Interspeech 2018
2-6 September 2018, Hyderabad

Online Incremental Learning for Speaker-Adaptive Language Models

Chih Chi Hu, Bing Liu, John Paul Shen, Ian Lane

Electrical and Computer Engineering, Carnegie Mellon University, USA
{chihhu,liubing,jpshen,lane}@cmu.edu

Abstract

Voice control is a prominent interaction method on personal computing devices. While automatic speech recognition (ASR) systems are readily applicable for large audiences, there is room for further adaptation at the edge, ie. locally on devices, targeted for individual users. In this work, we explore improving ASR systems over time through a user's own interactions. Our online learning approach for speaker-adaptive language modeling leverages a user's most recent utterances to enhance the speaker dependent features and traits. We experiment with the Large-Vocabulary Continuous Speech Recognition corpus Tedlium v2, and demonstrate an average reduction in perplexity (PPL) of 19.18% and average relative reduction in word error rate (WER) of 2.80% compared to a state-of-the-art baseline on Tedlium v2.
Index Terms: Automatic Speech Recognition, Online Learning, Language Modeling, Speaker-Adaptation, Speaker-Specific Modeling, Recurrent Neural Networks

1. Introduction

Voice control is becoming an increasingly popular interaction method on personal computing devices, where interactions are limited to a single or a handful of users. Phones, laptops, and even vehicles, now support services in providing personalized recommendations and advertisements according to user's interests and needs. Speech recognition is one such area that can be leveraged to build user profiles and through real-time speaker adaption provide further enhancements based on user specific phrases, usage, and style. Voice assistants such as Apple Siri, Google Now, Microsoft Cortana, and Amazon Alexa, could provide a better interactive experience for all users if they could learn from its users through interactions and its own hypotheses (or references if available). We focus on adapting to the syntactic, semantic, and pragmatic characteristics of speech, which by design is captured by the language model. By leveraging the user's most recent utterances, we enhance speaker dependent features and traits in the recurrent neural network language model as well as implicitly capture the context to improve speech recognition for designated user(s) over time.

Prior work in speaker adaptation has broadly explored fine tuning or freezing various parameters or components in the ASR model. We utilize a simple approach of continuous mini batch training with varying epochs and batch sizes. With small defined epochs, a standard training strategy, and continuous online learning, our experiments show positive improvement for individual speakers over time.

In this paper, we use online incremental learning for language model speaker adaptation to improve performance and enhance robustness of automatic speech recognition systems. We train a state-of-the-art RNN language model, then during live inferences, we re-train on mini batches of utterances in an incremental and continuous fashion in real time. After each segment, an updated model is produced to be evaluated on the next segment. We utilize both ASR hypotheses and reference utterances for mini batch training to explore the effectiveness of online incremental learning. With online learning on reference utterances, our RNN model obtained an average of 20.09% reduction in PPL (Figure 2) and a reduction of up to 0.75% in absolute WER (Figure 1), corresponding to a relative WER reduction of 9.93%. Furthermore, online learning with references shows additional 0.98% relative improvement in average PPL and 0.10% absolute improvement in average WER on top of online learning with ASR hypotheses.

Our main contribution is to show that through online incremental learning ASR systems can adapt to users over time and show significant improvements in PPL along with reductions in WER. This improvement is consistent for online incremental learning on ASR hypotheses and references across three state-of-the-art baseline acoustic models.

The paper is organized as follows. In Section 2 we introduce relevant work for speaker-adaptive acoustic modelling and language modeling. We then describe the baseline language model and our online incremental learning methodology in Section 3. In Section 4, we discuss the experimental setup, results, and findings of online incremental learning applied to speech recognition systems. Finally, in Section 5 we conclude our work and discuss future ideas.

2. Related Work

Language model adaption has been widely studied in literature. Hsu et al. [1] explored the iterative use of the ASR hypotheses for unsupervised parameter estimation for n-gram language models. Similarly, [2, 3, 4] proposed unsupervised adaptation methods for presentation lecture speech recognition. Recent work on RNNLM adaptation [5, 6] explored using utterance topic information extracted with Latent Dirichlet Allocation and Hierarchical Dirichlet Processes in adapting language models for multi-genre broadcast transcription tasks. These works reported significant perplexity reduction by doing RNNLM adaptation, and small (0.1-0.2%) reduction on WER. Deena et al. [7] studied RNNLM adaptation with combined feature and model-based adaptation. Gangireddy et al. [8] explored model-based adaptation by scaling forward-propagated hidden activations and direct fine-tuning all RNNLM parameters. Ma et al.[9] proposed adapting the softmax layer of the neural network and showed improved performance on both perplexity and WER.

