Improving ASR output for endangered language documentation

Robbie Jimerson[1], *Kruthika Simha*[1], *Raymond Ptucha*[1], *Emily Prud'hommeaux*[1,2]

[1]Rochester Institute of Technology, Rochester, New York, USA
[2]Boston College, Chestnut Hill, Massachusetts, USA

{rcj2772,kps2151,rwpeec}@rit.edu, prudhome@bc.edu

Abstract

Documenting endangered languages supports the historical preservation of diverse cultures. Automatic speech recognition (ASR), while potentially very useful for this task, has been underutilized for language documentation due to the challenges inherent in building robust models from extremely limited audio and text training resources. In this paper, we explore the utility of supplementing existing training resources using synthetic data, with a focus on Seneca, a morphologically complex endangered language of North America. We use transfer learning to train acoustic models using both the small amount of available acoustic training data and artificially distorted copies of that data. We then supplement the language model training data with verb forms generated by rule and sentences produced by an LSTM trained on the available text data. The addition of synthetic data yields reductions in word error rate, demonstrating the promise of data augmentation for this task.

Index Terms: speech recognition, under-resourced languages, data augmentation, endangered language documentation

1. Introduction

Nearly half of the world's 7000 languages are considered endangered, and it is predicted that the majority of these languages will not survive into the next century. Collecting comprehensive audio and textual evidence of these languages will facilitate the preservation of the languages and the cultures of the people who speak them. Unfortunately, many communities of speakers lack the time and financial resources necessary to carry out this work. Automatic speech recognition (ASR) has the potential to support language documentation efforts by both field linguists and community stakeholders, but building robust ASR models from the limited training data that is typically available for such languages presents challenges.

In the past several years, the use of deep neural networks (DNN) for training ASR acoustic models has resulted in substantial reductions in word error rate (WER) for high-resource languages. DNN frameworks, however, typically require very large amounts of data, making them less useful for the low-resource scenarios typically encountered with endangered languages. Some of the more successful work on using a DNN ASR frameworks in low-resource scenarios has focused on adapting robust acoustic models built using abundant multilingual data to the target language. Acoustic data augmentation via speech signal distortion has also been used to supplement in-domain acoustic training data. There has been little work, however, on improving the language models within the DNN framework.

In this paper, we compare several ASR frameworks to identify ways of improving an ASR system for Seneca, an endangered language of the Northern Iroquoian language family spoken natively by fewer than 100 people primarily in Western New York State and Ontario. In order to determine the optimal approach for working with the sort of extreme low-resource situations typically associated with endangered languages in need of documentation, we explore both a traditional GMM/HMM framework and a DNN framework that has been shown to outperform the GMM/HMM approach in high-resource languages.

We find that the GMM/HMM ASR framework outperforms DNN-based ASR when acoustic training is limited to the small amount of available Seneca audio data. Using a prebuilt DNN acoustic model trained on large quantities of English data and adapting that model to Seneca via transfer learning results in significant reductions in word error rate. Data augmentation via duplication of existing audio data modified by speech signal distortion results in further improvements. The GMM/HMM framework trained only on the small amount of available Seneca audio data can be improved by supplementing the text data used to built the language model. These results highlight the potential for improving the quality of ASR output under low-resource conditions in both DNN and GMM/HMM frameworks via acoustic and text data augmentation while also revealing weakness in the DNN framework when applied to an under-resourced language.

2. Background

There are estimated to be fewer than 100 native speakers of Seneca, with an additional one hundred to two hundred second-language learners. Like all Iroquoian languages, Seneca is characterized by a relatively small segment inventory, a challenging accentual system, and a highly complex morphology, in which a single verb can have thousands of possible forms and in which a noun can be inserted between a verb and its affixes Seneca is fairly well documented descriptively, but the amount of naturalistic and spontaneous text and speech data is minimal, with just a few hours of transcribed speech and very little additional digitized text available to researchers outside the Seneca community. The scarcity of textual and audio documentation of Seneca, combined with the dwindling number of native speakers, presents challenges not only to future attempts to revitalize the language, but also to the development of tools and technologies for preserving the language.

The last few years have seen an increased interest in developing robust automatic speech recognition (ASR) systems for low resource and under-resourced languages like Seneca. The majority of this recent research has focused on optimizing the acoustic model in order to overcome the constraints of a limited amount of labeled audio training data. Researchers have explored modifications in approaches used to train the acoustic models [1, 2, 3]; improvements in the features included in the models [4, 5, 6, 7, 8]; and supplementing the acoustic train-

ing data with data from other languages [5, 1, 9, 7]. Most of this work, however, has generally not focused specifically on endangered language documentation but instead on tasks like spoken term detection (e.g., [8, 10]) and phonetic transcription for languages without a writing system (e.g., [11, 12, 13, 14]).

