Stochastic Multiple Choice Learning for Acoustic Modeling

Bin Liu[1,2] Shuai Nie[1,2] Shan Liang[1] Zhanlei Yang[1] Wenju Liu[1]
[1] *National Laboratory of Patten Recognition, Institute of Automation, Chinese Academy of Sciences*
[2] *School of Artificial Intelligence, University of Chinese Academy of Sciences*
Beijing, China
{bin.liu2015,shuai.nie,sliang,zhanleiyang,lwj}@nlpr.ia.ac.cn

***Abstract*—Even for deep neural networks, it is still a challenging task to indiscriminately model thousands of fine-grained senones only by one model. Ensemble learning is a well-known technique that is capable of concentrating the strengths of different models to facilitate the complex task. In addition, the phones may be spontaneously aggregated into several clusters due to the intuitive perceptual properties of speech, such as vowels and consonants. However, a typical ensemble learning scheme usually trains each submodular independently and doesn't explicitly consider the internal relation of data, which is hardly expected to improve the classification performance of fine-grained senones. In this paper, we use a novel training schedule for DNN-based ensemble acoustic model. In the proposed training schedule, all submodels are jointly trained to cooperatively optimize the loss objective by a Stochastic Multiple Choice Learning approach. It results in that different submodels have specialty capacities for modeling senones with different properties. Systematic experiments show that the proposed model is competitive with the dominant DNN-based acoustic models in the TIMIT and THCHS-30 recognition tasks.**

***Index Terms*—Stochastic Multiple Choice Learning, acoustic modeling, automatic speech recognition, ensemble learning**

I. Introduction

Acoustic model is the key component of the hybrid automatic speech recognition (ASR) system. It is used to compute the posterior probabilities over HMM states. A large vocabulary continuous speech recognition system generally needs to model thousands of fine-grained senones, which requires the strong acoustic modeling capability. At present, deep learning techniques have dominated the acoustic modeling fields. Many deep models, including deep neural networks (DNNs) [1] [2], convolutional neural networks (CNNs) [3] [4] and recurrent neural networks (RNNs) [5] [6] have widely been applied into acoustic modeling and achieved significant performance improvement. In addition, many novel variants, such as highway long short-term memory RNN [7] and residual LSTM [8], are developed in order to further promote the acoustic modeling capabilities. With the increasing depth and complicating architecture of the network, the modeling capability is constantly growing. But even so, it is still a challenging task to indiscriminately model thousands of fine-grained senones only by a single model. Moreover, it's hard to train very deep and complicated networks in practice [9].

In fact, the speech phones may be spontaneously aggregated into several clusters due to the intuitive perceptual properties of speech, such as vowels and consonants. It can be expected to improve the classification performance of fine-grained senones by an ensemble of acoustic models, in which different submodels have specialty capacities for modeling senones with different properties.

Ensemble learning is a well-known technique that is capable of concentrating the strength of different models to facilitate the complex task. It has attracted more and more attentions of speech recognition community [10] [11] [12] [13]. For instance, Microsoft latest speech recognition system uses the combination of various convolutional and LSTM acoustic models and reaches human parity on the Switchboard corpus [14]. However, these ensemble schemes usually train each submodule independently and don't explicitly consider the internal relation of data. It can be expected to further improve the classification performance of fine-grained senones by mining the underlying distribution of data. Kirchoff and Bilmes [15] discuss the problem of acoustic model combination, using a soft version of the "min" function which is differentiable and can thus be jointly backpropagated through the entire ensemble. But this technique needs to backpropagate for all submodules of the ensemble during training, which is time consume and not applicable to the large acoustic models.

In this paper, we use a novel training schedule for DNN-based ensemble acoustic model. In the proposed training schedule, all submodels are jointly trained to cooperatively optimize the loss objective by a Stochastic Multiple Choice Learning (SMCL) approach. Compared with boosting methods that train each submodel of ensemble sequentially by data resampling [16], SMCL trains all ensemble members concurrently and is time efficient. It results in that different submodels have specialty capacities for modeling senones with different properties.

The rest of this paper is organized as follows: the related work is discussed in Section II. We present the Stochastic Multiple Choice Learning for acoustic modeling in Section III. In Section IV, we demonstrate the effectiveness of the SMCL method. Finally, we draw conclusions and discuss future work in Section V.

II. Related Work

Ensemble learning focuses on diversity between submodules and expects to combine the benefits of both various model

architectures and different input features. Moreover, submodels can be trained complementarily by re-sampling data [17], negative correlation learning [18] and Adaboost [19], but these iterative methods that require costly retraining aren't applicable to the DNN-based acoustic models. In addition, [15] uses a soft version of the "min" function which can be jointly backpropagated through the entire ensemble. [21] describes a system composed of several different expert networks plus a gating network that decides which of the experts should be used for each training case. [22] extends the mixture of experts to a stacked model with multiple sets of gating and experts. However, these techniques need to backpropagate for all submodules of the ensemble during training, which is time consume. In practice, various acoustic models are combined for score fusion [20] [23], which obtain WER improvements.

There is a large amount of work on the topic of extracting multiple diverse solutions from a single model [24] [25] [26]. The Multiple Choice Learning (MCL) scheme is formalized in [27], which explicitly minimizes the joint loss over the outputs of an ensemble. [28] presents a max-margin formulation in which ensemble members are learned sequentially like k-means method to minimize an upper-bound on the loss function. [29] introduces a stochastic gradient descent based algorithm to train ensemble members concurrently.

III. Stochastic Multiple Choice Learning for Acoustic Modeling

In this section, we briefly describe the training process of the ensemble learning. In addition, we elaborate the technology of stochastic multiple choice learning for acoustic modeling, consisting of the joint loss formulation and training algorithm.

