INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Unbiased semi-supervised LF-MMI training using dropout

Sibo Tong[1,2], *Apoorv Vyas*[1,2], *Philip N. Garner*[1], *Hervé Bourlard*[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{sibo.tong, apoorv.vyas, phil.garner, bourland}@idiap.ch

Abstract

The lattice-free MMI objective (LF-MMI) with finite-state transducer (FST) supervision lattice has been used in semi-supervised training of state-of-the-art neural network acoustic models for automatic speech recognition (ASR). However, the FST based supervision lattice does not sample from the posterior predictive distribution of word-sequences but only contains the decoding hypotheses corresponding to the Maximum Likelihood estimate of weights, so that the training might be biased towards incorrect hypotheses in the supervision lattice even if the best path is perfectly correct. In this paper, we propose a novel framework which uses Dropout at the test time to sample from the posterior predictive distribution of word-sequences to produce unbiased supervision lattices for semi-supervised training. We investigate the dropout sampling from both the acoustic model and the language model to generate supervision. Results on Fisher English show that the proposed approach achieves WER recovery of $\sim 51.6\%$ over regular semi-supervised LF-MMI training.

Index Terms: Automatic Speech Recognition, Semi-Supervised learning, Dropout, LF-MMI

1. Introduction

The current acoustic models for Automatic Speech Recognition (ASR) are based on Deep neural networks (DNN). Sequence level training criteria such as Connectionist Temporal Classification (CTC) [1], Lattice-free Maximum Mutual Information (LF-MMI) [2] and state-level Minimum Bayes Risk (sMBR) [3, 4] are preferred over frame-level objectives as they exploit sequential information. However, these methods are known to be data hungry.

Although it is difficult and costly to obtain large amount of supervised data, abundant unsupervised audio is often easily available. A typical approach to exploit unsupervised data is to train a seed model using supervised data and use the seed model to automatically transcribe the unsupervised data [5, 6]. Of course, the automatic transcripts are not perfect and the unsupervised training data is usually selected based on confidence measure on frame level [7], word level [6, 8, 9] or utterance level [10, 11, 12].

More recently, lattice-based supervision has been combined with lattice-free MMI objective for semi-supervised training [13]. Instead of using only the best path as supervision, training with the whole decoding lattice for the unsupervised data allows the model to learn from alternative hypotheses when the best path is not accurate. Although it has shown significant improvement, directly learning from the whole lattice can deteriorate the performance in cases where the best path hypothesis has much lower word error rate (WER) than the alternate hypotheses. This is because all the paths in the supervision lattice are considered equally likely and thus the training can be biased towards incorrect hypotheses.

To this end, we propose to use a novel approach to sample alternate hypothesis from the approximate posterior-predictive distribution instead of using decoding lattice which contains the most competitive hypothesis for the Maximum Likelihood estimate of the weights. Given an already trained neural network (NN) based acoustic model, we use dropout during inference to compute the frame-level state posterior probabilities. We then use these frame level posterior probabilities to generate a decoding hypothesis for the test utterance. We repeat this process N times for the same utterance with different random selection of active neurons to generate N decoded hypotheses. As shown in [2], this process leads to a Bayesian inference over the acoustic model weights and thus approximates sampling from the posterior predictive distribution over word sequences. As shown in [14], when the acoustic model is uncertain at a certain word, we observe variations in the predicted Monte Carlo hypotheses; we see the same hypothesis sampled when the model is confident. Given this observation, we can combine all the N hypotheses to form an unbiased supervision lattice for the corresponding unlabeled utterance. Similarly, this dropout-based sampling can be applied to language model as well by re-scoring the same decoding lattice multiple times using a NN-based language model while keeping dropout on.

