs. The Mozilla RNN had a 76.88% WER, illustrating the mismatch between its training and our data was substantial. Additionally, due to the lack of training data, the LR RNN resulted in even worse performance, with a 93.96% WER. With the integration of a language model, improvements were seen with all three systems. However, since the Mozilla language model was not trained on any inmate-like speech, only a slight improvement to 74.89% WER was seen. The LR system improved to 79.82% WER and the Mozilla-LR system improved to 42.27% WER. Due to our specialized vocabulary, the language models trained with our data played an important role in producing more accurate transcripts.

The LR and Mozilla-LR systems perform worse on the female speaker than on either male speaker, as might be expected with only 30% female representation in the dataset.

Table 2. Speech Recognition Results (%WER) for the ASR Systems

System	Male 1	Male 2	Female	All
Mozilla (RNN)	68.46	80.99	77.62	76.88
LR (RNN)	90.74	90.47	97.87	93.96
Mozilla-LR (RNN)	44.30	50.60	54.94	51.33
Mozilla (RNN + LM)	68.13	79.91	74.24	74.89
LR (RNN + LM)	74.82	76.51	84.40	79.82
Mozilla-LR (RNN + LM)	35.91	42.26	45.08	42.27
wav2vec 2.0	54.64	64.00	59.87	60.18

Thus, for the LR RNN, the number of training instances was likely inadequate for female speakers. For the Mozilla-LR RNN, transfer learning with mostly male speech may have caused the RNN to forget some female characteristics from the original Mozilla RNN, known as catastrophic forgetting [20]. However, in our case the degree of forgetting is not severe. The female speaker had a slightly higher WER (45.08%) than Male1 (35.91%) and Male2 (42.26%).

As expected, both LR and Mozilla-LR systems performed better with speakers included in their training. The LR system, with the integrated language model, had a 69.50% WER for seen speakers and a 79.82% WER for unseen speakers, a difference of 10.32% WER. On the other hand, the Mozilla-LR system had a difference of only 6.33% WER from 35.94% WER for seen speakers to 42.27% WER for unseen speakers. Thus, a greater degree of speaker adaptation occurred with LR than Mozilla-LR, perhaps due to the lack of diversity among the speakers in the training set for the LR RNN. The LR RNN was exposed to only 29 speakers, whereas the Mozilla-LR was originally trained with thousands of speakers in Mozilla’s training set before it was adapted to the speakers of our dataset. Furthermore, the level of speaker adaptation that occurred with the Mozilla-LR system is small. By looking at Male 1, who has the clearest diction, we see a 35.91% WER, which is comparable to the performance with previously seen speakers (35.94% WER). Speakers Male 2 and Female display a greater decrease in performance but also had more challenging speech in general as seen by higher WER for those speakers with all ASR systems. These results indicate our proposed method generalizes well to unseen speakers.

Table 3. Speaker Adaptation Results (% WER) for the ASR Systems

System	Test Set	System	Test Set	System	Test Set
Mozilla (RNN)	73.48	Mozilla (RNN + LM)	72.17	wav2vec 2.0	52.59
LR (RNN)	85.00	LR (RNN + LM)	69.50		
Mozilla-LR (RNN)	44.24	Mozilla-LR (RNN + LM)	35.94		

6. Conclusions and Future Work

Our principle contribution is the creation and evaluation of ASR suited to a novel dataset containing speech from Canadian correctional facilities. These speakers have little to no representation in existing systems and thus they do poorly on this speech. Additionally, the speech displays disfluency, accents, unusual word choices, and has degraded audio quality, which makes it challenging to recognize. Our proposed method uses a combination of transfer learning and language model adaption to tailor generic models to the unique characteristics of our data. Although previous work has demonstrated the effectiveness of transfer learning for low resource languages, there is limited research on its use for under-represented speakers like the ones in our dataset. Furthermore, there is limited research into the effectiveness of combining transfer learning and language model adaption for low resource speech.

