Interspeech 2018
2-6 September 2018, Hyderabad

Active Learning for LF-MMI Trained Neural Networks in ASR

Yanhua Long[1], *Hong Ye*[1], *Yijie Li*[2], *Jiaen Liang*[2]

[1] SHNU-Unisound Joint Laboratory of Natural Human-Computer Interaction, Shanghai Normal University, Shanghai, China
[2] Beijing Unisound Information Technology Co., Ltd. Beijing, China
{yanhua, yeeho}@shnu.edu.cn, {liyijie, liangjiaen}@unisound.com

Abstract

This paper investigates how active learning (AL) effects the training of neural network acoustic models based on Lattice-free Maximum Mutual Information (LF-MMI) in automatic speech recognition (ASR). To fully exploit the most informative examples from fresh datasets, different data selection criterions based on the heterogeneous neural networks were studied. In particular, we examined the relationship among the transcription cost of human labeling, example informativeness and data selection criterions for active learning. As a comparison, we tried both semi-supervised training (SST) and active learning to improve the acoustic models. Experiments were performed for both the small-scale and large-scale ASR systems. Experimental results suggested that, our AL scheme can benefit much more from the fresh data than the SST in reducing the word error rate (WER). The AL yields $6\sim13\%$ relative WER reduction against the baseline trained on a 4000 hours transcribed dataset, by only selecting 1.2K hrs informative utterances for human labeling via active learning.

Index Terms: active learning, acoustic modeling, heterogeneous neural network, speech recognition

1. Introduction

In recent years, the success of acoustic modeling using deep neural networks (DNNs) in ASR is great. The performances of ASR systems have been greatly improved. This significant progress has accelerated the successful applications of ASR techniques in many industrial services [1, 2, 3]. At the same time, it also brings new challenges to ASR, because for different applications, the data property of users' speech utterances may deviate far from the given acoustic models (AMs), either in the acoustic environments or in the linguistic conditions, etc. Therefore, in order to improve the performance for each industrial ASR service, it is necessary to update the acoustic model (AM) constantly using fresh speech data collected from the latest production traffic.

There are two main acoustic modeling techniques to use fresh data in the ASR community, the semi-supervised training (SST) [4, 5, 6] and the active learning (AL) [7, 8, 9, 10, 11, 12]. The advantage of SST is the ability of creating automatic transcriptions for a large amount of untranscribed data. However, these transcriptions may still contain many errors, these errors are very sensitive to the DNN-based acoustic modeling techniques using discriminative training criterions [5]. Moreover, the selected fresh data with high confidence tends to have similar acoustic properties with the ones for seed AM training. Because all the automatic transcriptions are highly depend on the seed ASR systems used for decoding. Thus, the SST techniques are normally useful at the beginning of untranscribed data labeling, its gains are rapidly degraded, when the selected homogeneous data reaches to some extent, especially for updating the AMs in some low-resource and large-scale ASR tasks.

Different from SST, the advantage of AL is that, it can guarantee the transcription quality of untranscribed data through human labeling. And we can label any type of speech data with diverse acoustic properties. Therefore, it has been receiving much attentions in the recent ASR literature [9, 10, 11, 12]. However, creating high quality transcription of speech data manually is a time-consuming and costly process, it becomes impossible for a very large amount of data. Therefore, the key challenge of AL is to select the most informative examples from untranscribed fresh data for human labeling with minimum manual labeling costs. Many studies in AL have been focused on this issue. They are mainly differ in the criterion used for informative data selection. Such as, the utterance selection using the traditional confidence measures (CMs) and their variants [7, 11, 13], the min-max framework [14], the delta-AUC selection approach [10], and the HMM-state or N-best entropy [9, 15], etc.