The proposed approach has similarities to Negative Conditional Entropy (NCE) [15] for semi-supervised training where the authors minimize the expected risk over the uncertain decoding of the unsupervised data. However, in contrast to [15], where the decoding lattice with forward-backward likelihood computation estimates the likelihood of word-sequence, in this work, we directly sample from the approximate posterior-predictive distribution using dropout to generate the supervision lattice. The approach proposed in [16] also shares some similarities in the sense that the labels of unlabeled data are the decoding output from multiple seed models to incorporate the diversity. An ensemble of models is trained in parallel using these diverse labels, and then averaged as the final model. In the context of our framework, these multiple seed models can be considered as the dropout-based neural network samples and all the diverse labels are combined into one supervision lattice used for LF-MMI training. Thus, the proposed method is simpler and more rigorous.

The remainder of this paper is organized as follows: in Section 2, we introduce the regular semi-supervised training for LF-MMI and discuss the proposed framework in more detail. Experimental results and analysis are provided in Section 3. Finally, Section 4 concludes the paper.

2. Proposed Method

In this section, we explain our approach to maximize the expected LF-MMI objective for unlabelled data by sampling target

http://dx.doi.org/10.21437/Interspeech.2019-2678

word-sequences from the posterior-predictive distribution for a given utterance. Our proposed loss for semi-supervised training is given as follows :

$$\mathcal{F}_{\text{MMI}} = \max_{\theta} \sum_{u=1}^{U} \log \left(\mathop{\mathbb{E}}_{\mathbf{W} \sim P(W|O^u, \mathbf{D}_s)} P(W|\mathbf{O}^{(u)}, \theta) \right) \quad (1)$$

where $\mathbf{O}^{(u)}$ is the sequence of acoustic observations for utterance u, $\mathbf{D}_s$ is the supervised training data. W is the sampled target word sequence for the utterance. In this work, we use dropout to decode the same utterance multiple times to perform approximate Bayesian inference over the model parameters. This allows to sample from the approximate posterior-predictive distribution $P(W|O^{(u)}, \mathbf{D}_s)$. In contrast to this, the regular semi-supervised LF-MMI objective proposed in [13] is given by:

$$\begin{aligned} \mathcal{F}_{\text{MMI}} &= \max_{\theta} \sum_{u=1}^{U} \log \left(\sum_{W \in \mathcal{G}_{\text{num}}^{(u)}} P(W|\mathbf{O}^{(u)}, \theta) \right) \\ &\approx \max_{\theta} \sum_{u=1}^{U} \log \left(\mathop{\mathbb{E}}_{\mathbf{W} \sim \mathcal{G}_{\text{num}}^{(u)}} P(W|\mathbf{O}^{(u)}, \theta) \right) \end{aligned}$$

where $\mathcal{G}_{\text{num}}^{(u)}$ is the decoding lattice for the utterance u.

This can be seen an approximation to the proposed loss (1) where the expectation is taken over the word-sequences in the decoding lattice and each output word sequence in the lattice is assumed to be equally likely. Using the whole lattice as supervision provides additional information especially when the seed network is not confident on the unsupervised data. However, these alternative paths can also spoil the supervision in some cases. For instance, when the best decoding is quite accurate or when the utterance is short (containing only 1 or 2 words), the supervision might be biased towards the incorrect paths. One decoding lattice example from the unlabeled data is shown in Fig. 2(a). The example utterance is quite clean. Although the model is quite confident on the sentence, the decoding lattice still contains many incorrect paths which will deteriorate the supervision quality.

Therefore, in this paper, we propose to employ dropout at test time and decode the unlabeled data multiple times. Sampling from the posterior-predictive distribution will lead to an unbiased estimate to (1). We investigate Dropout-based sampling for both the acoustic and the language model.

2.1. Dropout-based sampling

While dropout is typically used during training to prevent overfitting of DNNs, it was recently shown in [17] that dropout during inference can lead to Bayesian inference over the model parameters and thus provide a way to sample from posterior-predictive distribution as well as to compute the models uncertainty on its predictions. The present work is a novel attempt to study the usage and utility of dropout uncertainty in the context of semi-supervised training for ASR systems.