Another line of research that is closely related to our work is the cache language model [10, 11]. A cache language model stores a representation of the recent text and uses it for next word prediction. Jelinek et al. [12] proposed a cache trigram language model by using trigram frequencies estimated from the recent history of words. Grave et al. [13] proposed RNNLM with a continuous cache to adapt the word prediction to the recent history by storing past hidden activations as memory. They showed effective reductions of language model perplexity with the cache model and smaller reductions on WER.

10.21437/Interspeech.2018-2259

3. Models

We describe the acoustic models, our baseline language model, and proposed online incremental learning method.

3.1. Acoustic Models

Acoustic models capture the phonemes and various linguistic units of speech. In this work, we use state-of-the-art models and recipes available on Kaldi [14]. The three acoustic models we experiment with are Tri2, Tri3, and TDNN. Tri2 uses Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transforms (MLLT) [15] also known as Semi-Tied Covariance (STC) [16]. Tri3 is Speaker Adaptive Training (SAT) [17] on top of Tri2. The time delay nueral network (TDNN) [18] models long term temporal dependencies between acoustic events. The three baseline acoustic models along with their baseline WERs are summarized in Table 3, where TDNN shows the lowest WER, followed by Tri3, and Tri2.

3.2. RNN Language Model

Language models assign probabilities to a sequence of words. Let $\mathbf{w} = (w_0, w_1, ..., w_{T+1})$ be a sequence of words, where w_0 and w_{T+1} are the beginning-of-sentence token and end-of-sentence token. Using the chain rule, likelihood of the word sequence $\mathbf{w}$ can be factorized as:

$$P(\mathbf{w}) = \prod_{t=1}^{T+1} P(w_t|w_0, w_1, ..., w_{t-1}) \quad (1)$$

RNN-based language model [19], especially its variant using Long Short-Term Memory (LSTM) [20], has shown advantageous performance comparing to n-gram based models in applications such as speech recognition and machine translation. In this work, we use LSTM cell as the basic RNN unit for its stronger capability in capturing long term dependencies between words in the sequence. The LSTM cell state contains summarized information of previous observations, the propagation of which is regulated by cell gates. The cell output $h(t)$ is used to generate next word prediction probability distribution via a linear transformation:

$$P_{rnn}(w_t|w_{<t}) = P_{rnn}(w_t|h_t) = \text{softmax}(W_w h(t) + b_w) \quad (2)$$

where W_w and b_w are the weights and biases at the word output layer.

In ASR n-best list rescoring, we apply linear interpolation of the RNN based language model with the n-gram based language model for better generalization performance [21, 22]:

$$P(w_t|w_{<t}) = \lambda P_{ng}(w_t|w_{<t}) + (1-\lambda) P_{rnn}(w_t|w_{<t}) \quad (3)$$

where P_{ng} is the word probability from the n-gram model, and λ is the interpolation weight of the RNN language model.

3.3. Regularized Online Adaptation

During un-adapted base RNN language model training, we optimize model parameters to minimize cross-entropy of the predicted and true probability distributions at each time step.

$$\mathcal{D} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T+1} \log P(w_t^n|w_{<t}^n) \quad (4)$$

where N is the number of training samples in the mini-batch, and w_t^n denotes the word at index t of the nth training sample.

During online RNN language model adaptation, we apply conservative training [23] in order to prevent over fitting adaption data. We constrain model parameters to avoid high deviations from the pre-trained language model. Specifically, during model adaptation training, we fix the word embeddings learned from the offline model training, and apply additional regularization on the word output layer weights W_w to penalize large deviations from the un-adapted model parameter values. The adaptation optimization function becomes:

$$\widehat{\mathcal{D}} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T+1} \log P(w_t^n|w_{<t}^n) + \beta \left\| \widehat{W}_w - W_w \right\|^2 \quad (5)$$

where N is the number of training samples in the adaptation mini-batch, and $\widehat{W}_w$ denotes the new adapted word output layer weights. $\left\| \widehat{W}_w - W_w \right\|^2$ is the L2 norm. Instead of regularizing $\widehat{W}_w$ towards zero, we regularize it towards the un-adapted weights W_w so as to penalize large deviation of model parameters during adaptation.