In this paper, we focus on a new method to select the informative utterances for AL based on the heterogeneous neural networks (HNN). In this method, the confidence measures and a word matching error rate (WMER) are combined together to form the criterion of data selection. Our idea is inspired by [16], which is a recent proposed approach used for semi-supervised training. Effectiveness of the proposed AL will be validated for the LF-MMI criterion [17] based deep acoustic model training. In our experiments, different data selection methods for AL to minimize the human efforts are investigated, and the relationship among the manual transcription cost, example informativeness and data selection criterion is also studied particularly. Furthermore, we compare the AL with SST approach, using the unified HNN framework, to examine their effectiveness for improving acoustic models. Experimental results indicate that, the proposed AL scheme can benefit more from the fresh data than the recent proposed SST, and it still can obtain $6\sim13\%$ relative WER reduction by only required 1.2K hours (hrs) selected informative utterances for human labeling, even the baseline was already trained on a 4000 hrs large-scale transcribed dataset.

2. HNN-based Active Learning

The scheme of HNN-based AL approach is illustrated in Figure 1. Our target is to improve the deep acoustic model of LTDNN (Interleaving Time-Delay Neural Network (TDNNs) and unidirectional Long Short-term Memory LSTMs), which has been proposed in [18] and implemented in Kaldi speech recognition toolkit [19]. This architecture has been shown to not only outperforms the state-of-the-art low frame rate (LFR) BLSTM models, but also computationally more efficient.

As illustrated in Figure 1, given a large amount of un-

10.21437/Interspeech.2018-1162

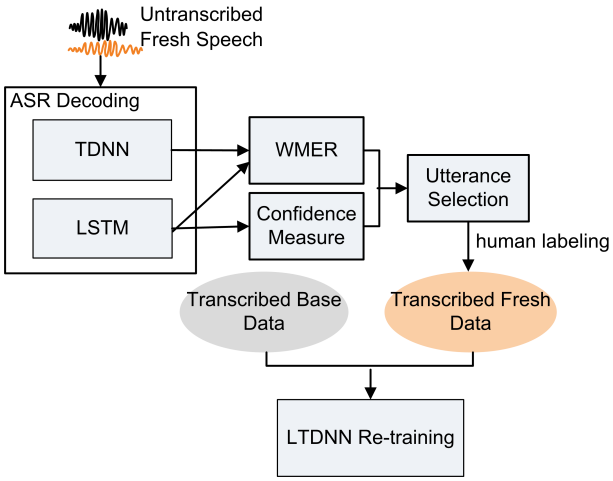

Figure 1: *Schematic diagram of HNN-based Active Learning.*

transcribed fresh speech utterances (UTFD), we first recognize them using two ASR systems with HNN acoustic models of LSTM and TDNN. Then the WMER of each utterance is calculated, by comparing its two decoding hypotheses. The way to compute WMER is the same as WER, but taking hypotheses from LSTM as reference instead of the ground truth. Meanwhile, the confidence measure of each utterance is computed. As observed in [11], using sophisticated confidence measures did not help to yield any data selection gain, so we choose to use only the lattices from LSTM ASR system to compute the CMs. Both CM and WMER are combined to form the criterion of fresh utterance selection for human labeling, such as, we may use a criterion of "WMER=[30,40] $\oplus$ CM=[0, 0.7]". It means that, those utterances with WMERs lie in 30% to 40% range and CM lie in 0 to 0.7 range will be selected as informative examples for human labeling. Finally, the manually transcribed fresh data (TFD) and the transcribed base data (TBD) are jointly used for re-training of the target LTDNN model.

Although the proposed scheme for AL seems very simple and easy to implement, it proved to be very effective in our experiments. In our scheme, two points are different from previous works: 1) We use two HNNs to train the decoding ASR systems (seed systems) on the whole TBD, instead of using the same type AMs as target LTDNN model. Such as, the committee-based works using the same type AMs as target one to train multiple ASR systems on subsets of TBD dataset in [20]. Our design is motivated by [16], and its advantage is, it can alleviate the data homogeneity between the TFD and TBD, and enhance the generalization ability of the improved target AM. Because the use of heterogeneous models could bring large diversity to produce informative and complementary examples. Furthermore, using the whole TBD can build better HNN models to reduce the degree of disagreement of those fresh utterances with similar acoustic properties to TBD. 2) We use the combination of WMER and CM to produce a better measurement of disagreement degree between the recognition results, because we know that, estimating accurate CM is also very challenging in ASR literature. Therefore, we expect that, it could outperform the conventional AL data selection techniques which depend on only the CM from a single target AM.

