Semi-Supervised GMM and DNN Acoustic Model Training with Multi-system Combination and Confidence Re-calibration

Yan Huang, Dong Yu, Yifan Gong, and Chaojun Liu

Microsoft Corporation, One Microsoft Way, Redmond, WA 98052

{yanhuang; dongyu; ygong; chaojunl}@microsoft.com

Abstract

We present our study on semi-supervised Gaussian mixture model (GMM) hidden Markov model (HMM) and deep neural network (DNN) HMM acoustic model training. We analyze the impact of transcription quality and data sampling approach on the performance of the resulting model, and propose a multi-system combination and confidence re-calibration approach to improve the transcription inference and data selection. Compared to using a single system recognition result and confidence score, our proposed approach reduces the phone error rate of the inferred transcription by 23.8% relatively when top 60% of data are selected.

Experiments were conducted on the mobile short message dictation (SMD) task. For the GMM-HMM model, we achieved 7.2% relative word error rate reduction (WERR) against a well-trained narrow-band fMPE+bMMI system by adding 2100 hours of untranscribed data, and 28.2% relative WERR over a wide-band MLE model trained from transcribed out-of-domain voice search data after adding 10K hours of untranscribed SMD data. For the CD-DNN-HMM model, 11.7% and 15.0% relative WERRs are achieved after adding 1K hours of untranscribed data using random and importance sampling, respectively. We also found using large amount of untranscribed data for pre-training does not help.

Index Terms: semi-supervised acoustic model training, system combination, confidence re-calibration, importance sampling

1. Introduction

The unlimited live data harvested from deployed systems contains valuable information for acoustic model training for tasks such as mobile voice search (VS) and short message dictation (SMD). Transcribing large amount of live data, however, is both costly and time-consuming. For this reason, much effort has been devoted to the unsupervised and semi-supervised acoustic model training [1, 2, 3, 4, 5, 6, 7, 8] to exploit the untranscribed data. Most of these previous works focus on generating better quality hypothesis with various offline decoding techniques and on improving confidence measure for better data selection.

Automatically inferred transcription is never perfect. It is in general believed that discriminative modeling techniques tend to be more sensitive to transcription errors than the generative modeling techniques. However, we have not seen any previous work reporting quantitative comparison results. Our simulation results in this paper show that the sensitivity to label quality is highly correlated with the discriminability of the model itself. This analysis result is used as a reference operating point for transcription inference and data selection in our study.

We further propose a ROVER-based multi-system combination and committee-based confidence re-calibration approach to improve the transcription inference and data selection. Compared to using a single system recognition result and confidence score, our approach reduces the phone error rate (PER) of the inferred transcription by 23.8% relatively when the top 60% of data are selected.

Typical data selection in semi-supervised training strives to select the data with the most accurate inferred transcription. From the active learning point of view, these data are usually less valuable [9, 10, 11]. The main challenge here is how to select the most valuable data with good quality transcription and without skewing the prior distribution. This is especially important in optimizing the semi-supervised deep neural network (DNN) acoustic model training with gigantic amount of available untranscribed data. We investigate the importance sampling technique based on the re-calibrated confidence in the Gaussian mixture model (GMM)-hidden Markov model (HMM) and apply the result to the context dependent (CD)-DNN-HMM [12] semi-supervised acoustic model training.

The rest of this paper is organized as follows: Section 2 analyzes the impact of transcription quality to the resulting model performance. Section 3 introduces the multi-system combination and confidence re-calibration approach for transcription inference and data selection. Importance data sampling approach is discussed in Section 4. Section 5 presents the experimental results of using untranscribed data in GMM-HMM and CD-DNN-HMM training. Section 6 concludes this paper.

2. Analysis on the Impact of Transcription Quality to Model Performance

In order to understand the impact of transcription quality on the resulting model performance, we conducted a simulation to investigate how different modeling approaches are affected by erroneous transcription.

