A Comparative Analytic Study on the Gaussian Mixture and Context Dependent Deep Neural Network Hidden Markov Models

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

Microsoft Corporation, One Microsoft Way, Redmond, WA 98052

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

Abstract

We conducted a comparative analytic study on the context-dependent Gaussian mixture hidden Markov model (CD-GMM-HMM) and deep neural network hidden Markov model (CD-DNN-HMM) with respect to the phone discrimination and the robustness performance. We found that the DNN can significantly improve the phone recognition performance for every phoneme with 15.6% to 39.8% relative phone error rate reduction (PERR). It is particularly good at discriminating certain consonants, which are found to be "hard" in the GMM. On the robustness side, the DNN outperforms the GMM at all SNR levels, across different devices, and under all speaking rate with nearly uniform improvement. The performance gap with respect to different SNR levels, distinct channels, and varied speaking rate remains large. For example, in CD-DNN-HMM, we observed 1∼2% performance degradation per 1dB SNR drop; 20∼25% performance gap between the best and least well performed devices; 15∼30% relative word error rate increase when the speaking rate speeds up or slows down by 30% from the "sweet" spot. Therefore, we conclude the robustness remains to be a major challenge in the deep learning acoustic model. Speech enhancement, channel normalization, and speaking rate compensation are important research areas in order to further improve the DNN model accuracy.

Index Terms: GMM-HMM, CD-DNN-HMM, noise robustness, channel compensation, speaking rate normalization

1. Introduction

The success of the deep neural network in the large vocabulary speech recognition is one of the greatest breakthroughs in speech recognition technologies over the last decade. Following the seminar work of applying the deep learning and context-dependent deep neural network hidden Markov model (CD-DNN-HMM) to the large vocabulary continuous speech recognition tasks (LVCSR) [1], different research groups published consistent performance gain over the best discriminatively trained context-dependent Gaussian mixture hidden Markov model (CD-GMM-HMM) on the public research benchmark switchboard (SWB) [2, 3] task. Meanwhile, the industry world quickly launched this technology in their production systems and reported comparable results [4, 5]. To this end, [6] summarized the shared views of four research groups on the deep neural network for acoustic modeling.

As is known, diverse acoustic environment, distinct channels, various speaking style are the fundamental challenges in the LVCSR acoustic modeling. With the significant performance gain widely reported in the deep neural network acoustic modeling, we would like to find out, amongst these acoustic model challenges, what have been solved gracefully and to what degree; what are still left as intriguing problems? The ultimate goal is to identify several important research areas to further improve the deep learning acoustic model performance.

Previous study [7] shows that the gain of the deep learning acoustic model is mostly due to the fact that DNN can extract more invariant and selective features through many layers of nonlinear feature transformation. In this paper we answer the question whether this property helps to boost performance differently or equally for different phonemes and under different SNRs, channels, and speaking rates.

Specifically, we adopted an analytic methodology and conducted a deep error pattern analysis on a pair of CD-DNN-HMM and CD-GMM-HMM with respect to the phone discrimination and the performance pattern on three selected robustness factors. Each addresses one acoustic model challenge described earlier. To the best of our knowledge, this kind of analytic study for the deep learning acoustic model is not available.

Our study shows that the deep neural network acoustic model can significantly improve the phone discrimination with 27.9% PER reduction. It performs particularly well in discriminating certain consonants, which are found to be "hard" in the GMM. On the robustness side, we found that the DNN outperforms the GMM at all SNR levels, across all devices, and under all speaking rate. This is consistent with the study in [7].

Nevertheless, the DNN seems to generate uniform performance improvement under different conditions. Our study shows 1∼2% performance degradation per 1dB SNR drop; 20∼25% word error rate gap between the best and the least well performed devices; 15∼30% accuracy drop when the speaking rate speeds up or slows down by 30% from the "sweet" spot. This suggests that the noise robustness, channel normalization, and speaking rate compensation remain to be the important areas in the deep learning acoustic model.