2.1.1. Dropout Sampling from Acoustic Model

Given an already trained DNN-based acoustic model, for each utterance, we forward-pass it N times through a dropout enabled neural network acoustic model. Each of the N acoustic model outputs is then processed though the decoding pipeline to generate N dropout-lattices. As shown in our previous work [14], the acoustic model uncertainty about a test utterance is reflected in the variations observed in the predicted hypotheses for each Monte Carlo sample. Moreover, the variations in different decoded hypotheses for any utterance are often highly localized at certain word positions and depict locations where the ASR decoding might be inaccurate.

Therefore, we can generate an unbiased supervision lattice for each unlabeled utterance by composing the predicted hypotheses from the Monte Carlo samples. More specifically, as shown in Fig. 1, for each unlabeled utterance, we prune the dropout-lattices with a very small beam and combine them together to create the supervision lattice for semi-supervised training. Optimizing $P(W|\mathbf{O}^{(u)}, \theta)$ over this lattice leads to an unbiased estimate of (1). We keep the rest of the training steps the same as proposed in [13].

Fig. 2(b) shows the lattice for the same example utterance, generated using the proposed approach. As we can find, most of the paths in this lattice correspond to the correct transcription since the model is confident on this clearly spoken utterance (high $P(W|\mathbf{O}^{(u)}, \theta_s)$ for the decoded sequence). If the model is uncertain about an utterance, more variations will appear in each decoding sample [14] so that the combined lattice can still retain alternative paths to provide additional information. We hypothesize that the unbiased lattice combined from different dropout-based decoding samples better reflects the uncertainty of the acoustic model and is able to foster the more likely word sequences, while keeping variations for uncertain utterances, thus improving the semi-supervised training performance.

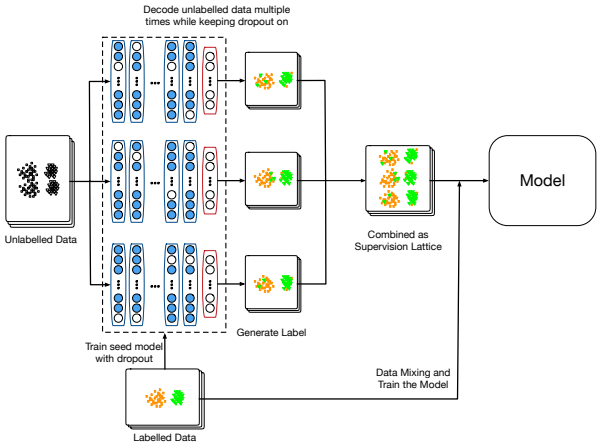
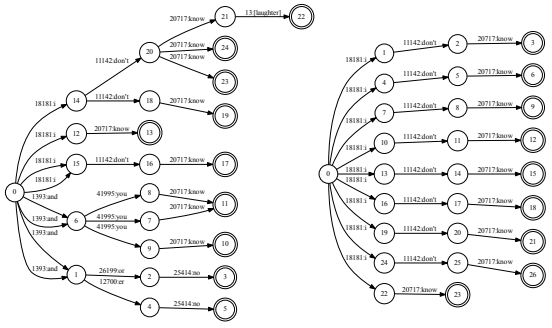


Figure 1: *Flow-chart of the proposed method. Each network in the figure represents one network sample because of a different random selection of the active nodes. The white nodes denote that they are dropped out.*

2.1.2. Dropout Sampling from Language Model

It is not straight forward to apply the dropout-based sampling in N-gram language model (LM) that is used in decoding. Instead, we investigated the same framework for neural network-based language models during re-scoring. For each unlabeled utterance, we first obtain the decoding lattice using the acoustic model with dropout off. The lattice is then re-scored N times by using a dropout enabled neural network language model. These re-scored lattices are then pruned and combined together to generate the supervision lattice which reflects uncertainties in the

language model. Similarly, we keep everything else the same in the semi-supervised training setup and evaluate the performance. Additionally, we hypothesize that the combination of the dropout sampling from acoustic model and the sampling from language model could help further because it covers the uncertainties from each of the two major components of an ASR system. This combination will also be investigated in Section 3.