3.4. Online Incremental Learning

Online incremental learning is the process of evaluating and improving an algorithm or model in real time from a continuous data stream. Here we leverage this concept and continuously train our ASR model on mini batches of the most recent ASR hypotheses or references. The baseline model is trained offline with parameters as described in Section 3.2. The continuous data stream refers to data that is intended only for inference by the baseline model. During online incremental learning, this continuous data stream is split into small segments. Once we have enough data for a segment and it has already been used for inference in real-time, we utilize this segment to mini batch train the online model before evaluating the next segment of data in real-time. The online model at time k_i is trained with the data in segment k_{i-1} or initially the baseline model, while it is used for inference on the data in the current segment at k_i. In other words, online model at time k_i is used to evaluate the data in segment k_{i+1} and then trained on it to create the online model at time k_{i+1}. The online model is retrained on one new segment of data at any given point in time and used to evaluate the next segment. In this paper, segments are units of multiple utterances. Refer to Equation 5 in Section 3.3 for the optimization function and training details for online adaptation.

4. Experiments

Experiments are conducted with the Kaldi speech recognition toolkit [14] and Tensorflow [24].

4.1. Data Preparation

We utilize the Large-Vocabulary Continuous Speech Recognition corpus Tedlium v2 [25] for our experiments. Tedlium v2 consists of 1495 transcribed Technology, Entertainment, Design (TED) talks, 207 hours in duration, given by 1242 unique speakers, with a total of 92,976 utterances and 2.6M words. TED talks are twenty minute influential talks that are highly focused on a specific topic or domain. To capture speaker articulation and speaker style, each dataset should contain unique speakers. The online model will then pick up specific features and enhance certain designated speaker traits in the neural network. Thus we split the speakers in the original Tedlium v2 training set at random into a new training set and test set.

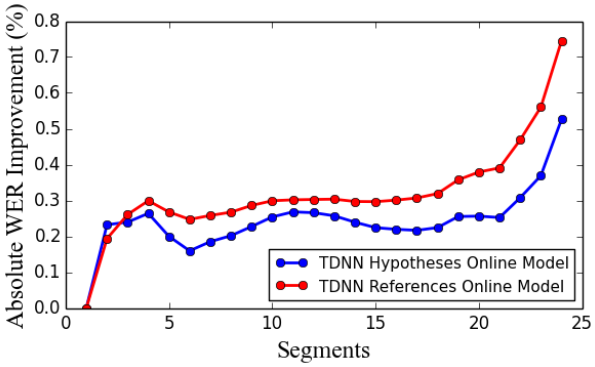

Figure 1: *The absolute Word Error Rate (WER) improvement (%) for TDNN + RNN LM ASR hypotheses and references online models per segment for the 10 test speakers.*

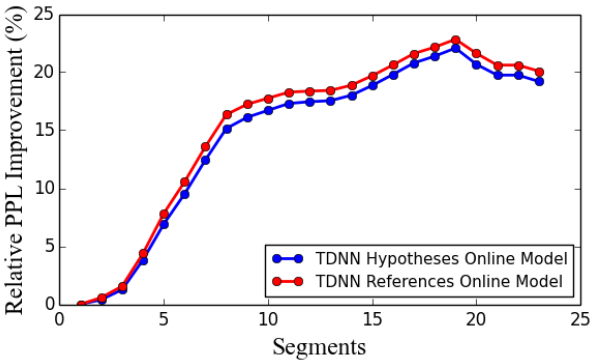

Figure 2: *The relative model perplexity (PPL) improvement (%) for TDNN + RNN LM ASR hypotheses and references online models per segment for the 10 test speakers.*

The test set is a non-overlapping subset of 10 unique speakers, each with a single talk from the original Tedlium v2 training set. The training set includes the remaining speakers of the original Tedlium v2 training set after extracting out the 10 speakers in the test set. We do not use a traditional validation set in our experiments since the online model is consistently trained with a small number of epochs to mimic real-time online learning. Table 1 contains a summary of the modified dataset.