The remainder of this paper is organized as follows: Section 2 introduces the analytic methodology and the LVCSR task in this study; Section 3 compares the overall phone discrimination of the CD-GMM-HMM and CD-DNN-HMM; Section 4 analyzes the performance pattern of the CD-GMM-HMM and CD-DNN-HMM with respect to the different SNR level, channel, and speaking rate. Section 5 concludes this study.

2. Analytic Methodology and Task

The mobile voice search (VS) and short message dictation (SMD) serves vast types of mobile devices used by millions of users with distinct speaking styles in diverse acoustic environments. This real world speech application embeds almost all key LVCSR acoustic model challenges and therefore was cho-

sen as an ideal task for this study. Moreover, besides our interest in analyzing and improving such a practical system, the large volume of the available analytic material with a broad coverage of the real life acoustics can ensure the statistical significance. We don't know of any good alternative choice from the public domain speech database more suitable for this analytic study.

Specifically, we trained a pair of GMM and DNN models with a comparable setup using 400 hr VS/SMD data. The GMM is a discriminative model trained with the feature-space minimum phone error rate (fMPE) [10] and the boosted MMI (bMMI) [9] criteria. The front-end is the 39-dimension MFCC feature. The DNN was trained using the cross entropy (CE) criteria and the front-end is the 87-dimension log filter bank (LFB) feature with a context window of 11 frames. The two models shared the same training data, decision tree, and the same MLE seed model used for the lattice generation in the GMM and the senone state alignment in the DNN.

The analytic material consists of 100 hr VS/SMD test data randomly sampled from the deployment with roughly the same distribution as the training data. A list of interested meta tags, extracted from the search log or generated offline, were used to partition the analytic material into disjoint "condition" specific analytic sets. The "condition specific analytic data sets were then used to evaluate and compare the distinct error patters for the pair of GMM and DNN models. In particular, for the continuous valued meta tags, e.g. the SNR and the speaking rate, we implemented some simple smoothing for a more consistent and smoothed error pattern.

The overall performance comparison of the CD-GMM-HMM and CD-DNN-HMM is summarized in Table 1. We obtain 20.3% and 25.1% WERRs in the DNN compared to the baseline GMM for the VS and SMD task respectively.

Next, we will first analyze the distinct phone error pattern in the CD-GMM-HMM and CD-DNN-HMM; then compare the robustness performance pattern of the two models with respect to the different SNR, channel, and speaking rate.

Table 1: *Overall performance comparison of the 400 hr CD-GMM-HMM and CD-DNN-HMM VS/SMD models.*

Task	GMM(%)	DNN(%)	WERR (%)
VS	30.4	24.3	20.3
SMD	19.9	15.0	25.1

3. Phone Error Pattern Analysis

Figure 1 illustrates the phone error rate (PER) of the CD-GMM-HMM and CD-DNN-HMM rendered by the decreasing order of the phone error rate reduction (PERR). It can be seen that the phone error rate for every phoneme was reduced in the DNN model and the PERRs range from 15.6% to 39.8%. On average, the CD-DNN-HMM yields 27.9% PERR compared to the CD-GMM-HMM. This indicates that the DNN can generate significantly better classification boundary than the GMM discriminatively trained with the fMPE and bMMI criteria.

We further observed that the DNN model is more effective in modeling consonants comparing to the GMM. Certain consonants, which are "hard" to discriminate in the GMM, obtain notably larger performance boost in the DNN model. Overall, the DNN model exhibits a much smoother PER contour. In particular, "[zh]" and "[dn]" have significantly higher PERs compared to all other phonemes in the GMM. Their PERs drop from 21.9% and 20.9% to 14.9% and 14.5% respectively in the DNN

