INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Robust Speech Recognition using Generalized Distillation Framework

Konstantin Markov[1], *Tomoko Matsui*[2]

[1]University of Aizu, Fukushima, Japan
[2]Institute of Statistical Mathematics, Tokyo, Japan

markov@u-aizu.ac.jp, tmatsui@ism.ac.jp

Abstract

In this paper, we propose a noise robust speech recognition system built using generalized distillation framework. It is assumed that during training, in addition to the training data, some kind of ”privileged” information is available and can be used to guide the training process. This allows to obtain a system which at test time outperforms those built on regular training data alone. In the case of noisy speech recognition task, the privileged information is obtained from a model, called ”teacher”, trained on clean speech only. The regular model, called ”student”, is trained on noisy utterances and uses teacher's output for the corresponding clean utterances. Thus, for this framework a parallel clean/noisy speech data are required. We experimented on the Aurora2 database which provides such kind of data. Our system uses hybrid DNN-HMM acoustic model where neural networks provide HMM state probabilities during decoding. The teacher DNN is trained on the clean data, while the student DNN is trained using multi-condition (various SNRs) data. The student DNN loss function combines the targets obtained from forced alignment of the training data and the outputs of the teacher DNN when fed with the corresponding clean features. Experimental results clearly show that distillation framework is effective and allows to achieve significant reduction in the word error rate.

Index Terms: speech recognition, noise robustness, DNN-HMM acoustic model, generalized distillation, privileged information

1. Introduction

Recently, there has been a surge in studies of automatic speech recognition (ASR) using DNN-HMM acoustic models, and they have shown that such models can achieve much higher performance than traditional GMM-HMM combination [1]. One particular challenge in ASR is robustness against environmental noises and a lot of various techniques and methods have been developed during the last decades. Naturally, with the increased popularity of DNN, their noise robustness has also been investigated. For example, Seltzer et.al.[2] reported that DNN obtains comparable performance to the best GMM system with various noise reduction, feature enhancement and model compensation methods. This is attributed to the property of DNN to learn higher level representation of the features which is inherently less prone to environment variations [3]

In some studies, researchers try to apply noise robust methods developed for GMM-HMM systems in DNN based models. Thus, Abe et.al.[4] combines classical spectral subtraction method before feeding feature vectors to the DNN. In addition, they perform noise estimation and use it during DNN training. This approach, called noise-aware training has been previously proposed in [2] and [5]. A way to estimate noise robust features using deep denoising autoencoders (DAE) is studied in [6]. Such neural networks learn the mapping between noisy and clean features. Robustness can also be improved by explicitly modeling left and right temporal contexts in features windows [7].

The paradigm of *machines-teaching machines* has been investigated in studies of Vapnik [8][9] and Hinton [10]. Motivated by the principles of human education, authors incorporate an ”intelligent teacher” into machine learning. It is assumed that for each feature-label pair, there is an additional information about it provided by a teacher to support the learning process. However, teacher information will not be available at test time. This framework is also known as *learning using privileged information*. Such approach allows to build a classifier which is better than those built on the regular features alone. On the other hand, Hinton proposed the concept of distilling the knowledge in neural networks [10], where a simple machine learns a complex task by imitating the solution of a more complicated and flexible machine. This can be applied in cases when a fast or real time operation is required, but using the flexible machine is computationally prohibitive.

In a recent study [11], the learning using privileged information and the distillation methods have been combined into a *Generalized Distillation* framework which utilizes the strengths of both methods. Here, the teacher who has access to the privileged information plays the role of the more complicated machine in the distillation process. After the simpler student is learned through the distillation process, it is used for testing when no privileged information is available. Generalized Distillation is closely related to applications in methods such as semi-supervised learning, domain adaptation, transfer learning, Universum learning [12] and curriculum learning [13].

In this study, we apply the Generalized Distillation in the speech recognition task in order to improve the noise robustness of the ASR system. As privileged information we utilize clean speech data to learn the teacher machine. The student machine is learned on noisy speech and guided during the training by the teacher which has access to the clean version of the same speech. Our ASR system is a DNN-HMM hybrid where DNN is used to predict HMM state probabilities. Such systems are popular since they allow to utilize the high performance of the DNN with the conventional and well established decoding and language modeling methods. The HMM transition probabilities are obtained from a traditional GMM-HMM acoustic model and the DNN is learned using the above mentioned generalized distillation framework.