Data augmentation, in which unattested or synthetic data is generated from existing data, was originally used in image processing tasks to increase the size of a dataset in order to avoid overfitting and to improve robustness of the models. This idea has recently been extended to speech data, and several teams have explored different kinds of acoustic augmentation techniques for speech data.

Hannun et al. [15] explored the impact of corrupting clean speech with noise and found that it improved the robustness of the ASR system against noisy speech. Jaitly and Hinton [16] successfully experimented with Vocal Tract Length Perturbation (VTLP) as an augmentation technique on the TIMIT phoneme recognition task, using DNN-based acoustic modeling. VTLP was further successfully tested by Cui et al. [17] and Ragni et al. [18]. Kanda et al. [19] used similar augmentation methods on low-resource languages, with acoustic training data of less than 10 hours. Ko et al. [20] experimented with the Switchboard benchmark task, using speed augmentation with various speeds for augmentation.

Comparatively little work has been done on augmentation of the text data used to train the language model. Some work on high-resource languages predating the introduction of DNNs for ASR explored using machine translation (MT) to create artificial text data for discriminative language modeling for reranking n-best lists rather than augmenting the language model training data itself [21, 22]. More recent work on low resource languages has used MT-generated [23] and recurrent neural network (RNN) generated [24] synthetic data to augment the language model training data.

3. Data

The Seneca data used to train our acoustic models consists of 155 minutes of recorded and transcribed spontaneous conversations and narratives provided by seven adult first-language Seneca speakers from two different reservations. Recordings were made over many years under a variety of conditions using various pieces of recording equipment, yielding a diverse set of audio data. Table 1 shows the number of minutes, utterances, and words per speaker.

The text used to train the language model totals 13585 words and 1843 utterances. This includes the transcripts of the audio used to train the acoustic model, several short texts produced from audio recordings that are no longer available, a few brief texts not derived from audio recordings, and the text content of the Seneca Reference Guide, a document created by the Seneca community for supporting language learning.

The held-out test data used to evaluate the various recognizers consists of 35 minutes of audio from two of the seven speakers.

Speaker A is from the Cattaraugus Seneca Reservation which is approximately 30 miles south of Buffalo, NY. He tells a short story about his Grandfather who hunts bears without using a gun. This recording was recorded and transcribed by Wallace Chafe, an emeritus professor of linguistics at UC Santa Barbara. Speaker B is from the Allegany Seneca Reservation which is located near Salamanca, NY. This speaker gives a brief description of his garden. This recording was also recorded and transcribed by Dr. Chafe. Speaker C, also from the Cattaraugus Seneca reservation, discusses the habits of deer. This was also recorded and transcribed by Dr. Chafe. Both speaker D and speaker E are from the Cattaraugus Seneca Reservation. These two speakers are recorded together discussing various topics as springtime, working and retirement. This data was recorded and transcribed by Dr. Chafe. Speaker F is from the Allegany Seneca reservation. This data totals 60 minutes of conversations in which the speaker discuss with other Seneca speakers a wide range of topics, including personal narratives and Seneca culture and folklore. This data was recorded and transcribed by the first author who is a member of the Seneca nation and a second-language speaker of Seneca. Speaker H is from the Cold Spring portion of the Allegany Seneca reservation. Her audio data consists of 75 minutes of conversations in Seneca with the first author. The topics in this recording are wide ranging and include the speaker's family and upbringing, various stories from her childhood, and current events.

	Minutes	**Words**	**Utterances**
Speaker A	3	139	20
Speaker B	2	126	21
Speaker C	4	265	20
Speaker D	11	451	133
Speaker E	11	1011	257
Speaker F	60	5931	491
Speaker H	75	4343	475
Total	**155**	**12266**	**1417**

Table 1: *Amount of acoustic training data by speaker.*

Sentences	**Words**	**Types**
1843	13584	2973

Table 2: *Amount of text to be used in language model training by sentence, word, and type.*

4. Methods

We explore two different ASR frameworks: the GMM/HMM framework of Kaldi [25] and the DNN framework of Deep Speech [15].

The Deep Speech model is a five layer recurrent neural network (RNN), with the first, second, third and the fifth layer being non-recurrent, and only the fourth layer being bi-directional recurrent. Each layer has 2048 hidden units. Within the Deep Speech framework we train on three different acoustic training data configurations: (1) training on the 155 minutes of Seneca data alone; (2) initialization of weights with a pre-trained Deep Speech English acoustic model, then fine-tune to the existing 155 minutes of Seneca data; and (3) modifying the transfer learning approach in (2) by augmenting the Seneca data with the addition of synthetic data of various types created by distorting the original Seneca audio data, as described below in Section 4.1.