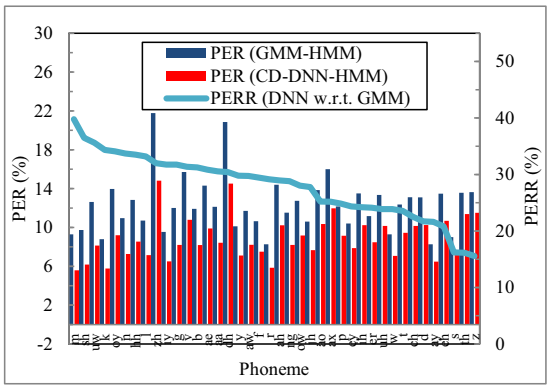

Figure 1: *Phone error rate (PER) of the CD-GMM-HMM and CD-DNN-HMM rendered by ordering the phonemes by the decreasing order of phone error rate reduction (PERR).*

model. These two phonemes remain on the top PER phoneme list in the DNN model though with much smaller performance gap with other phonemes.

We conducted the phone error analysis for the VS and the SMD task separately and obtained the similar results. This further confirmed the observation we made is task independent.

4. Robustness Performance Analysis

In this section, we will discuss and compare the robustness performance pattern of the GMM and DNN models with respect to the different SNR, channel, and speaking rate using the methodology described in Section 2.

4.1. Noise Robustness

The environmetal noise can significantly degrade the speech recognition performance. Technologies that allow the speech recognition perform well in the diverse acoustic conditions is criticall for the success of the mobile speech recognition.

To study the noise robustness of the deep learning acoustic model, we compared the error pattern of the GMM and DNN models under different SNR levels with the results summarized in Figure 2 and Figure 3 for the VS and SMD respectively. The CD-DNN-HMM significantly outperforms the CD-GMM-HMM at all SNR levels. The consistent performance gain across all SNR levels suggests that the DNN is in general a more powerful model which can improve the ASR performance not only on the clean speech but also on the noisy speech with a wide range of noise levels. Further comparing the performance of the CD-DNN-HMM across the different SNRs, we found that the CD-DNN-HMM yields almost the uniform performance gain over the CD-GMM-HMM. This distinct pattern is shared between the VS and SMD tasks.

To measure the noise robustness of the DNN, we calculated the relative performance degradation per 1dB SNR drop. For the VS, as the SNR drops from 40dB to 0dB, the WERs increase from 18% to 34% and the SNR per dB drop introduces about 2% relative performance degradation. For the SMD, within the same SNR range, the WERs increase from 12% to 18% and the SNR per dB drop results in 1% relative performance degradation. The quantitative difference of the sensitivity to the noise level between these tasks is due to the fact that the SMD has much lower LM perplexity.

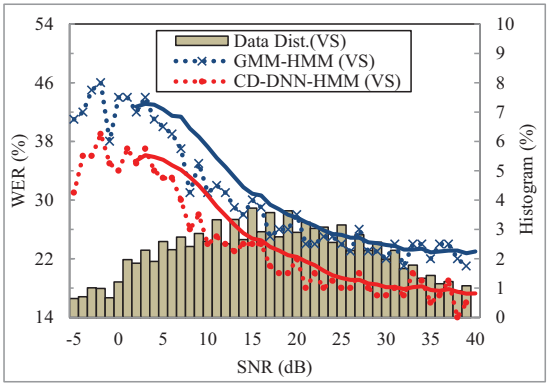

Figure 2: *Performance comparison of CD-GMM-HMM and CD-DNN-HMM at different SNR levels for the VS task.*

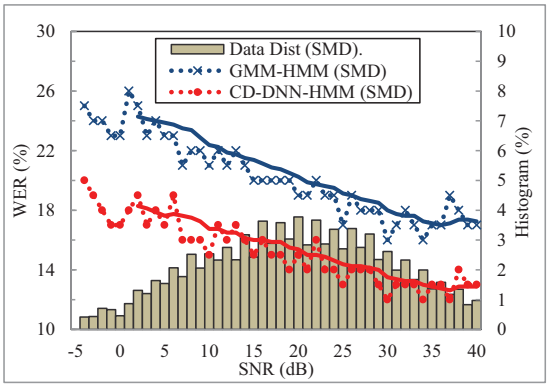

Figure 3: *Performance comparison of the CD-GMM-HMM and CD-DNN-HMM at different SNR levels for the SMD task.*

The speech recognition performance of the DNN drops significantly as the noise level increases within the normal range for the mobile speech application. This suggests the noise robustness remains as an important research area. Speech enhancement, noise robust acoustic features, or other multi-condition learning technologies need to be explored to bridge the performance gap and further improve the overall performance of the deep learning based acoustic model.