Table 1: *Modified Tedlium v2 Data Set, where the Tedlium v2 training set is split into a new training set and test set.*

Characteristic	**Train**	**Test**
Number of unique speakers	1222	10
Number of unique talks	1475	10
Number of utterances	91709	936
Number of words	11786861	21020
Number of total segments	-	191

For each speaker in the test set, we split their utterances into **segments** - units of utterances. Instead of fixing the number of segments, we experiment with varying segment sizes, noting that this is dependent on the total number of utterances available for the given TED talk. Since we are evaluating the online model at each segment as well as at the end of the TED talk, we keep the segment size constant and allow for varying number of segments per speaker. The results presented in this paper utilize segment size of 5 utterances, with 11 to 24 segments per talk based on the available number of transcribed utterances.

4.2. Model Configurations

We simulate the online learning process by splitting 10 speaker talks in the test set into segments of 5 utterances each. Segments are preserved in order, as if the speaker were giving the talk live. This maximizes the information gained by speaker articulation and speaker style, and also provides indirect, partial topic and domain knowledge for future segments. The initial online model at segment 1 is the RNN baseline model described in Section 3.2 RNN Language Models. From segment 2 onwards, the online model is retrained on the most recent utterances after the evaluation of each segment.

We train a base RNN language model using Cantab-Tedlium text data. The training set has 14.5M sentences (251.8M words). The validation set (171.9K words) and a test set (174.9K words) each contains 10K sentences. The vocabulary is defined with the top frequent 20K words. We use LSTM cell as the basic RNN unit for its stronger capability in capturing long-range word dependencies. The LSTM state size is set as 1024. We conduct mini-batch training using Adam optimization method [26] with batch size of 128. Initial learning rate is set as 1e-3. Dropout regularization is applied to non-recurrent LSTM layers with a dropout keep probability of 0.8.

During online incremental learning, we train the model with the latest segment's utterances (either ASR hypotheses or references). We apply a small learning rate of 1e-4 and additional regularization on word output layer weights to prevent over fitting of adaptation data during online incremental learning. For model adaptation training on each segment, we use a batch size of 5 (the segment size) and train for 20 epochs. Note that in the last segment the batch size can be less than the segment size.

4.3. Results and Analysis

We evaluate our online incremental learning method on perplexity (PPL) and word-error-rate (WER), and analyze these findings across hypotheses, references, various acoustic models, segments, and speakers. Overall, our experimental results for 10 test speakers show a small positive correlation between model PPL and WER. Some strong improvements in the language model PPL are reflected as small improvements in WER. This is also consistent with the observations by Mikolov and Zweig in [5], where model PPL improvements correspond to at least small reductions in WER.

Online Learning with ASR Hypotheses vs. References To obtain relative improvement and explore the potential upper bound of online learning, we evaluate our online incremental learning method on both ASR hypotheses and reference utterances. While both online learning models show significant amount of improvement in model perplexity and WER, the online model trained on references shows slightly more improvement as expected. Across three different acoustic models, the online references models achieved an additional reduction of 0.03% to 0.12% in WER compared to the online hypotheses models. This suggests that we can leverage online learning in today's voice control systems that use an RNNLM to achieve reasonable improvements over time. With this, we simply collect the ASR hypotheses and use mini batch training on the model in real time. It would be slightly more difficult to utilize references, as corrections aren't always readily available. Both approaches are possible, on average, utilizing ASR hypotheses alone can achieve about half of the improvement of using references within a reasonable amount of segments.