Specifically, five versions of transcription for 100 hour transcribed voice search data are simulated at 2%, 4%, 6%, 12%, and 16% PER level using a recognition system with different decoding configurations. The 100 hour voice search data with six versions of transcription including the human transcription are used to train MLE (maximum likelihood estimation), fMPE (feature-minimum phone error) [13], and fMPE+bMMI (boosted maximum mutual information) [14] models separately. In particular, the following treatments were adopted to guarantee the rigor of the comparison: First, all MLE models trained from erroneous transcription share the same model structure and model size as the model trained from the human transcription; Second, this MLE model is also used as the common MLE seed model for all discriminative training in this experiment.

It is to be noted that the human transcription is not perfect

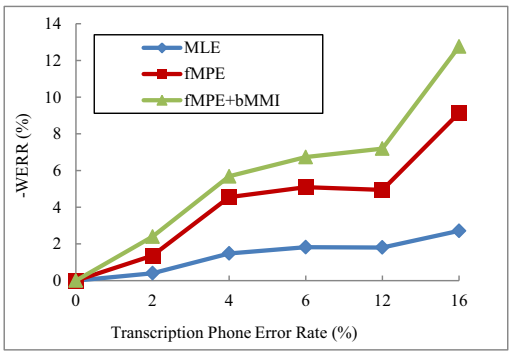

Figure 1: *Relative model performance increase versus transcription quality as compared to the model trained from the human transcription in MLE, fMPE, and fMPE+bMMI.*

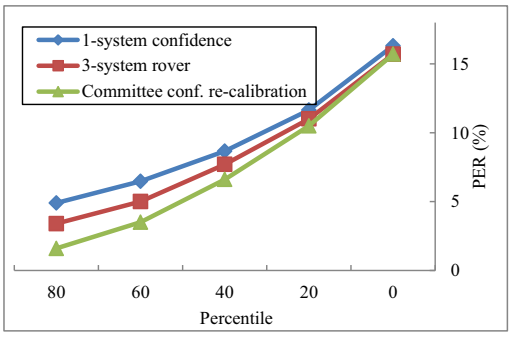

Figure 2: *Transcription phone error rate (PER) via system combination and confidence re-calibration.*

either. Here we just treat it as a reference, which is used to measure the quality of all simulated transcription.

Fig. 1 presents the relative WER increase of the models trained from transcription at different quality level compared to the model trained from the human transcription. It can be seen that the sensitivity to the label quality is highly correlated with the discriminability of the model itself. We further observe:

- At the 6% transcription PER level, the fMPE+bMMI model has about 7% relative WER increase comparing to the human transcription, while the MLE model has only about 1% relative WER increase.
- We are mostly interested in transcription quality at 6% PER level. This is our typical data operating points for untranscribed data in our interested tasks. About 5∼7% WER increase is expected at this range for the fMPE+bMMI models as compared to the manual transcription. We use Fig. 1 as a reference to choose data selection operating point for different semi-supervised acoustic model training.

3. Transcription Inference and Data Selection

Generally speaking, the higher quality the transcription, the smaller is the gap between semi-supervised training and supervised training as suggested in Fig. 1. Generating and selecting high quality transcription are the two fundamental issues in using untranscribed data for acoustic model training. Here, we propose an approach with ROVER-based multi-system combination followed by committee-based confidence re-calibration for transcription inference and data selection.

First, three ASR systems with different acoustic models and language models are used to generate initial hypothesis with word and sentence level confidence. The ROVER-based system combination is used to generate improved new hypothesis with new word and sentence level confidence. We use a linear interpolation of the confidence scores from the original systems and the degree of agreement on the hypothesis to generate word-level confidence after ROVER, which is further used to calculate the sentence level confidence. This is the standard ROVER-based system combination approach [15].

Next, a committee system consisting of a major system and three supplementary systems are formed to re-calibrate the confidence. Unlike the traditional committee approach, which treats each system equally in voting, our committee is completely biased in the following way: only the hypothesis from the major system is considered as a potential candidate, the supplementary systems are only used to re-calibrate the confidence of the major system.