4.2. Channel Mismatch

The channel mismatch is another major source of the speech recognition performance degradation. The channel robustness issue is a traditional speech recognition robustness topic which has been researched for many years. It is also particularly important for the mobile speech application since typically the mobile application serves a large number of different devices from many different phone manufactures. The channel robustness is an indispensable feature for a successful mobile speech recognition system.

In this session, we discuss whether the channel mismatch issue still exists as a distinct speech recognition robustness problem or it has been largely resolved with the invariant and selective feature learning in the deep learning technology. Specifically, we compared the performance of the CD-GMM-HMM and CD-DNN-HMM on four different mobile devices from different manufactures with the comparison results summarized in

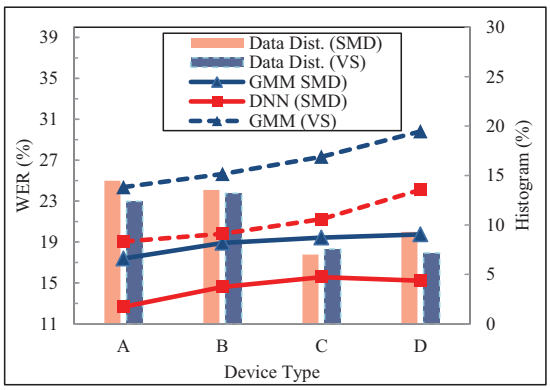

Figure 4: *Performance comparison of CD-GMM-HMM and CD-DNN-HMM models on four different devices.*

Figure 4.

In the CD-DNN-HMM, we observed consistent word error rate reduction for all four selected devices both on the VS and SMD tasks comparing to the CD-GMM-HMM. The DNN yields nearly uniform performance gain on all four devices and the pattern is shared between the VS and SMD tasks. This once again verifies that the DNN is a generally more discriminative model compared to the GMM. The uniform performance improvement across all four devices suggests a nice property in the deep learning that it helps improving the best performed device as much as it helps the least well performed device.

We further compared the performance gap across different devices on the CD-DNN-HMM. On the VS task, the WERs of the selected four devices range from 19% to 23% or 20% relative WER difference between the best and the least well performed device. Similar trend was observed in the SMD task. The performance variance across devices appears to be as large as in the GMM.

To this end, we think the channel robustness still exists as a distinct robustness issue and it remains to be further researched in the deep learning acoustic model. Besides the traditional channel normalization methodologies, developing channel normalization technologies within the deep learning framework is promising given the deep learning capability and the deep neural network capacity.

4.3. Varied Speaking Rate

Speaking rate variation is known to affect the speech intelligibility and degrade the speech recognition performance especially under the mismatched training and testing condition [11, 12, 13]. In the mobile speech recognition applications, the speaking rate varies largely depending on the different speakers, speaking mode, and speaking styles. This requires the acoustic model to gracefully handle speech with varied speaking rate. We would like to find out how the DNN model performs on varying speaking rate compared to the GMM.