In our experiments, we used the Aurora2 database [14] which is a popular corpus for researching noise robustness and provides parallel clean / noisy speech utterances for training. Previously, this database has been also utilized to investigate

http://dx.doi.org/10.21437/Interspeech.2016-852

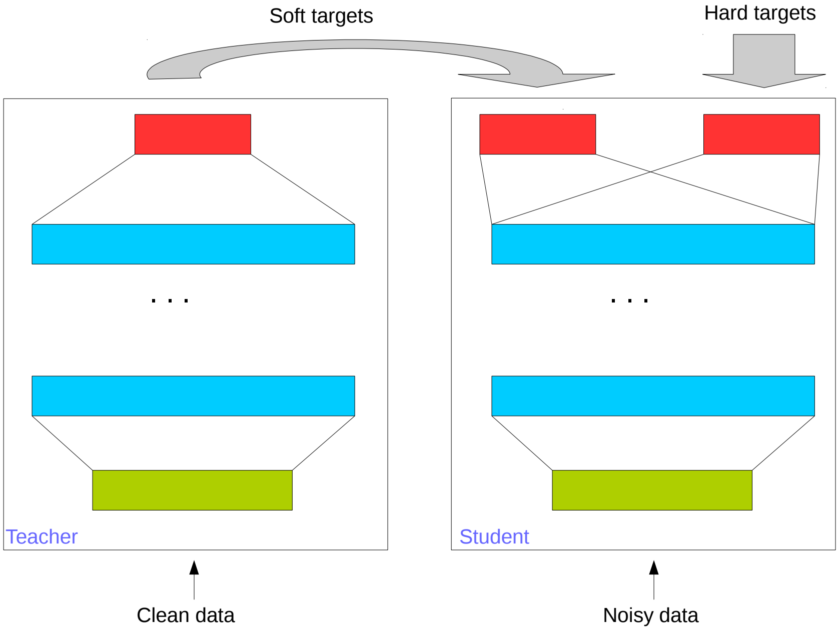

Figure 1: *Student training block diagram.*

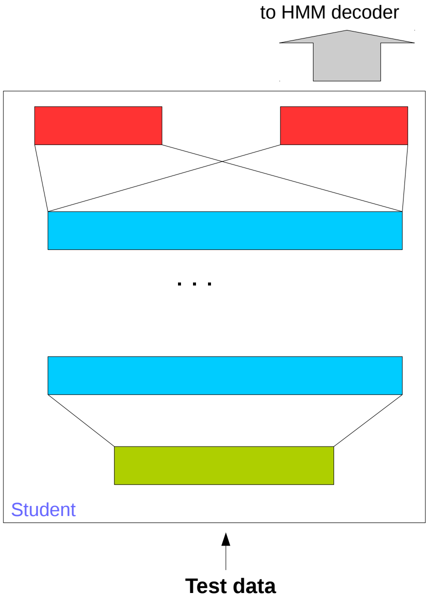

Figure 2: *Testing with student DNN.*

DNN robustness. In [15] and [16], Vinyals et.al have trained DNN to predict phone posteriors which are further concatenated with the MFCC features (Tandem features) and fed to a standard GMM-HMM system. Various training approaches and DNN structures are compared and big improvements are reported.

2. Generalized Distillation

Generalized distillation has been termed in [11] to frame two techniques of Hinton's distillation[10] and Vapnik's privileged information[9] that enable machines to learn from other machines. While a simple machine learns a complex task by imitating the solution of a flexible machine in Hinton's distillation, a machine learns from other machines in Vapnik's privileged information. In the framework, an "intelligent teacher"is incorporated into machine learning and the training data is formed by a collection of triplets

$$(x_1, x_1^*, y_1), \ldots, (x_n, x_n^*, y_n) \sim P^n(x, x^*, y),$$

where x_i, y_i is a feature-label pair and x_i^* is additional information about x_i, y_i provided by an intelligent teacher. The teacher is assumed to develop a language that effectively communicates information to help the student come up with better representation and to enable to learn characteristics about the decision boundary which are not contained in the student samples.