Table 2: *The absolute perplexity (PPL) and Word Error Rate (WER) for TDNN + RNN LM baseline, ASR hypotheses and references online models for the 10 test speakers. Speaker segments range from from 11 to 24, corresponding to 55 to 120 utterances.*

Speaker		1	2	3	4	5	6	7	8	9	10
Baseline Model	PPL	222	265	230	250	342	248	225	245	141	269
	WER	6.20	7.31	7.65	14.87	7.57	10.58	7.16	8.89	10.85	9.00
Hypotheses Online Model	PPL	194	226	191	213	231	193	172	238	113	206
	WER	6.25	7.05	7.14	14.69	6.92	10.06	7.24	8.89	10.67	8.79
References Online Model	PPL	192	224	189	208	229	191	172	237	113	203
	WER	6.13	6.96	6.92	14.55	6.82	10.15	7.20	8.89	10.71	8.63

Online Learning with Different Acoustic Models For all three baseline acoustic models - LDA + MLLT, LDA + MLLT + SAT, and TDNN, we observe various degrees of improvement in language model perplexity and WER. This suggests that regardless of the acoustic model and the baseline WER, we will see small improvements in WER that correspond to significant improvements in perplexity reduction. These results and highlighted examples are similar to those that have been shown with traditional and neural cache models [27, 13]. With online incremental learning the RNN language model is prioritizing recent context, but unlike cache models, it is not purely count based as speaker semantics are also picked up.

Table 3: *Absolute Word Error Rate (WER) for all acoustic models paired with our RNN LM.*

Acoustic Model	**Baseline Model**	**Online Hypotheses**	**Online References**
LDA+MLLT	22.66	22.55	22.43
+SAT	18.56	18.52	18.49
TDNN	9.30	9.04	8.97

Online Learning across Segments We observe improved performance for both language model perplexity and WER over the sequence of segments during online incremental learning. Figure 2 shows the average relative model perplexity improvement across segments for 10 test speakers and Figure 1 shows the average absolute WER improvement for the online model across segments for all 10 test speakers. Both graphs exhibit an upward trend in relative improvement, which indicates that the online model is learning specific speaker feature traits. Within the test set, the average relative improvement has the steepest improvement within the beginning segments and later segments. This sharp growth is visible in both the relative improvement in model perplexity and WER. The continued improvement over segments shows promise in online incremental learning on both ASR hypotheses and references.

Online Learning across Speakers We evaluate online learning for a speaker over the speaker's whole talk and compare the final WER for all segments. We observe improved language model perplexity for all speakers and improved WER for eight out of ten speakers during online incremental learning. On average, the online model shows a 20.42% relative reduction in model perplexity and a 3.55% relative reduction in WER, for all speakers. For the eight speakers that showed WER improvement, the model achieves on average 4.06% relative reduction. Table 2 shows a summary of the relative model perplexity improvement and relative WER improvement for all 10 test speakers. We see consistent improvements in model perplexity for all speakers, which shows that the online model is picking up and effectively adapting to specific speaker feature traits. For all speakers, the relative model perplexity improved by a range of 3.24% to 30.2%. Figure 3 shows the absolute WER for the TDNN + RNN language model baseline and online models.

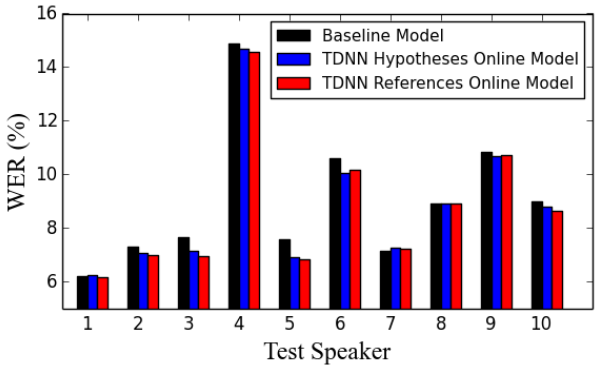

Figure 3: *The average Word Error Rate (WER) per speaker for TDNN + RNN LM baseline and ASR hypotheses and references online models per segment for the 10 test speakers.*

Example Phrases We highlight a few interesting phrases that surfaced during our online incremental learning experiments. In Ben Kacyra's talk on the preservation of cultural heritage sites in "Ancient Wonders Captured in 3D", the baseline model confuses "sight(s)" and "site(s)" through out the talk while the online model did not.

5. Conclusion

We explored language model adaptation for ASR systems through online incremental learning in which a speaker's most recent utterances are used in mini batches for online training of a RNN language model. During evaluation we studied the effects of using both ASR hypotheses and reference utterances across multiple baseline models. Comparing to a state-of-the-art baseline on Tedlium v2, our model trained on ASR hypotheses showed an average of 19.18% reduction in model perplexity and 2.80% relative reduction in word error rate. We also discovered a potential upper bound improvement of an additional 30.5% by training on references.