3. Experiments

3.1. Datasets

Our experiments are designed to simulate the possible gains obtained by the updated AM, using the proposed AL to achieve the TFD. For this simulation, we had about 4000 hrs TBD with Mandarin speech, which was used to train the large-scale ASR baseline and the seed LSTM and TDNN models for active learning. In addition, we randomly selected two different 600 hrs datasets from 4000 hrs. The first 600 hrs was used to build the small-scale ASR baseline AMs. The second 600 hrs was used as additional transcribed base data (TBD-add) in experiments in section 3.3.4.

About 30K hrs untranscribed fresh speech utterances are collected from the ASR voice search engine of Unisound Corporation (https://www.unisound.com/). These utterances are considered as UTFD. Three test sets are used to do the system evaluation, they are about 3 hrs Point-of-Interest speech (POI), 2.5 hrs speech about general voice search topic (GTopic) and 2 hrs Children's speech (Child).

3.2. Model description

The structure of LSTM acoustic model used here is the fast deep projected LSTM (LSTMP), recurrent neural network (RNN). It has 5 LSTM hidden layers, where each has 1024 memory cells, and the cell outputs were fed into the 256-unit projection layers. The output label delay was set to 5. The TDNN model is a 7 layers sub-sampled TDNN with splicing indexes are set to -2,-1,0,1,2 -1,0,1 -1,0,1 -3,0,3 -3,0,3 -3,0,3 -3,0,3 and the output ReLU dimensions of the weight matrices is set to 1024. Our target LTDNN model is a mixture architecture of LSTMPs and sub-sampled TDNNs, using 3 fast-LSTMP layers interleaved with 7 spliced TDNN layers. For the detail neural network configurations, the reader is directed to [21] for the TDNN-LSTMP model used for SWBD corpus. Both LSTM and TDNN (seed models) are trained from 4000 hrs TBD.

We perform phone-level sequence training for all the AMs, without frame level pretraining, using the LF-MMI training criterion as in [17]. The nnet3 toolkit in Kaldi speech recognition toolkit [19] is used to perform all the neural network training. All the AMs in our experiments use the same 80-dimensional FBANK features, including plus 3-dimensional pitch (raw pitch and its first and second derivates).

The language model (LM) is the same trigram LM for all speech decoding. It is trained from 160M words collected from the TBD and texts from a variety of web search engines.

3.3. Results

Results of two complementary seed models and the target LTDNN model trained from 4K hrs TBD are presented in Section 3.3.1. In Section 3.3.3 to 3.3.4, we validate the proposed AL approach for small-scale ASR, in which the baseline LTDNN model was trained from the 600 hrs TBD. Then, we try to generalize observations from these sections to large-scale ASR task in Section 3.3.5.

3.3.1. Baseline results

Table 1 shows a comparison of three types LF-MMI based models. The LTDNN results are taken as baseline for system comparison. And TDNN and LSTM models are used to decode the UTFD. It can be seen that the LSTM obtains much better performances than TDNN, that's why we choose automatic transcrip-

tions from LSTM system as reference in Figure 1. Moreover, the mixture architecture of LTDNN leads to more than 7% relative improvements. We can see that, these models have different behaviors on test sets. It indicates that using heterogeneous neural networks for AL data selection may produce complementary training examples to the TBD dataset.

In addition, it is clear to obtain that the POI and Child tasks are much more difficult to recognize than GTopic, by comparing their WERs across three models. This is due to the fact that, the TBD provides a much better match of the speech data property for the general voice search topic than the Point-of-Interest and children's speech.