The process is as follows:

1. Learn teacher $f_t \in \mathcal{F}_t$ in eq. (1) using $\{(x_i^*, y_i)\}_{i=1}^n$.

$$f_t = \arg\min_{f \in \mathcal{F}_t} \frac{1}{n} \sum_{i=1}^{n} l(y_i, \sigma(f(x_i^*))) + \Omega(||f||) \quad (1)$$

   Here, $x_i^* \in \mathcal{R}^d$, $y_i \in \Delta^c$, Δ^c is the set of c-dimensional probability vectors, F_t is a class of functions from $\mathcal{R}^d$ to $\mathcal{R}^c$, $\sigma : \mathcal{R}^c \to \Delta^c$ is a soft-max function, l is a loss function and Ω is an increasing function which serves as a regularizer.

2. Compute teacher soft labels $\{\sigma(f_t(x_i^*)/T\}_{i=1}^n$ using temperature parameter $T > 0$.

3. Learn student $f_s \in \mathcal{F}_s$ in eq. (2) using $\{(x_i, y_i)\}_{i=1}^n$, $\{(x_i, s_i)\}_{i=1}^n$ and imitation parameter $\lambda \in [0, 1]$.

$$f_s = \arg\min_{f \in \mathcal{F}_s} \frac{1}{n} \sum_{i=1}^{n} [(1-\lambda) l(y_i, \sigma(f(x_i))) + \lambda l(s_i, \sigma(f(x_i)))] \quad (2)$$

$$s_i = \sigma(f_t(x_i)/T) \in \Delta^c \quad (3)$$

Here, $\mathcal{F}_s$ is a function class simpler than $\mathcal{F}_t$.

In this paper, we utilize DNN to learn representation of both f_t and f_s.

3. System Description

Our system is a hybrid DNN-HMM system, where DNN is used to predict HMM state posterior probabilities given an input data vector. These probabilities are converted to likelihoods using state priors and standard decoding is performed to obtain the recognition result.

We apply the Generalized Distillation framework for the DNN training only. Targets for the DNN learning are obtained by first training a conventional GMM-HMM system using only clean data. Then, target states are identified by forced alignment.

Next, we learn the teacher DNN according to the first step of the procedure described in the previous section. The train data x_i^* are clean speech vectors and the targets y_i are one-hot vectors where the component corresponding to the target state is 1 and all other components are set to 0. After training, the parameters of the teacher DNN are fixed.

Student DNN learning procedure is illustrated in Fig.(1). Outputs of the teacher DNN are used as soft targets s_i and together with the hard targets y_i act as arguments of the student DNN loss function as in Eq.(2). In addition, teacher DNN outputs are smoothed with the temperature parameter T according to Eq.(3). The input training data for the student DNN are noisy (mix of several SNRs and clean speech), and are fed in batches. The corresponding clean data, also in batches, are given to the teacher DNN input. However, only student DNN parameters are updated during this procedure.

During the test, only student DNN is used and the state probability predictions from the ”hard” output, i.e. the output that was compared with the hard targets during training, are fed to the HMM decoder as shown in Fig.(2).

4. Experiments

For experiments with the generalized distillation framework, we adopted the Aurora2 database [14] which provides parallel clean and noisy training data. There are 8440 clean speech utterances from 55 male and 55 female speakers. They are equally split into 20 subsets and 4 different noises (train, babble, car and exhibition) at 5 different SNRs (20dB, 15dB, 10dB, 5dB and clean condition) are added to each subset respectively. The test data are divided into three sets, A, B and C. Set A has the same types of noise as the training data and set B has four new noises - restaurant, street, airport and train station. For set C, there are only two noise conditions - train and street, but with additional channel distortions. For all three test sets, the noise SNR ranges from 0dB to clean, where only the 0dB condition is not present in the training data.

Speech signal is processed in a standard way. We use 12 MFCC coefficients with power component and their first and second derivatives. For comparison, FBANK features are extracted from 24 log filter-bank energies and their delta and delta-delta coefficients. All feature vectors are mean and variance normalized on per utterance level.

A conventional GMM-HMM is built using clean training data according to the Aurora2 recipe. It uses word level HMMs with 16 states and 3 Gaussians per state and the MFCC features. The silence model has 3 states with 6 Gaussians each. In total there are 179 states. The language model is a simple equal probability digit loop network. This system achieves an average WER of 0.87% for clean test data and 25.68% for the multi condition data. Using the clean training data, with the GMM-HMM system we generate frame level DNN training labels.