(a) (b)

Figure 2: *Lattices of a clearly spoken utterance. (a) represents the pruned decoding lattice from a dropout-off acoustic model. (b) denotes the unbiased lattice combined from multiple dropout decoding samples.*

3. Experiments

3.1. Experimental Setup

Similar to [13], we report our results on the Fisher English corpus [18]. A randomly chosen subset of speakers (250 hours) from the corpus is used as unsupervised data. The transcripts from the remaining 1250 hours are used to train the language models for decoding and re-scoring the unsupervised data. We use a 50 hours subset from the corpus as the supervised data to train the seed model. The results are reported on separately held-out development and test sets (about 5 hours each), which are part of the standard Kaldi [19] recipe for Fisher English. WER Recovery Rate (WRR) [20] is used as an additional metric to evaluate the WER improvements from semi-supervised training:

$$\text{WRR} = \frac{\text{BaselineWER} - \text{SemisupWER}}{\text{BaselineWER} - \text{OracleWER}}.$$

Following the standard Kaldi recipe, we first train a GMM system using only the supervised data and use this to get supervision to train a seed LF-MMI time-delay neural network (TDNN) [21] system. The TDNN consists of 8 hidden layers, with 450 hidden units in each layer. Dropout is applied on top of each layer. We use i-vector [22] for speaker adaptation of the neural network. The i-vector extractor is trained using the combined supervised and unsupervised datasets. Also, for comparison purposes, we use statistics from only the supervised data to train the context-dependency decision tree. Following [13], the phone LM used for creating the denominator FST is estimated using phone sequences from both supervised and unsupervised data with a higher weight to the phone sequences from supervised data (1.5 for the 50 hours supervised dataset and 1 for the unsupervised data).

In addition to N-gram language models, A neural network-based language model is trained on the same data. The network consists of 3 temporal convolutional layers [23], with 600 units in each layer. The size of the word embeddings is fixed to 600 and the kernel size is taken to be 3. Similarly, dropout is applied on top of each layer. The language model was trained using Pytorch.

3.2. Results

3.2.1. Effect of Dropout Sample Numbers from Acoustic Model

As a hyper-parameter, N denotes the number of dropout samples needed to represent the posterior-predictive distribution. Although more posterior samples can better represent the distribution, it is more time consuming. Therefore, it is of importance to investigate appropriate value of N for a good trade-off. Here, we have only applied the dropout-based sampling on the acoustic model. To generate the supervision lattice of the unsupervised data, we decoded the data N time while keeping dropout on and varied N from 5 to 40. As a baseline, we use the decoding lattice generated from the same acoustic model in a standard way (with dropout off), following [13]. The decoding lattices were not re-scored and the performance was evaluated on development set. As shown in Fig.3, the performance of the

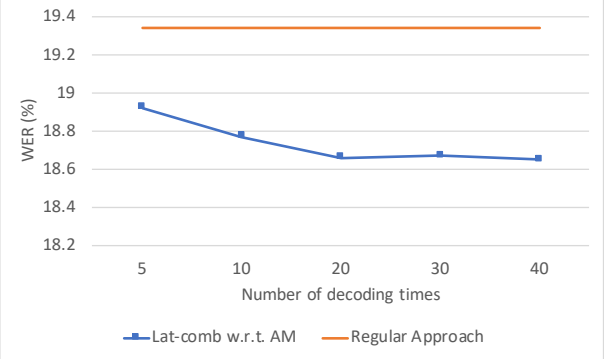


Figure 3: *WER (%) of different semi-supervised training setup by varying the value of N. The dropout-based sampling is only applied on the acoustic model. The red line denotes the regular semi-supervised training approach [13].*

proposed method first gets improved as we combined more decoding lattices. It seems to saturate after reaching 20 times of decoding. Therefore, we keep using $N = 20$ for the following experiments except when explicitly stated.

3.2.2. Quality Analysis of the Supervision Lattices

From Fig.3, we can also see that the unbiased lattices yield better word error rate (WER) than the regular semi-supervised training approach. We analyzed the averaged WER and the sentence error rate (SER) of the unbiased lattices with $N = 20$ and compared it with the regular decoding lattice on the whole unsupervised data. We evaluated the WER of each lattice by averaging the WER of the N-best hypotheses for each utterance. The regular decoding lattice was generated from the dropout-off model and was pruned before this evaluation.