Table 1: *WERs% of different systems trained on 4000 hrs TBD.*

System	GTopic	POI	Child
TDNN	13.0	16.0	33.5
LSTM	12.6	15.3	32.4
LTDNN	11.9	14.4	31.0

3.3.2. Criterions of utterance selection for HNN-based AL

For reasons of space, we only show a subset of our experiments for the examination of criterions for utterance selection, which were very extensive. From these tryout experiments, we found that those utterances in the UTFD pool with WMERs $> 50\%$, are extremely difficult to be accurately labeled by human, and those utterances with WMERs $< 20\%$ are less informative. Therefore, to balance the human labeling costs and informativeness of the selected UTFD, in Table 2, we tried four different combination ways of the CM and WMER statistics to form the criterion for AL utterance selection (indicated using $\oplus$).

Table 2: *Evaluation of different criterion for data selection. WMER refers to the value of WMER%.*

CM $\oplus$ WMER	hrs/RL	WER%		
		GTopic	POI	Child
LTDNN (600 hrs)	-	13.0	16.5	33.2
$[0, 0.7] \oplus [20, 30]$	322/0.88	12.6	14.5	31.7
$(0.7, 0.8] \oplus [20, 30]$	477/0.92	12.4	16.1	33.0
$[0.8, 0.9] \oplus [30, 40]$	375/0.91	12.6	14.8	31.6
$[0.8, 0.9] \oplus [40, 50]$	226/0.88	12.5	14.2	31.5

Numbers in the 2nd column of Table 2 give the total hours of selected UTFD utterances (indicated using hrs) and the ratio of these utterances successfully labeled by human (indicated using RL), using different data selection criterions. Such as, for the CM$= [0, 0.7]\oplus$ WMER$= [20, 30]$, 322 hrs utterances are selected, but with only $322 * 0.88 = 283.36$ hrs can be successfully labeled by human as TFD. These numbers indicate that the harder to be recognized by ASR systems of these selected UTFD utterances, the more difficult to be transcribed of them by human, and we need to pay higher manual transcription costs for them.

Furthermore, we select 100 hrs randomly from the TFD under each CM $\oplus$ WMER criterion separately. The right part of Table 2 shows the performances of LTDNNs trained from the joint 600 hrs TBD and each 100 hrs TFD. It can be seen that, comparing rows 2,4,5 with 3, the selected utterances using criterion in row 3 are much less informative than others. In addition, from rows 2,4,5, we can see that the relative gains are smaller for GTopic($3\sim3.8\%$), compared with the ones for POI ($10\sim14\%$), and Child ($5\sim7\%$). Compared with the baseline, all the significant WER improvements derived from these three rows proved the effectiveness of the proposed AL scheme. Considering both observations from the left and right part of Table 2, we can conclude that, the degree of TFD informativeness is proportional to the difficulty of manual labeling. That's to say, we need to pay a higher transcription cost to get the more informative TFD. Therefore, we choose to use the first and last two criterions in Table 2 for all the fresh data selection, and we obtained about 823 hrs TFD from the total 30K hrs UTFD.

3.3.3. Comparing different methods for AL data selection

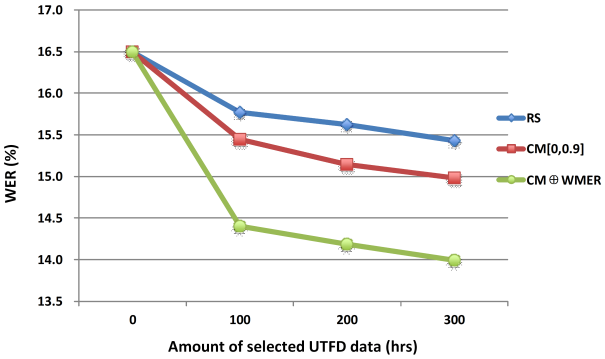

Figure 2: *Performance comparison on POI test set using different AL data selection methods.*