Within the Kaldi GMM/GMM framework, we augment only the language model training data with: (1) unseen verb forms generated by a deterministic algorithm designed to account for phonological changes across morpheme boundaries; and (2) novel utterances generated by a long short term memory (LSTM) language model built on the full set of available text data.

4.1. Acoustic data augmentation

We experiment with three augmentation techniques: noise addition, pitch augmentation, and speed augmentation. For noise addition, ten background noises were used: waves at the beach, riding a bicycle, birds chirping, sounds of doing dishes, cat noises, sounds at the gym, table fan, rain, running water and subway. The noise signals were chosen such that the duration of each noise signal was longer than the longest speech sample (i.e., utterance). In order to distort an input speech sample, the speech sample was combined with a randomly chosen interval of the selected noise signal with duration equal to that of the speech sample, with a signal-to-noise ratio of 30dB.

For pitch augmentation, the pitch of the speech signal was varied in fractions of octaves. We experimented with 0.10, 0.15, 0.20, 0.25, and 0.30 fractions of an octave. The fraction of octave was chosen at random from the above values each time a speech signal was augmented.

Speed augmentation was done by re-sampling the speech data at 0.75, 0.80, 0.85, 0.90, 0.95, 1.00, 1.05, 1.10, 1.15, 1.20, and 1.25 multiples of the sampling frequency of the utterance. The multiple for the sampling frequency was chosen at random from the above values each time a speech signal was augmented.

In order to determine the impact of acoustic data augmentation on ASR output quality, we created two new augmented datasets: one in which each speech sample (i.e., utterance) was augmented 15 times, and one in which each speech sample was augmented 25 times. When combining the augmentation techniques, the augmentation technique and the associated parameters for that technique (e.g., the nature of the noise applied, the pitch modification, the change in speed) were applied at random for each augmentation of each sample.

A baseline acoustic model ("unaugmented Seneca") was trained on only the 155 minutes of unaugmented Seneca audio data using Deep Speech with default parameter settings. Three additional transfer-learning acoustic models (unaugmented transfer learning, augmented-15, and augmented-25) were trained under Deep Speech by importing the weights from a pre-trained English speech model and setting the $initialize\_from\_frozen\_model$ flag to the path of the output graph of the pre-trained model. We used the pre-trained English model provided by Mozilla, which was trained on the Fischer corpus, LibriSpeech, and Switchboard. These three transfer-learning Seneca models were then retrained from the pre-trained model using (1) the 155 minutes of Seneca data alone (unaugmented); (2) the 155 minutes of Seneca data plus the 15 augmentations of each speech sample totaling approximately 4500 minutes of Seneca data (augmented-15) ; and (3) the 155 of minutes of Seneca data plus the 25 augmentations of each speech sample totaling approximately 7500 minutes of Seneca data (augmented-25). The models were trained for 25 epochs on the pre-trained English Deep Speech model, with early stopping enabled.

4.2. Text data augmentation

In addition to augmenting the data used to train the acoustic model, we also investigated generating synthetic text data to supplement the small amount of available data used to train the language model. These experiments were carried out on our GMM/HMM-based baseline ASR system built using the Kaldi framework.

The first set of experiments consisted of generating synthetic Seneca sentences from a language model trained on the available text data and adding these sentences to the ASR language model training corpus. A word-based RNN/LSTM [26] was trained using the original 1843 Seneca sentences in our corpus. In order to generate the synthetic Seneca sentences from this model, a seed is used to produce the rest of the sentence. We selected as seeds the three most common utterance-initial words in our existing Seneca text data. We generated 1834 synthetic sentences, doubling the size of the corpus, with one third of the total number of synthetic sentences produced from each of the three seeds. When experimenting with adding a smaller number of synthetic sentences, sentences were randomly chosen from the total set of 1834 synthetic sentences.

One of the challenges developing an ASR system for a polysynthetic language is the unusually high out of vocabulary (OOV) rate. In the second experiment, we tried to address this by adding synthetic verb forms to the lexicon. Verbs roots were identified in the original Seneca corpus and the most common forms of these verb roots were computationally generated using a rule-based algorithm for generating verb forms from a verb root, while observing the complex phonological changes that occur across the various morpheme boundaries. Verb forms were similarly generated for a number of commonly used verbs roots that were lacking in the existing corpus. In total, 6013 verbs forms were added to the to the lexicon, bringing the total number of words in the lexicon to 8551.