4.1. Teacher DNN

Selecting optimal DNN structure and training parameters can be quite time consuming. Following some other studies [7] and [4], we set the input window of 17 feature vectors, resulting in 663 or 1224 input nodes when using MFCC or FBANK features. The output layer always has 179 nodes as the number of HMM states. We varied the number of hidden layers from 3 to 5 and the number of nodes in each hidden layer from 1024 to 3072.

In contrast to some other approaches, we don’t use layer-wise pre- training. Weights in each layer are uniformly initialized and the activation function is Rectifying Linear (ReLU). The output layer uses SoftMax activation. Since the DNN operates in classification mode, the standard objective is Categorical Cross-entropy. The optimization method is Stochastic Gradient Descent (SGD) with learning rate of 0.01 and momentum of 0.9. No learning rate decay is used and the training is stopped if validation data (10% of the training data) loss starts increasing or the maximum of 100 iterations is reached.

First, we tested the teacher DNN in frame level state classification mode. The number of hidden layers did not have big effect on the accuracy which was around 83.5%. We found, however, that dropout influences the performance and it is highest when 20%-30% of the nodes are removed.

Next, we did speech recognition experiment using teacher DNN to provide the HMM state probabilities. For the clean only test data, the average WER was only 0.33%, and for the all multi-condition tests - 11.24%. This is 2 to 3 times better than the GMM-HMM system.

We also compared the performance of MFCC and FBANK features. Although it was quite similar, in most cases MFCC gave little bit better results. Other studies have found FBANK to be superior, but we didn’t observe such phenomenon. In all the following experiments MFCC features were utilized.

4.2. Student DNN

The structure and training parameters of the student DNN are similar to the teacher DNN. However, we found out that the number of hidden layers and amount of dropout have bigger influence on the student DNN frame level state classification performance. Consequently, the variation of speech recognition WER was bigger than the one from the teacher.

First, we learned the student DNN without distillation, i.e. without the help of the teacher DNN. The dropout percentage was set to 20%. All other training parameters were the same as for the teacher DNN. As training data, we use the whole multi-condition training set - 5 different SNRs and 4 different noise types. The frame level HMM state classification rate of the student was 74.6%, 78.0%, 79.7% and 80.1% for 3,4,5 and 6 hidden layers respectively. Table 1 shows the student DNN performance for different number of hidden layers and SNR conditions averaged over all noise types.

Table 1: *Student DNN performance on the test set in terms of WER when trained alone.*

SNR	Number of hidden layers			
	3	4	5	6
clean	0.73	0.69	0.69	0.70
20dB	0.79	0.81	0.88	0.88
15dB	1.10	1.15	1.10	1.18
10dB	2.28	2.24	2.21	2.24
5dB	6.37	6.00	5.88	5.89
0dB	21.93	20.65	20.31	20.32
Average	5.53	5.26	5.18	5.20

Based on the results from this table, for the following experiments with distillation, we choose student DNN with 5 hidden layers. The temperature parameter in Eq.(3) was varied from 1 to 5 and the imitation value was changed from 0 to 1 in steps of 0.2. Learning of the distilled student was performed as illustrated in Fig.1.

Since the teacher is trained on clean data and its performance on noisy data has little meaning, we first compare the results using the clean only part of all the test sets. Note that for test C, even the clean data have channel distortion. Figure 3 shows the performance of the teacher, student and distillation training. The dashed line shows the result of the student when trained alone, so it doesn't depend on the temperature or imitation parameters. The teacher result serves as the higher bound distilled student can achieve. As can be seen from the figure, for $T = 1$ and $\lambda = 0.8$, distillation result is very close to the teacher. Higher temperature values, however, perform worse which can be explained with the smoothing effect they have on the teacher output.