Figure 2 compares three techniques of UTFD data selection on POI test set: unfiltered random selection (RS), confidence filtering with CM=[0, 0.9], and the proposed CM $\oplus$ WMER. The baseline is the same LTDNN trained from 600 hrs TBD, as shown in Table 2. We found that there was around 36% of utterances with a confidence measure lower than 0.9, it was much higher than the data ratio of selected UTFD in the total 30K hrs. All the selected datasets in this figure were the ones successfully labeled by human. For CM $\oplus$ WMER, the 100, 200 and 300 hrs were randomly selected from the total 823 hrs TFD which was obtained in Table 2. It can be seen that the combination criterion of CM $\oplus$ WMER leads to above relative 6% WER improvements over the conventional data selection only using CM, and even larger gains are obtained over the RS.

3.3.4. Comparing effectiveness of TFD and TBD

It is interesting to perform experiments to improve the LTDNN AMs, by adding the TFD and TBD-add data to the baseline 600 hrs TBD as training examples. Additional 100 and 600 hrs were randomly selected from the 823 hrs TFD, TBD-add datasets were tried in the experiments. Figure 3 demonstrates the comparison of relative WER reductions (WERRs) obtained from the improved AMs, compared with the ones of LTDNN baseline.

From Figure 3, it is clear to see that by adding the same amount of 100 or 600 hrs data to the baseline, the TFD can provide much bigger performance gains over the TBD-add, especially for the POI test set which may reflect users' actual needs from the ASR engine. In fact, to meet the industrial application needs, the TBD was usually well designed in advance to build

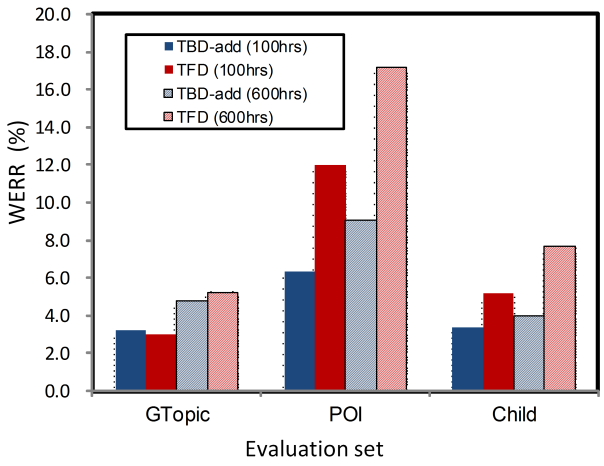

Figure 3: *Performance comparison of the LTDNN systems. The AMs were separately trained using the joint datasets of baseline 600 hrs TBD and TBD-add, TFD. WERR refers to the relative WER reduction.*

the baseline ASR system. However, for those applications as voice search, the rapid change of users' activities can result in extremely diverse data properties of test speech, such as new POI words, audio recording conditions or background acoustics, etc. It is impossible to design a TBD to fully cover every case and all these new factors. Therefore, It is indeed necessary to develop an efficient AL or semi-supervised approach to overcome new challenges.

3.3.5. AL vs SST for large-scale LTDNN training

To validate the generalization ability of the proposed AL scheme from small-scale to large-scale ASR tasks, we performed an initial experiment of large-scale active learning for LTDNN training, and compared it with the SST training based on the same heterogeneous neural networks. We increased the 30K hrs UTFD utterances to 100K hrs from the voice search engine for both large-scale SST and AL.

The baseline LTDNN was trained from the whole 4K hrs TBD. The SST approach is similar to [16], and its HNN models are the same two ones used for our AL, except using the criterion of WMER=0 $\oplus$ CM=[0.6, 0.95] to do the automatic selection of UTFD utterances and their corresponding transcriptions (the best utterance selection criterion we have tried). The same AL data selection criterion as in Section 3.3.2 to obtain the 823 hrs from the 30K hrs UTFD is directly applied for the 100K hrs UTFD.