Figure 5 and Figure 6 illustrate the performance comparison across different speaking rate for the VS and SMD task. Here the speaking rate was measured by the number of phones per second. We have also adopted some of its variations such as the number of vowels per second or the normalized speaking rate by taking into account the average duration for different phonemes in this study. Similar performance pattern with respect to the speaking rate was observed. Therefore, we simply

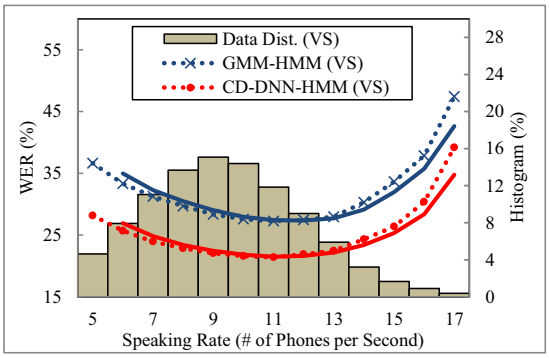

Figure 5: *Performance comparison of the CD-GMM-HMM and the CD-DNN-HMM at different speaking rate for the VS task.*

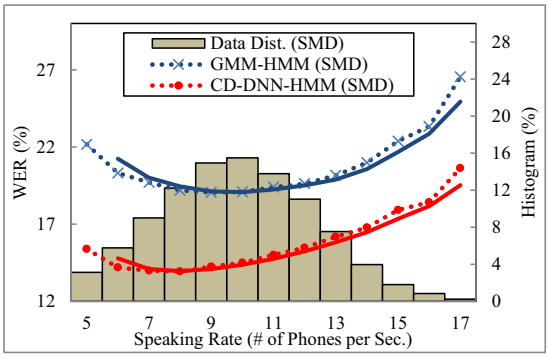

Figure 6: *Performance comparison of the CD-GMM-HMM and the CD-DNN-HMM at different speaking rate for the SMD task.*

adopted the number of phones per second as a measure for the speaking rate in this paper.

We found that the CD-DNN-HMM consistently outperforms the CD-GMM-HMM with almost uniform performance gain across all speaking rate. This again proves that the DNN model is a better discriminative model. In comparing the performance of the CD-DNN-HMM across different speaking rate, we observe the "U"-shaped pattern for both VS and SMD. On the VS, the speaking rate "sweet spot" is around 10 to 12 phones per second. When the speaking rate deviates 30% from the "sweet spot" (either speeds up or slows down), 30% relative word error rate increase is observed.

In the shared acoustic model scenario for the voice search and short message dictation, the speaking rate varies both within and between the two tasks. They have different "sweet spots" and also exhibit slightly different error pattern with respect to the speaking rate change. On the SMD, we observe 15% relative word error rate increase when the speaking rate deviates 30% from the "sweet spot".

Extremely fast or slow speech may result in speech recognition performance degradation due to the following reasons: First, it may change the acoustic score dynamic range. Second, the fixed frame rate, frame length, and context window size may be inadequate to capture the dynamics in transient speech events for fast or slow speech and therefore result in sub-optimal modeling. Third, the extremely fast or slow speech may result in slight formant shift due to the human vocal instrumentation

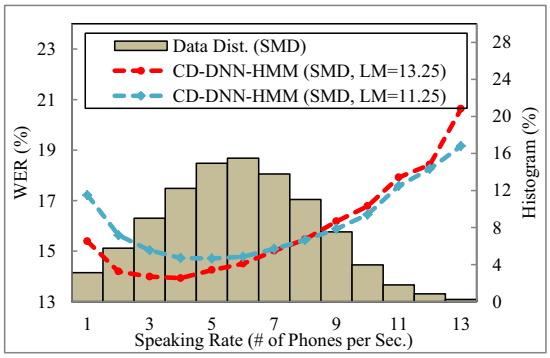

Figure 7: *Performance comparison of the CD-GMM-HMM and the CD-DNN-HMM at different speaking rate for the SMD task.*

limitation. Last, other phonological changes such as the phone deletion and the fragmented word may accompany with the extremely fast speech.

We conducted an initial experiment to investigate the effect of adjusting the LM interpolation weight for the extremely fast or slow speech on the SMD task. As shown in Figure 7, decreasing the LM interpolation weight can yield moderate WER reduction for the fast speech and result in small performance degradation for the slow speech. Overall, the speaking rate "sweet" spot shifts slightly to the faster speech region. This verified our hypothesis on the effect of the acoustic score dynamic range change on the ASR performance of the extremely fast or slow speech. Nevertheless, the small performance change suggests that the speaking rate compensation problem is a modeling issue requiring the model level solution.