Through metric improvements as well as highlighted examples in the speaker's text, the online models are able to better capture context and speaker specific traits. The results in this paper confirm the benefit and demonstrate the potential of using online incremental learning for language model speaker adaptation in ASR systems. As next steps, we can explore the rate and efficiency of online learning over longer periods of time and for larger sampling of speakers.

6. References

[1] B.-J. Hsu and J. Glass, "Language model parameter estimation using user transcriptions," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4805–4808.

[2] T. Niesler and D. Willett, "Unsupervised language model adaptation for lecture speech transcription," *Training*, vol. 144, no. 38h, p. 413K, 2002.

[3] H. Nanjo and T. Kawahara, "Unsupervised language model adaptation for lecture speech recognition," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[4] v. H. Nanjo and T. Kawahara, "Language model and speaking rate adaptation for spontaneous presentation speech recognition," *IEEE Transactions on speech and Audio Processing*, vol. 12, no. 4, pp. 391–400, 2004.

[5] T. Mikolov and G. Zweig, "Context dependent recurrent neural network language model." *SLT*, vol. 12, pp. 234–239, 2012.

[6] X. Chen, T. Tan, X. Liu, P. Lanchantin, M. Wan, M. J. Gales, and P. C. Woodland, "Recurrent neural network language model adaptation for multi-genre broadcast speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[7] S. Deena, M. Hasan, M. Doulaty, O. Saz, and T. Hain, "Combining feature and model-based adaptation of rnnlms for multi-genre broadcast speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Sheffield, 2016, pp. 2343–2347.

[8] S. R. Gangireddy, P. Swietojanski, P. Bell, and S. Renals, "Unsupervised adaptation of recurrent neural network language models." in *Interspeech*, 2016, pp. 2333–2337.

[9] M. Ma, M. Nirschl, F. Biadsy, and S. Kumar, "Approaches for neural-network language model adaptation," *Proc. Interspeech 2017*, pp. 259–263, 2017.

[10] R. Kuhn, "Speech recognition and the frequency of recently used words: A modified markov model for natural language," in *Proceedings of the 12th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 1988, pp. 348–350.

[11] R. Kuhn and R. De Mori, "A cache-based natural language model for speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 12, no. 6, pp. 570–583, 1990.

[12] F. Jelinek, B. Merialdo, S. Roukos, and M. Strauss, "A dynamic language model for speech recognition," in *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*, 1991.

[13] E. Grave, A. Joulin, and N. Usunier, "Improving neural language models with a continuous cache," *arXiv preprint arXiv:1612.04426*, 2016.

[14] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[15] R. A. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2. IEEE, 1998, pp. 661–664.

[16] M. J. Gales, "Semi-tied covariance matrices for hidden markov models," *IEEE transactions on speech and audio processing*, vol. 7, no. 3, pp. 272–281, 1999.

[17] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 2. IEEE, 1996, pp. 1137–1140.

[18] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts." in *INTERSPEECH*, 2015, pp. 3214–3218.

[19] T. Mikolov, S. Kombrink, L. Burget, J. H. Černockỳ, and S. Khudanpur, "Extensions of recurrent neural network language model," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5528–5531.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] H. Schwenk, "Continuous space language models," *Computer Speech & Language*, vol. 21, no. 3, pp. 492–518, 2007.

[22] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model." in *Interspeech*, vol. 2, 2010, p. 3.

[23] D. Albesano, R. Gemello, P. Laface, F. Mana, and S. Scanzio, "Adaptation of artificial neural networks avoiding catastrophic forgetting," in *Neural Networks, 2006. IJCNN'06. International Joint Conference on*. IEEE, 2006, pp. 1554–1561.

[24] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[25] A. Rousseau, P. Deléglise, and Y. Estève, "Enhancing the ted-lium corpus with selected data for language modeling and more ted talks." in *LREC*, 2014, pp. 3935–3939.

[26] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[27] S. Della Pietra, V. Della Pietra, R. L. Mercer, and S. Roukos, "Adaptive language modeling using minimum discriminant estimation," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 103–106.