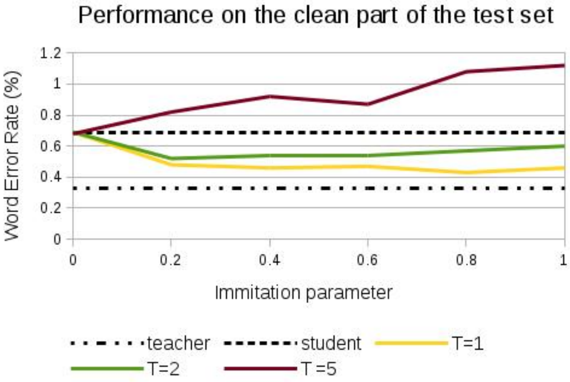

Figure 3: *Results using only clean test data.*

Of course, we are interested in the result of distillation on noisy data. It is summarized in Fig.4, where the average performance over all SNR conditions and noise types is shown. Again, the result of the student when trained alone is denoted with dashed line. The effect of distillation is clear in this case as well. The best performance is again at $T = 1$ and $\lambda = 0.8$ and is 10.2% better than the result of the student alone.

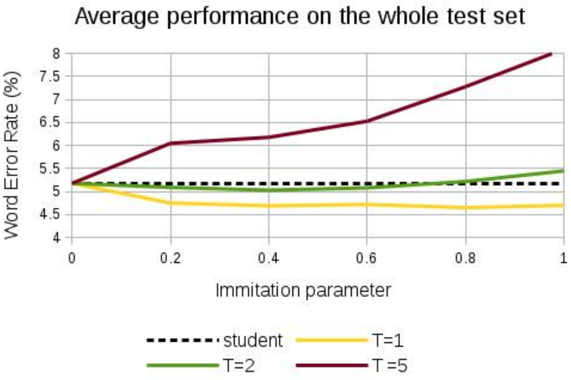

Figure 4: *Results on all multi-condition test sets.*

5. Conclusions

In this work, we proposed a noise robust ASR system, where acoustic model DNN is trained using the Generalized Distillation framework. It is an example of the *machines-teaching-machines* paradigm where machines learn from other machines. The teacher DNN trained on ”good” clean data provides guidance to the student which learns from noisy data. Using DNN in the acoustic model provides big performance boost compared to the conventional GMM-HMM systems and we have confirmed this observation is our experiments as well. The student DNN trained without distillation achieves 5.18% average WER, while the GMM-HMM system result is 11.24%. When we applied the distillation framework, additional 10.2% performance improvement was achieved leading to WER as low as 4.65%.

This is the first attempt to apply the generalized distillation framework for noisy speech recognition. We believe there are other issued to be investigated within this framework including the effect of the teacher performance on training data, more sophisticated ways of ”teaching”, not just linear combination of loss functions, etc.

6. References

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[2] M. L. Seltzer, D. Yu, and Y. Wang, “An investigation of deep neural networks for noise robust speech recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7398–7402.

[3] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, “Feature learning in deep neural networks-studies on speech recognition tasks,” *arXiv preprint arXiv:1301.3605*, 2013.

[4] A. Abe, K. Yamamoto, and S. Nakagawa, “Robust speech recognition using dnn-hmm acoustic model combining noise-aware training with spectral subtraction,” in *INTERSPEECH*, 2015.

[5] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “Dynamic noise aware training for speech enhancement based on deep neural networks.” in *INTERSPEECH*, 2014, pp. 2670–2674.

[6] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, “Extracting and composing robust features with denoising autoencoders,” in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1096–1103.

[7] B. Li and K. C. Sim, “Modeling long temporal contexts for robust dnn-based speech recognition,” in *INTERSPEECH*, 2014, pp. 353–357.

[8] V. Vapnik and A. Vashist, “A new learning paradigm: Learning using privileged information,” *Neural Networks*, vol. 22, no. 5, pp. 544–557, 2009.

[9] V. Vapnik and R. Izmailov, “Learning using privileged information: Similarity control and knowledge transfer,” *Journal of Machine Learning Research*, vol. 16, pp. 2023–2049, 2015.

[10] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.

[11] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik, “Unifying distillation and privileged information,” in *ICLR*, 2016.

[12] J. Weston, R. Collobert, F. Sinz, L. Bottou, and V. Vapnik, “Inference with the universum,” in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 1009–1016.

[13] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 41–48.

[14] H.-G. Hirsch and D. Pearce, “The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[15] O. Vinyals, S. V. Ravuri, and D. Povey, “Revisiting recurrent neural networks for robust ASR,” in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4085–4088.

[16] O. Vinyals and N. Morgan, “Deep vs. wide: depth on a budget for robust speech recognition.” in *INTERSPEECH*, 2013, pp. 114–118.