The large performance gap across different speaking rate in the CD-DNN-HMM suggests it is possible to further improve the DNN model performance via the effective speaking rate compensation methodologies.

5. Conclusion

In summary, we conducted an analytic error analysis on a pair of GMM and DNN models using significant amount of analytic material on the mobile VS/SMD task. Our study suggests that the DNN acoustic model is a generally more discriminative model. The DNN can significantly improve the phone discrimination with the phone error rate reduction ranging from 15.6% to 39.8%. It is particularly good at discriminating certain consonants, which are found to be "hard" in the GMM.

On the robustness side, the DNN outperforms the GMM at all SNR levels, across all devices, and under all speaking rate with nearly uniform improvement under different conditions. Nevertheless, the performance gap with respect to different SNR levels, distinct channels, and varied speech rate remains large. For example, in CD-DNN-HMM, we observed 1∼2% performance degradation per 1dB SNR drop; 20∼25% relative WER gap between the best and least well performed devices; 15∼30% WER increase when the speaking rate speeds up or slows down by 30% from the "sweet" spot.

Therefore, we conclude that robustness remains as a major challenge in the deep learning acoustic model. Speech enhancement, channel normalization, and speaking rate compensation are important areas to further improve the DNN model accuracy.

6. References

[1] Dahl, G.E., Yu, D., Deng, L., and Acero, A., "Large Vocabulary Continuous Speech Recognition With Context-Dependent DBN-HMMS", in the Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2011.

[2] Seide, F., Li, G., and Yu, D., "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks", in the Proceedings of Interspeech 2012.

[3] Kingsbury, B., Sainath, N. T., and Soltau, H., "Scalable Minimum Bayes Risk Training of Deep Neural Network Acoustic Models Using Distributed Hessian-free Optimization", in the Proceedings of Interspeech 2012.

[4] Jaitly, N., Nguyen, P., Senior, A., Vanhoucke, V., "Application of Pretrained Deep Neural Networks to Large Vocabulary Speech Recognition", in the Proceedings of Interspeech 2012.

[5] Deng, L., Li, J., Huang, J., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., He, X., Williams, J., Gong, Y., and Acero, A., "Recent Advances in Deep Learning for Speech Research at Microsoft", in the Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

[6] Hinton G., Deng, L., Yu, D., Dahl G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath,, and Kingsbury, B., "Deep Neural Networks for Acoustic Modeling in Speech Recognition", IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97, November 2012.

[7] Yu, D., Seltzer, M., Li, J. ,Huang, J., Seide, F., "Feature Learning in Deep Neural Networks - Studies on Speech Recognition Tasks", in the Proceedings of 2013 International Confernece on Learning Representation, 2013.

[8] Li, D., "Three Classes of Deep Learning Architectures and Their Applications: A Tutorial Survey", APSIPA Transactions on Signal and Information Processing, 2013.

[9] Povey, D, Kingsbury, B., Ramabhadran, B., Saon, G., Soltau H., and Visweswariah, K., "Boosted MMI for model and feature-space discriminative training", in the Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2008.

[10] Povey, D, Kingsbury, B., Mangu, L., Saon, G., Soltau, H., and Zweig, G., "fMPE: Discriminatively Trained Features for Speech Recognition", in the Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2005.

[11] Chu, S. and Povey, D. "Speaking Rate Adaptation Using Continuous Frame Rate Normalization", in the Proceedings of the 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2010.

[12] Yeh, C., Lee, H., and Lee. L, "Speaking Rate Normalization with Lattice-based Context-dependent Phoneme Duration Modeling for Personalized Speech Recognizers on Mobile Devices", in the Proceedings of Interspeech 2013.

[13] Zhu, Q. and Alwan, A., "On the use of Variable Frame Rate Analysis in the Speech Recognition", in the Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2000.