The acoustic model for these experiments was created following the Kaldi for Dummies tutorial recipe, which uses the standard thirteen dimensional Cepstral mean-variance normalized MFCCs, plus their first and second derivatives, within a GMM framework. The recipe was extended to apply LDA transformation and Maximum Likelihood Linear Transform to the features. Other training techniques included boosted Maximum Mutual Information (bMMI) and Minimum Phone Error (MPE). Both bMMI and MPE were trained over 4 iterations and bMMI used a boost weight of 0.5.

5. Results

The first results listed in Table 3 (WER 95.03) and Table 4 (60.43) were produced using unaugmented, purely Seneca data under the two different ASR frameworks. These results represent the baseline of the deep learning and HMM models, respectively. They are also the results against which we compare the results of subsequent experiments.

Table 3 shows the word error rate (WER) obtained from the four acoustic models. We see that training a Deep Speech model on the very minimal amount of available Seneca data (155 minutes) results in a WER so high that it is likely that there is no usable output. After adapting via transfer learning an English speech model to this small amount of Seneca data, the WER is substantially reduced. Augmenting the Seneca data to which the model is adapted via signal distortion results in further reductions in WER, yielding a minimum WER of 65.9%. Although the results are not presented here, we note that when applied individually rather than jointly, only the speed augmentation method results in a WER reduction over the unaugmented transfer learning baseline.

Although these results show that both transfer learning and data augmentation yield significant reductions in WER in a DNN framework, the traditional GMM/HMM framework trained on only the 155 minutes of Seneca audio data results in a lower overall WER, as shown in Table 4. Table 4 also shows the results of adding synthetically generated text data for building the language model. Unlike earlier work [24] in which

Acoustic Model	WER
unaugmented Seneca	95.03
unaugmented transfer learning	70.43
augmented-15	68.33
augmented-25	65.84

Table 3: *Word error rates (WER) using Deep Speech with the four acoustic models described in Section 4.1.*

Language Model	WER
No augmentation	60.43
LM augmentation by 50%	61.57
LM augmentation by 75%	63.22
LM augmentation by 100%	65.36

Table 4: *Word error rates (WER) using Kaldi with three levels of augmentation of synthetic text to the corpus used to train the language model (LM).*

synthetic data produced in this way resulted in WER improvements, we found that adding the LSTM-generated sentences increased WER rate.

Table 5 shows the results obtained when synthetic verb forms were generated and added to the lexicon of the GMM/HMM model. We see a small decrease in WER, consistent with our previous findings on an even smaller Seneca dataset [27].

6. Conclusions

Although deep learning ASR approaches are demonstrably superior to traditional GMM/HMM approaches for high resource languages such as English, it remains to be seen whether the utility of deep learning can be fully exploited in languages like Seneca that have extremely limited audio and text resources. In our work using deep learning, we were unable to achieve the WER of a very basic GMM/HMM ASR system even with substantial modifications to the baseline DNN models via transfer learning and data augmentation. We also were unable to reduce WER by supplementing the language model training data with synthetic data produced by an LSTM language model, which we suspect is again due to the very limited amount of training data.

As the size of our Seneca training corpus increases in the course of our current language documentation project, we anticipate that the performance gap between the two approaches will decrease. In the meantime, we plan to explore in more depth methods for improving the language model and reducing the number of potential OOVs. One obvious first step is to add the synthetic data we have created so far to the text corpus used to build the Deep Speech recognizer.

Until now we have been using a word-based LSTM to generate synthetic sentences. In our future work, we will use a character-based LSTM, which has the potential to generate not only novel utterances but also novel word forms, which could mitigate the OOV problem that makes working with polysynthetic languages so challenging.

Finally we would like to explore the use of machine translation for synthesizing language model training data, which has been shown to improve ASR output for both high- and low-resource languages. Although there currently is very little parallel data available that could be used to train a machine translation system, we are currently working to organize and digitize a number of historic Seneca texts that have been translated into English, and religious English texts that have been translated to Seneca.

Language Model	WER
No augmentation	60.43
Augmented lexicon	59.11

Table 5: *Word error rates (WER) using Kaldi before and after supplementing the lexicon with unseen verb forms generated via a phonologically-sensitive deterministic algorithm.*

As many of the world's languages become endangered, the demand for tools and technologies for language documentation will grow. Although ASR has the potential to serve as one of these tools, it may not be possible to rely on the frameworks typically used for building ASR systems for high-resource languages like English or even for under-resourced but widely spoken languages like Haitian Creole or Vietnamese. The work presented here is a first step in determining the most fruitful approaches for ASR for endangered languages.