Formally, let $\{S_i(h_i, c_i), i = 1, \cdots, N\}$ denotes the committee with N systems, where i is the index of the system, N is the total number of systems in the committee, h_i is the hypothesis generated from the i-th system, and c_i is the confidence of the corresponding hypothesis. $S_1(h_1, c_1)$ is the primary system and the rest are supplementary systems. We denote the decision of the committee as $S(h, c)$, where h always equals to h_1 since only the hypothesis from the primary system is considered as a potential final hypothesis. c is calculated via a re-calibration process as

$$c = \left\{ \begin{array}{ll} c_1^{\frac{\beta}{n}}, & \text{if } n > 0; \\ \gamma^{\frac{\sum_{i=1}^{N} c_i}{c_1}}, & \text{if } n = 0, \end{array} \right\}, \quad (1)$$

where β and γ are the warping factors used to adjust the re-calibrating rate. n is the number of supplementary systems which agree with the primary system. No special tuning is performed on β and γ in our experiments, which are set to 1.

The original three systems and the system after ROVER form a committee. In the third step, we run the committee at word and sentence level in a round-robin fashion, i.e. each system gets a chance to be the primary system and uses the committee to re-calibrate its own confidence. The ROVER system is assigned to be the primary system in the last committee run, when the confidence of all three original systems have been re-calibrated. The ROVER hypothesis with the re-calibrated confidence in the last committee run is used for data selection.

Fig. 2 presents the performance comparison results of 1-system confidence, 3-system ROVER, and committee-based confidence re-calibration after ROVER at different data selection operating points. At 40-percentile selection point (top 60%), 11.1% PER reduction is obtained using the ROVER hypothesis and the ROVER confidence compared to one system recognition results and one system confidence. After running the committee-based confidence re-calibration, the overall PER is not changed compared to ROVER. But at 40-percentile selection level, further 12.7% PER reduction is obtained compared to using the ROVER confidence. In total, 23.8% relative PER is obtained comparing to using one system recognition results and confidence.

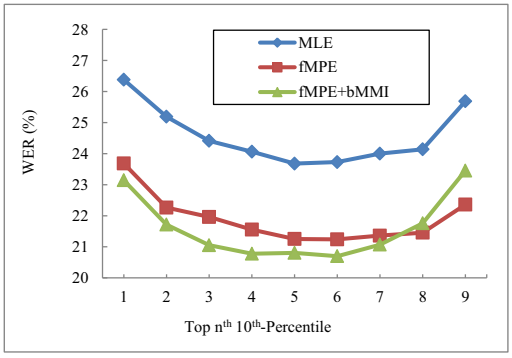

Figure 3: *Data sampling via re-calibrated confidence in MLE, fMPE, and fMPE+bMMI.*

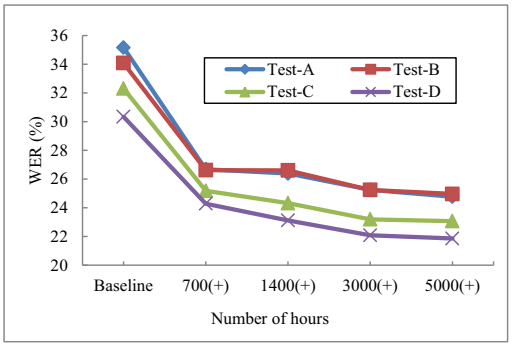

Figure 4: *Results of semi-supervised MLE model on a wide-band SMD task. 28.2% WERR is achieved after adding 10K hours live SMD data compared to the baseline MLE model trained from 400 hours of out-of-domain mobile VS data.*

4. Importance Data Sampling

Effective data sampling is important in optimizing the semi-supervised acoustic model training with nearly unlimited amount of available untranscribed data. The importance sampling here involves three different but related factors: the value of each datum to the model training, the label quality, and the change of prior distribution. We investigate the importance data sampling using re-calibrated confidence in MLE, fMPE, and fMPE+bMMI.