Table 3: *Performance of large-scale LTDNN acoustic model training with different amount of training data from the SST and proposed AL based on heterogeneous neural networks.*

Method	Data size	WER%		
		GTopic	POI	Child
Baseline	4K hrs (TBD)	11.9	14.4	31.0
SST	+7K hrs	11.4	13.4	29.4
AL	+1.2K hrs	11.2	12.5	28.1
SST+ AL	+8.2K hrs	10.9	12.1	27.5

It is surprising to find that, we obtained about 7K hrs utterances with high confidence automatic transcriptions from SST, however, only around 3% utterances (around 3K hrs) were finally selected using the proposed AL approach, because the HNN models were already well trained using the 4K hrs TBD. Due to the high manual labeling cost, we only labeled 1.2K hrs selected UTFD as the TFD and added it to the 4K hrs baseline. It indicates that the manual transcription cost can be greatly reduced, when we compare the proposed data selection of AL with the random selection and the one only based on the CM as shown in Section 3.3.3. It is specially useful for very large-scale ASR tasks.

Table 3 shows the performances of LTDNN models trained from different large-scale datasets. Comparing the WERs of 1st row in Table 2 and the ones in Table 3, it can be seen that the large-scale LTDNN baseline was significantly improved. About relative $6\sim12\%$ WER reductions have been achieved on three evaluation sets. Comparing the first two rows in Table 3, it can be seen that when adding 7K hrs data from SST to the 4K hrs baseline, only around $4\sim7\%$ performance gains have been achieved. However, we can obtain relative $6\sim13\%$ WER gains by only adding 1.2K hrs TFD from AL. It indicates that, the training data size is greatly increased by the SST, but the complementary information it brings is limited, the TFD utterances selected by AL are more informative than the ones collected by SST, even for the large-scale ASR task. Furthermore, when we combined the utterances derived from SST and AL, and added them into the 4K hrs TBD, the retrained LTDNN was further improved. In addition, when we compare the improvements on GTopic with the ones on POI and Child, it can be seen that the proposed AL is very useful to improve the acoustic models for challenging evaluation tasks.

4. Conclusion and Future work

In this paper we applied ideas from recent semi-supervised training approach [16] to active learning of LF-MMI trained LTDNN acoustic models, using heterogeneous neural networks to select informative utterances. We used a combination of WMER and CM to form the data selection criterion, and experimental results proved that it outperforms the conventional CM and RS data selection significantly. We showed that the AL works very well in both small and large-scale ASR tasks. It still can produce $6\sim13\%$ relative improvements by only labeling 1.2% utterances from 100K hrs fresh data, even the baseline was already well trained from 4K hrs TBD. Furthermore, we performed the HNN-based SST and AL in an unified framework, using the same types of HNN models to do the fresh data decoding. Initial results showed that combing AL and SST can lead to further improvements. We guess that, if the HNN models used for AL data selection could be further improved, it can result in bigger reduction of the manual transcription cost and produce more useful data for improving AMs. We are currently investigating this observation on variety ASR tasks. The effects of AL to language model training is also our future work.

5. Acknowledgements

This work was funded by the Project 61701306 supported by National Natural Science Foundation of China, and the Shanghai Normal University Funds (Grant No. DCL201702). The authors would like to thank Beijing Unisound Information Technology Co., Ltd for providing part of training and test data sets.

6. References

[1] G. Hinton, D. Li, Y. Dong, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsburg, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] Y. Dong and D. Li, *Automatic Speech Recognition: A Deep Learning Approach*. Verlag London: Springer, 2015.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[4] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription," in *ASRU 2013 – IEEE Automatic Speech Recognition and Understanding Workshop, December 8-12, Olomouc, Czech Republic, Proceedings*, 2013, pp. 368–373.

[5] Y. Huang, Y. Wang, and Y. Gong, "Semi-supervised training in deep learning acoustic model," in *INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association, September 8-12, San Francisco, USA, Proceedings*, 2016, pp. 3848–3852.