7. Acknowledgements

The authors would like to thank the members of the Seneca Nation of Indians for their participation in this work. This material is based upon work supported by the National Science Foundation under Award No. 1761562. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

8. References

[1] F. Grézl, M. Karafiát, and K. Vesely, "Adaptation of multilingual stacked bottle-neck neural network structure for new language," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 7654–7658.

[2] Y. Miao, F. Metze, and S. Rawat, "Deep maxout networks for low-resource speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 398–403.

[3] S. Thomas, M. L. Seltzer, K. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6704–6708.

[4] X. Cui, B. Kingsbury, J. Cui, B. Ramabhadran, A. Rosenberg, M. S. Rasooli, O. Rambow, N. Habash, and V. Goel, "Improving deep neural network acoustic modeling for audio corpus indexing under the IARPA Babel program," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[5] M. J. Gales, K. M. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low-resource languages: Babel project research at CUED." in *SLTU*, 2014, pp. 16–23.

[6] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2494–2498.

[7] Z. Tüske, P. Golik, D. Nolden, R. Schlüter, and H. Ney, "Data augmentation, feature combination, and multilingual neural networks to improve asr and kws performance for low-resource languages," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[8] R. Prabhavalkar, K. Livescu, E. Fosler-Lussier, and J. Keshet, "Discriminative articulatory models for spoken term detection in low-resource conversational settings," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8287–8291.

[9] D. Imseng, P. Motlicek, H. Bourlard, and P. N. Garner, "Using out-of-language data to improve an under-resourced speech recognizer," *Speech communication*, vol. 56, pp. 142–151, 2014.

[10] A. Rosenberg, K. Audhkhasi, A. Sethy, B. Ramabhadran, and M. Picheny, "End-to-end speech recognition and keyword search on low-resource languages," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5280–5284.

[11] X. Kong, P. Jyothi, and M. Hasegawa-Johnson, "Performance improvement of probabilistic transcriptions with language-specific constraints," *Procedia Computer Science*, vol. 81, pp. 30–36, 2016.

[12] C. Liu, P. Jyothi, H. Tang, V. Manohar, R. Sloan, T. Kekona, M. Hasegawa-Johnson, and S. Khudanpur, "Adapting asr for under-resourced languages using mismatched transcriptions," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5840–5844.

[13] A. Das, P. Jyothi, and M. Hasegawa-Johnson, "Automatic speech recognition using probabilistic transcriptions in swahili, amharic, and dinka." in *INTERSPEECH*, 2016, pp. 3524–3528.

[14] M. A. Hasegawa-Johnson, P. Jyothi, D. McCloy, M. Mirbagheri, G. di Liberto, A. Das, B. Ekin, C. Liu, V. Manohar, H. Tang *et al.*, "Asr for under-resourced languages from probabilistic transcription," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 1, pp. 50–63, 2017.

[15] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[16] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013.

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

[18] A. Ragni, K. M. Knill, S. P. Rath, and M. J. Gales, "Data augmentation for low resource languages," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[19] N. Kanda, R. Takeda, and Y. Obuchi, "Elastic spectral distortion for low resource speech recognition with deep neural networks," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 309–314.

[20] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[21] K. Sagae, M. Lehr, E. Prud'hommeaux, P. Xu, N. Glenn, D. Karakos, S. Khudanpur, B. Roark, M. Saraclar, I. Shafran *et al.*, "Hallucinated n-best lists for discriminative language modeling," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 5001–5004.

[22] E. Dikici and M. Saraçlar, "Semi-supervised and unsupervised discriminative language model training for automatic speech recognition," *Speech Communication*, vol. 83, pp. 54–63, 2016.

[23] A. Gorin, R. Lileikyte, G. Huang, L. Lamel, J.-L. Gauvain, and A. Laurent, "Language model data augmentation for keyword spotting in low-resourced training conditions." in *Interspeech*, 2016, pp. 775–779.

[24] G. Huang, T. F. da Silva, L. Lamel, J.-L. Gauvain, A. Gorin, A. Laurent, R. Lileikyte, and A. Messouadi, "An investigation into language model data augmentation for low-resourced stt and kws," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5790–5794.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[26] S. Kim. (2017) word-rnn-tensorflow. [Online]. Available: https://github.com/hunkim/word-rnn-tensorflow

[27] R. Jimerson and E. Prud'hommeaux, "ASR for Documenting Acutely Under-Resourced Indigenous Languages," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018, pp. 4161–4166.