10M SMD utterances are first auto-transcribed and selected using the methodology described in Section 3. They are then partitioned into 1M utterances per partition based on the re-calibrated confidence. The partition with the lowest re-calibrated confidence is dropped. MLE, fMPE, and fMPE+bMMI models are trained using each one of the nine partitions. For fair comparison, the model structure and model size are kept the same across all MLE models; fMPE and fMPE+bMMI models also share the same MLE seed model.

Fig. 3 depicts the data sampling effects in MLE, fMPE, and fMPE+bMMI models:

- The resulting model performance exhibits clear "U" shaped pattern with respective to sampling effects for all three training criteria. The model trained from the 5th 10th-percentile partition outperforms the model trained from the 1st and 9th 10th-percentile by around 10% relative. The mid partitions result in better performed models compared to the two end partitions.
- It is interesting to observe that MLE and fMPE show very similar pattern of performance versus data sampling, while in our previous discussion on transcription quality factor, they exhibit distinct patterns. The prior change and the distinction of the value of each partition seem to be the dominant factors here.

We apply the importance sampling approach to select mid-partition (1M utterances) from the 10M utterances for semi-supervised DNN model training and the comparison results with the random sample will be discussed in Section 5.4.

5. Experiments and Results

Our semi-supervised training experiments were conducted on the mobile SMD task. 10K hour live SMD data collected from deployment is auto-transcribed and selected based on the multi-system combination and confidence re-calibration approach described in Section 3.

The results of using untranscribed live data to further improve the baseline MLE, fMPE+bMMI, and CD-DNN-HMM systems trained from transcribed live data and/or engineered data with various amount will be presented in the rest of this section. We will also report results on an SMD language expansion task, which indicates the proposed approach is applicable to different languages.

5.1. Semi-supervised MLE model results on a wide-band SMD task

Starting from a wide-band MLE model with 16kHz sample frequency trained from 400 hours of out-of-domain mobile voice search data, we incrementally add selected untranscribed live SMD data in the MLE model training. The model size is increased as more data is added. The baseline model and the semi-supervised MLE modes are evaluated using four different live test sets collected during different months of the year.

As shown in Fig. 4, when the first 700 hours untranscribed data are added, since this is the first time the model is exposed to the in-domain SMD data, we observe the largest performance jump with 22.0% WERR. When we incrementally add more untranscribed data in the amount of 1400, 3000, and 5000 hours, the performance keeps on improving but with much slower pace. When the total 10K hour untranscribed data is added, 28.2% WERR are achieved averaged among the four test sets.

We note that in addition to the wide-band model itself, two narrow-band models with significantly better initial performance are also used as seed models for the transcription generation. The seed models used in transcribing the four batches of untranscribed data are dynamically updated with improved acoustic models from the previous round. The combined system effectively speeds up the performance ramping up for the wide-band model. When significant amount of untranscribed data are added, the performance gap between the two models noticeably shrinks.

5.2. Semi-supervised fMPE+bMMI model results on a narrow-band SMD task

In this task, we start with a baseline narrow-band SMD fMPE+bMMI model with 8kHz sample frequency well-trained from large amount of transcribed data. After adding two batches of untranscribed data in the total amount of 2100 hours in the fMPE+bMMI training directly, we achieve 7.2% average WERR with the detailed results summarized in Table 1.

Table 1: *Results of semi-supervised fMPE+bMMI model on a narrow-band SMD task. 7.2% average WERR is achieved against a well-trained fMPE+bMMI baseline model after adding 2100 hours untranscribed live SMD data .*

Test Set	Word Count	WERR
Test-A	22809	7.9%
Test-B	16028	5.4%
Test-C	38874	6.7%
Test-D	41031	7.2%
Ave WERR	134735	7.2%