Table 1 shows that the unbiased lattice has a better WER and a much better SER than the regular lattice. The better WER and SER confirms our hypothesis that the lattice combination from different dropout samples can help reduce the effect of

Table 1: *Comparison the averaged WER(%) and SER (%) between combined lattice and regular decoding lattice.*

	avg. WER	SER
Regular Lat	23.6	87.8
Lat-comb	23.1	75.7

incorrect hypotheses in the supervision lattice when the acoustic model is confident on the unlabeled sentence, while keeping alternative paths to be exploited when the acoustic model is uncertain. It also explains the improvement on development set after semi-supervised training because the unbiased lattice provides supervision with better quality.

3.2.3. Effect of Number of Dropout Samples from Language Model

Similar to Section 3.2.1, in this section, we analyze the effect of N with respect to language model only. To generate the unbiased supervision with respect to language model, we first obtained the lattice by decoding the data in regular way (keeping dropout off). The lattice was then re-scored N times by the network-based language model while keeping dropout on. Similarly, we varied N from 5 to 40.

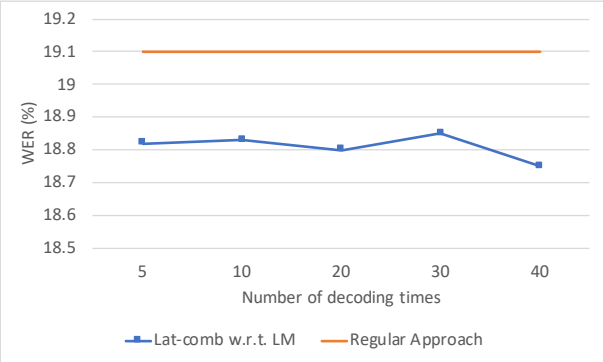


Figure 4: *WER (%) of different semi-supervised training setup by varying the value of N. The dropout-based sampling is only applied on the language model. The red line denotes the regular semi-supervised training approach [13] where the supervision lattices of unsupervised data were also re-scored using NN LM.*

As shown in Fig. 4, the performance on the development set does not change much with different values of N and the proposed approach yields very slight improvement. One of our previous hypotheses was that the Dropout-based Monte Carlo sampling can help reduce the confusion in the supervision lattice especially for shorter sentences. However, language model re-scoring for sentences with one or two words wouldn't make much difference by its nature. We found there are around one third of the unsupervised utterances containing only 3 words or less. Therefore, applying dropout sampling on language model only slightly improves the performance.

3.2.4. Complete Comparison

Table 2 shows a complete comparison of the alternatives we are exploring. The first row shows the performance of supervised training using only 50 hours supervised data. The last row shows supervised training results using oracle transcripts for the unsupervised data. All the supervision lattices for unlabeled data were re-scored using the network language model. For re-scoring the unbiased acoustic lattice in the proposed framework, we first generated the decoding lattice samples by keeping dropout on in the acoustic model. Then, each decoding lattice was re-scored before pruning and combination. In order to testify whether the dropout sampling from both acoustic model and language model can further improve the performance, we simply combined the lattice evaluated in Section 3.2.1 and Section 3.2.3 and tested the WER after semi-supervised training.

Table 2: *Comparison between combined lattice and regular decoding lattice in WER(%). The 50h supervised system is used as baseline to calculate WRR.*

System	Dev	Test	WRR
50h supervised	21.0	20.9	-
Regular Approach	19.1	19.2	53.7 %
Lat-comb w.r.t. AM	18.5	18.3	76.1%
Lat-comb w.r.t. LM	18.8	18.7	65.7%
Lat-comb w.r.t. AM+LM	**18.5**	**18.2**	77.6%
Oracle	17.7	17.5	

As we can see in the table 2, semi-supervised training approach as proposed in [13] yields around 8.6% relative WER reduction. Incorporated with uncertainty information from only the acoustic model, the unbiased supervision lattice improves over the supervised system by around 12.2%. Dropout sampling from network language model also brings improvement, although the improvement is not as much as the one from acoustic model. The combination cannot further improve the performance significantly. Most of the gains come from the acoustic part. In total, the proposed semi-supervised training approach yields approximately 12.4% relative improvement over the supervised setup. Compared with the regular LF-MMI semi-supervised training, the proposed approach gives 4.2% relative WER reduction and 51.6% WER recovery rate.