The results of transfer learning and language model adaption for low resource data are abundantly clear. The proposed Mozilla-LR system (42.27% WER) was significantly more

accurate than the original Mozilla system (74.89% WER), the LR system trained solely on the low resource data (79.82% WER), and a state of the art system, wav2vec 2.0 (60.18% WER). The results also indicate the efficacy of using an integrated language model with an end-to-end model. Improvements were made to three systems by adding a language model, but due to the specialized vocabulary of our dataset, the language models trained, in part or exclusively, with inmate-like speech provided the greatest improvements. With the integration of a language model, the LR system showed the greatest improvement, with a change from 93.96% WER to 79.82% WER.

Future work could improve our models by preventing catastrophic forgetting during the RNN training by incorporating more female speakers and by expanding the training corpora for our language model.

References

- [1] D. Wang, X. Wang, and S. Lv. “An Overview of End-to-End Automatic Speech Recognition”. In: *Symmetry* 11 (Aug. 2019), p. 1018.
- [2] D. Wang and T. F. Zheng. “Transfer Learning for Speech and Language Processing”. In: *APSIPA*. 2015, pp. 1225–1237.
- [3] C. Yu, Y. Chen, Y. Li, M. Kang, S. Xu, and X. Liu. “Cross-Language End-to-End Speech Recognition Research Based on Transfer Learning for the Low-Resource Tujia Language”. In: *Symmetry* 11 (Feb. 2019), p. 179.
- [4] M. Asadolahzade. “Transfer Learning for ASR to Deal with Low-Resource Data Problem”. Tehran, Iran, May 2019.
- [5] A. Agarwal and T. Zesch. “LTL-UDE at Low-Resource Speech-to-Text Shared Task: Investigating Mozilla DeepSpeech in Low-Resource Setting”. In: *SwissText & Konvens*. June 2020.
- [6] P. G. Shivakumar and P. Georgiou. “Transfer Learning from Adult to Children for Speech Recognition: Evaluation, Analysis and Recommendations”. In: *Computer Speech & Language* 63 (Sept. 2020), p. 101077.
- [7] A. Sethy, S. Chen, B. Ramabhadran, and P. Vozila. “Static Interpolation of Exponential N-gram Models Using Features of Features”. In: *IEEE ICASSP*. 2014, pp. 4878–4882.
- [8] *Project DeepSpeech*. Version 0.7.4. June 2020. URL: <https://github.com/mozilla/DeepSpeech>.
- [9] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng. *Deep Speech: Scaling Up End-to-End Speech Recognition*. 2014. arXiv: [1412.5567](https://arxiv.org/abs/1412.5567) [cs.CL].
- [10] R. Morais. *Release DeepSpeech 0.7.4 - Mozilla/DeepSpeech*. June 2020. URL: <https://github.com/mozilla/DeepSpeech/releases/tag/v0.7.4>.
- [11] C. Cieri, D. Miller, and K. Walker. “The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text”. In: *LREC*. May 2004.
- [12] J. J. Godfrey, E. C. Holliman, and J. McDaniel. “SWITCHBOARD: Telephone Speech Corpus for Research and Development”. In: *ICASSP*. Vol. 1. 4. 1992, pp. 517–520.
- [13] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. “LibriSpeech: An ASR Corpus Based on Public Domain Audio Books”. In: *ICASSP*. Apr. 2015, pp. 5206–5210.
- [14] 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*. Mar. 2020. arXiv: [1912.06670](https://arxiv.org/abs/1912.06670) [cs.CL].
- [15] M. Li, M. Soltanolkotabi, and S. Oymak. “Gradient Descent with Early Stopping is Provably Robust to Label Noise for Overparameterized Neural Networks”. In: *PMLR*. Vol. 108. Aug. 2020, pp. 4313–4324.
- [16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *JMLR* 15.1 (Jan. 2014), pp. 1929–1958.
- [17] A. Stolcke. “SRILM - An Extensible Language Modeling Toolkit”. In: *ICSLP*. July 2002, pp. 901–904.
- [18] K. Heafield. “KenLM: Faster and Smaller Language Model Queries”. In: *WMT*. July 2011, pp. 187–197.
- [19] A. Baevski, H. Zhou, A. Mohamed, and M. Auli. *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. 2020. arXiv: [2006.11477](https://arxiv.org/abs/2006.11477) [cs.CL].
- [20] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. “Overcoming Catastrophic Forgetting in Neural Networks”. In: *PNAS* 114.13 (2017), pp. 3521–3526.