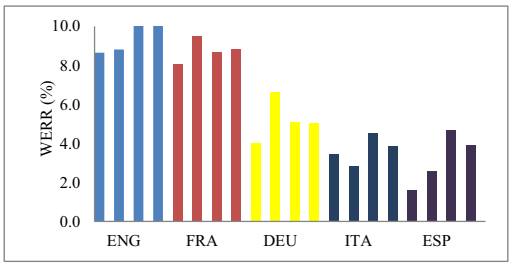

Figure 5: *Results of semi-supervised training on five European languages in the SMD language expansion task. Here each bar represents the WERR on a test set.*

It is to be noted that further performance gain is expected if we re-train the MLE model especially when significantly more untranscribed data is added. We are currently experimenting with a full model update using the 10K hour untranscribed data. Normal practice such as increasing model size and Gaussian component will be applied.

5.3. Semi-supervised training in SMD language expansion

Besides EN-US, we also applied the similar semi-supervised training approach to five European languages in the SMD language expansion task. The baseline models are fMPE+bMMI models trained from in-domain transcribed live data and engineered data. After adding 200 $\sim$1000 hours untranscribed live data, 5 $\sim$10% WERRs are obtained as summarized in Fig. 5.

The results indicate that the proposed approach is language independent and can be directly applied to non-English languages. It is particularly valuable for low resource languages, though a few more iterations of semi-supervised training and transcription re-generation may be needed due to lack of good seed models.

5.4. Semi-supervised training in CD-DNN-HMM on an SMD task

The baseline CD-DNN-HMM system is trained using 60K transcribed SMD utterances with 36K Gaussian components and 1812 senone states. The topology of the DNN net is 468-2048-2048-2048-1812 with three 2048-demension hidden layers. The neural net input is a 468-dimension feature vector formed by 52-dimension MFCC with 9-frame context windows.

1M supervised untranscribed SMD utterances are selected from 10M utterances either by random sampling or by importance sampling based on re-calibrated confidence as described in Section 4. These two sets will be referred to as "1M.Random" and "1M.Sample" in the rest of this section. The senone state alignment for both transcribed and untranscribed data are generated using the same MLE seed model trained from 60K transcribed SMD utterances. The model size is not increased as significantly more untranscribed data is added for DNN training. With a pre-trained DNN using 60K transcribed utterances, we conduct model refinement using the baseline 60K transcribed utterances, "1M.Random", and "1M.Sample" separately.

Table 2 presents the performance comparison results of the baseline model ("DNN.60K") and the semi-supervised DNN models trained by adding 1M untranscribed utterances by random sampling ("DNN.60K+1M.Random") or by importance sampling ("DNN.60K+1M.Sample"). After adding 1M randomly selected utterances to the baseline 60K transcribed data in the model refinement, we obtain 11.7% WERR. In comparison, adding 1M utterances using importance sampling results in 15.0% WERR. Importance sampling generates 3.6% further relative performance gain compared to random sampling.

Table 2: *Results of semi-supervised DNN model training in an SMD task. 11.7% and 15.0% relative WERRs are achieved after adding 1K hours of untranscribed data using random sampling or importance sampling, respectively.*

Model	Test-A	WERR
DNN.60K (Baseline)	25.1%	NA
DNN.60K+1M.Random	22.2%	11.7%
DNN.60K+1M.Sample	21.4%	15.0%

The results indicate that the importance sampling analysis results on GMM-HMM model as discussed in Section 4 is applicable to the DNN model. It would be interesting to see whether the DNN model will present the similar "U"-shaped curve as shown in Figure 3, which can helps us understand how different modeling approaches are affected differently by the change in the prior distribution, the value of the data, and the label quality.

Besides semi-supervised DNN training, we also experimented with large amount untranscribed data in DNN pre-training. We found it helps neither in improving the resulting model performance nor in speeding up the convergence rate.

6. Conclusions

In summary, we proposed a multi-system combination and confidence re-calibration approach to improve the transcription inference and data selection. Compared to using a single system recognition result and confidence score, when top 60% of data are selected, our proposed approach reduces the phone error rate of the inferred transcription by 23.8% relatively. The impact of transcription quality and data sampling approach on the performance of the resulting model was also analyzed.