[6] S. Li, X. Lu, S. Sakai, M. Mimura, and T. Kawahara, "Semi-supervised ensemble dnn acoustic model training," in *ICASSP 2017 – 42nd IEEE International Conference on Acoustics, Speech and Signal Processing, March 5-9, Orleans, USA, Proceedings*, 2017, pp. 5270–5274.

[7] G. Riccardi and D. Hakkani-Tur, "Active learning: Theory and applications to automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 504–511, 2005.

[8] D. Yu, B. Varadarajan, L. Deng, and A. Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," *Computer Speech and Language*, vol. 24, pp. 433–444, 2010.

[9] T. Fraga-Silva, J. Gauvain, L. Lamel, A. Laurent, V.-B. Le, and A. Messaoudi, "Active learning based data selection for limited resource stt and kws," in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association, September 6-10, Dresden, Germany, Proceedings*, 2016, pp. 3159–3162.

[10] K. Barnes, M. Snover, M. H. Siu, and H. Gish, "Importance sampling of delta-auc: A basis for active learning for improved keyword search," in *ICASSP 2016 – 41st IEEE International Conference on Acoustics, Speech and Signal Processing, March 20-25, Shanghai, China, Proceedings*, 2017, pp. 2244–2248.

[11] T. Drugman, J. Pylkknen, and R. Kneser, "Active and semi-supervised learning in asr: Benefits on the acoustic and language models," in *INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association, September 8-12, San Francisco, USA, Proceedings*, 2016, pp. 2318–2322.

[12] A. R. Syed, A. Rosenberg, and M. Mandel, "Active learning for low-resource speech recognition: Impact of selection size and language modeling data," in *ICASSP 2017 – 42nd IEEE International Conference on Acoustics, Speech and Signal Processing, March 5-9, Orleans, USA, Proceedings*, 2017, pp. 5315–5319.

[13] D. Hakkani-Tur and A. Gorin, "Active learning for automatic speech recognition," in *ICASSP 2002 – 27th IEEE International Conference on Acoustics, Speech and Signal Processing, May 13-17, Orlando, Florida, Proceedings*, 2002, pp. 3904–3907.

[14] S. Huang, R. Jin, and Z. Zhou, "Active learning by querying informative and representative examples," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 10, pp. 1936–1949, 2014.

[15] N. Itoh, T. Sainath, D. Jiang, J. Zhou, and B. Ramabhadran, "N-best entropy based data selection for acoustic modeling," in *ICASSP 2012 – 37th IEEE International Conference on Acoustics, Speech and Signal Processing, March 25-30, Kyoto, Japan, Proceedings*, 2012, pp. 4133–4136.

[16] N. Kanda, S. Harada, X. Lu, and H. Kawai, "Investigation of semi-supervised acoustic model training based on the committee of heterogeneous neural networks," in *INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association, September 8-12, San Francisco, USA, Proceedings*, 2016, pp. 1325–1329.

[17] 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 *INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association, September 8-12, San Francisco, USA, Proceedings*, 2016, pp. 2751–2755.

[18] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low latency acoustic modeling using temporal convolution and lstms," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 373–377, 2017.

[19] 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 *ASRU 2011 – IEEE Automatic Speech Recognition and Understanding Workshop, December 11-15, Hawaii, USA, Proceedings*, 2011.

[20] Y. Hamanaka, K. Shinoda, S. Furui, T. Emori, and T. Koshinaka, "Speech modeling based on committee-based active learning," in *ICASSP 2010 – 35th IEEE International Conference on Acoustics, Speech and Signal Processing, March 14-19, Dallas, Texas, USA, Proceedings*, 2010, pp. 4350–4353.

[21] G. Cheng, V. Peddinti, D. Povey, V. Manohar, S. Khudanpur, and Y. Yan, "An exploration of dropout with lstms," in *INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association, August 20-24, Stockholm, Sweden, Proceedings*, 2017, pp. 1586–1590.