4. Conclusion

We have proposed a novel way to exploit dropout uncertainty in context of semi-supervised LF-MMI training. It was demonstrated that the unbiased lattice combined from different dropout-based decoding samples is able to help reduce the confusion of the lattice paths, while keeping variations for uncertain unlabeled utterances. Experiments on the Fisher English shows that the proposed approach can further improve the WER over the regular semi-supervised training framework. While this paper primarily focused on LF-MMI training, it is clear that the idea can be further extended to other frameworks such as end-to-end based semi-supervised training. In the future, we also intend to apply the idea of NCE in LF-MMI training to reduce the time required for multiple times of decoding.

5. Acknowledgements

The research leading to these results has received funding from the European Community H2020 Research and Innovation Actionfunding, under "Scalable Understanding of Multilingual Media" (SUMMA) project No. 688139, and the Swiss National Science Foundation project SHISSM (Sparse and hierarchical Structures for Speech Modeling), grant agreement 200021-175589.

6. References

[1] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *international conference on machine learning*, 2006.

[2] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI." in *Proceedings of Interspeech*, 2016.

[3] J. Kaiser, B. Horvat, and Z. Kacic, "A novel loss function for the overall risk criterion based discriminative training of HMM models," in *Proceedings of the International Conference on Spoken Language Processing*, 2000.

[4] N. Kanda, Y. Fujita, and K. Nagamatsu, "Lattice-free state-level minimum Bayes risk training of acoustic models," in *Proceedings of Interspeech*, 2018.

[5] G. Zavaliagkos, M.-H. Siu, T. Colthurst, and J. Billa, "Using untranscribed training data to improve performance," in *Proceedings of the International Conference on Spoken Language Processing*, 1998.

[6] F. Wessel and H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, 2005.

[7] K. Veselỳ, M. Hannemann, and L. Burget, "Semi-supervised training of deep neural networks," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013.

[8] S. Thomas, M. L. Seltzer, K. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[9] K. Veselỳ, L. Burget, and J. Cernockỳ, "Semi-supervised DNN training with word selection for ASR." in *Proceedings of Interspeech*, 2017.

[10] S. Novotney, R. Schwartz, and J. Ma, "Unsupervised acoustic and language model training with small amounts of labelled data," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009.

[11] F. Grezl and M. Karafiát, "Semi-supervised bootstrapping approach for neural network feature extractor training," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013.

[12] P. Zhang, Y. Liu, and T. Hain, "Semi-supervised DNN training in meeting recognition," in *Proceedings of the IEEE Workshop on Spoken Language Technology*, 2014.

[13] V. Manohar, H. Hadian, D. Povey, and S. Khudanpur, "Semi-supervised training of acoustic models using lattice-free MMI," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[14] A. Vyas, P. Dighe, S. Tong, and H. Bourlard, "Analyzing uncertainties in speech recognition using dropout," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

[15] V. Manohar, D. Povey, and S. Khudanpur, "Semi-supervised maximum mutual information training of deep neural network acoustic models," in *Proceedings of Interspeech*, 2015.

[16] S. Li, X. Lu, S. Sakai, M. Mimura, and T. Kawahara, "Semi-supervised ensemble DNN acoustic model training," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

[17] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016.

[18] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: a resource for the next generations of speech-to-text." in *LREC*, 2004.

[19] 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 *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[20] J. Ma and R. Schwartz, "Unsupervised versus supervised training of acoustic models," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[21] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *Backpropagation: Theory, Architectures and Applications*, 1995.

[22] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, 2011.

[23] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.