For the GMM-HMM model, we achieved 7.2% relative word error rate reduction (WERR) against a well-trained narrow-band fMPE+bMMI system by adding 2100 hours of untranscribed data, and 28.2% relative WERR over a wide-band MLE model trained from transcribed out-of-domain voice search data after adding 10K hours of untranscribed SMD data. We also proved the proposed approach can be applied to other non-English languages as well.

For the CD-DNN-HMM model, 11.7% and 15.0% relative WERRs are achieved after adding 1K hours of untranscribed data using random sampling or importance sampling, respectively. We also found using large amount of untranscribed data for pre-training does not help.

7. References

[1] Lamel, L. F., Gauvain, J. L., Smith, J. O. and Adda, G., "Unsupervised Acoustic Model Training", 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Volume: 1, Page(s): 877 - 880, 13-17 May 2002.

[2] Lamel, L. F., Gauvain, J. L., Smith, J. O. and Adda, G., "Investigating Lightly Supervised Acoustic Model Training", 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Volume: 1, Page(s): 477 - 480, 7-11 May 2001.

[3] Wessel, F. and Herrmann, N. J., "Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition", IEEE Transactions on Audio, Speech, and Language Processing (TASLP), Volume: 13, Issue: 1, Page(s): 23 - 31, Jan 2005.

[4] Yu, K., Gales, M, Wang L., and Woodland, P. C., "Unsupervised Training and Directed Manual Transcription for LVCSR", Journal of Speech Communication,Vol. 52, Issue 7-8, Page(s): 652-663 , July 2010.

[5] Yu, K., Gales, M.J.F., and Woodland, P.C., "Unsupervised Training for Mandarin Broadcast News and Conversation Transcription", 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 4, Page(s): 353-356, 2007.

[6] Yu, K., Gales, M.J.F., and Woodland, P.C., "Unsupervised Training with Directed Manual Transcription for Recognizing Mandarin Broadcast Audio", 8th Annual Conference of the International Speech Communication Association (Interspeech), August 2007.

[7] Ma, J., Matsoukas, S., Kimball, O., and Schwartz, R., "Unsupervised Training on Large Amount of Broadcast News Data", 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 2, Page(s): 349-352, May 2006.

[8] , Zhang R. and Rudnicky E.I., "Improving the Rerformance of an LVCSR System through Ensembles of Acoustic Models", 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 1, Page(s): 876-879, April, 2003.

[9] Hakkani-tr, D. and Riccardi, G., "Theory and Applications to Automatic Speech Recognition", IEEE Transactions on Speech and Audio Processing (TASLP), Volume: 13, Issue: 4, Page(s): 504 - 511, 2005.

[10] Hakkani-tr, D., Riccardi, G., and Gorin, A., "Active Learning for Automatic Speech Recognition", 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Volume: 1, Page(s):3904-3907, 13-17 May 2002.

[11] Kamm, T.M.and Meyer, G.G.L., "Selective Sampling of Training Data for Speech Recognition", Proceedings of the 2nd International Conference on Human Language Technology Research, Page(s): 20-24, 2002.

[12] Dahl, G.E., Yu, D., Deng, L., and Acero, A., "Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition", IEEE Transactions on Audio, Speech, and Language Processing (TASLP) - Special Issue on Deep Learning for Speech and Language Processing, Volume: 1, No. 1, Page(s): 33-42, Jan 2012.

[13] Povey, D, Kingsbury, B., Mangu, L., Saon, G., Soltau, H., and Zweig, G., "fMPE: Discriminatively Trained Features for Speech Recognition", 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Volume: 1, Page(s): 961 - 964, March 2005.

[14] Povey, D, Kingsbury, B., Ramabhadran, B., Saon, G., Soltau H., and Visweswariah, K., "Boosted MMI for model and feature-space discriminative training", 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Page(s): 4057 - 4060, March 2008.

[15] Fiscus, J. G., "A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)", 1997 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Page(s): 347 - 354, 14-17 Dec 1997.