A. Ensemble Learning for acoustic modeling

Various neural networks for acoustic modeling are usually trained by optimizing the cross entropy (CE) objective function,

$$\ell_{\text{CE}} = -\sum_{i} \log P(\mathbf{s}_i^*|\boldsymbol{x}_i, \boldsymbol{\Phi}) \tag{1}$$

where $\boldsymbol{s}_i^*$ are the state-level forced alignment targets, $\boldsymbol{x}_i$ are the input features, i is frame index, and $\boldsymbol{\Phi}$ is the model. The models can then be trained by sequence training method with either the Maximum Mutual Information (MMI) criterion,

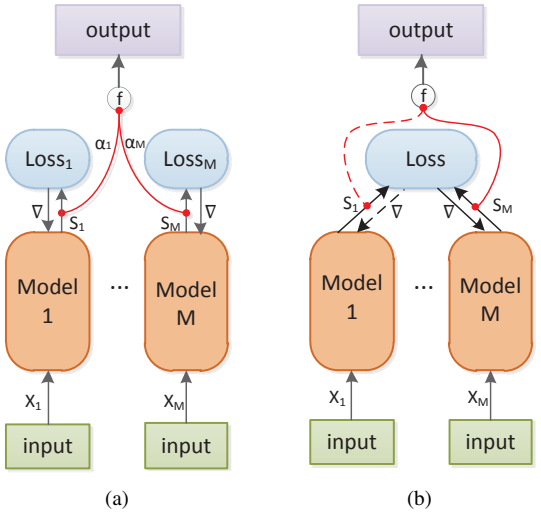
$$\ell_{\text{MMI}} = -\sum_{u} P(\boldsymbol{w}_u^*|\boldsymbol{x}_u, \boldsymbol{\Phi}) \tag{2}$$

or the Minimum Bayes Risk (MBR) criterion,

$$\ell_{\text{MBR}} = \sum_{u}\sum_{\boldsymbol{w}} A(\boldsymbol{w}_u, \boldsymbol{w}_u^*) P(\boldsymbol{w}_u|\boldsymbol{x}_u, \boldsymbol{\Phi}) \tag{3}$$

where u is the utterance index, $\boldsymbol{w}_u$ are the sentence hypotheses space, $\boldsymbol{w}_u^*$ are the corresponding correct hypotheses, and $A(\boldsymbol{w}_u, \boldsymbol{w}_u^*)$ is a loss function between $\boldsymbol{w}_u$ and $\boldsymbol{w}_u^*$.

Each submodule in the ensemble acoustic model is trained at frame or sequence level. After each submodule is well-trained independently, the frame-level combination used is the weighted average of the frame posteriors, which is illustrated in the Fig. 1(a). The bottom of the Fig. 1(a) represents the training process where each submodule has its own loss function and the red curve at the top is the weighted average score fusion for decoding,



(a) (b)

Fig. 1. Comparison of ensemble learning and stochastic multiple choice learning. (a)structure of ensemble learning; (b)structure of stochastic multiple choice learning.

$$P(\boldsymbol{s}|\boldsymbol{x}, \boldsymbol{\Phi}) = \sum_{m=1}^{M} \alpha_m f\left(P\left(\boldsymbol{s}_m|\boldsymbol{x}_m, \boldsymbol{\Phi}_{\boldsymbol{m}}\right)\right) \tag{4}$$

where $\boldsymbol{s}_m$ are the m^{th} submodel outputs, M is the ensemble size, $f(\cdot)$ presents linear or log-linear function, and α_m are the ensemble mixture weights, such that $\sum_m \alpha_m = 1$ and $\alpha_m \geqslant 0$.

B. Stochastic Multiple Choice Learning for Acoustic Modeling

For acoustic modeling, we consider training an ensemble of neural networks to together produce a set of outputs with the minimal joint loss.

Given inputs $\boldsymbol{x}_i \in \mathcal{X}$ and alignment targets $\boldsymbol{s}_i^* \in \mathcal{S}$, our goal is to learn a function $h : \mathcal{X} \rightarrow \mathcal{S}^M$ which maps each input to M outputs. We fix the form of h to be an ensemble of M models f such that $h(\mathbf{x}) = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \cdots, f_M(\boldsymbol{x}))$. For acoustic modeling, loss $\ell(\boldsymbol{s}^*, \boldsymbol{s})$ measures the error between true and predicted outputs $\boldsymbol{s}^*$ and $\boldsymbol{s}$. And we define the joint loss of h as

$$\mathcal{L}_O = \sum_{i} \min_{m\in[M]} \ell\left(\boldsymbol{s}_i^*, f_m\left(\boldsymbol{x}_i\right)\right) \tag{5}$$

In order to directly minimize the loss for an ensemble of learners, [27] present a max-margin formulation that minimizes an upper-bound on loss function (5). This objective replaces the min in the joint loss with flag variables $(\rho_{i,m})_{m=1}^{M}$

where $\rho_{i,m}$ is 1 if model m has the smallest error on sample i,

$$\begin{aligned} &\arg\min_{f_m,\rho_{m,i}} \sum_{i=1}^{n}\sum_{m=1}^{M} \rho_{i,m}\ell\left(\boldsymbol{s}_i^*, f_m\left(\boldsymbol{x}_i\right)\right) \\ &s.t. \sum_{m=1}^{M} \rho_{i,m} = 1, \rho_{i,m} \in \{0,1\} \end{aligned} \tag{6}$$

[27] also propose the block-coordinate descent algorithm where parameters of model and $\{\rho_{i,m}\}$ are optimized iteratively. However, this algorithm requires retraining models multiple times and is not applicable to neural networks, because it is time-consuming to train a deep model once. To overcome this shortcoming, [29] propose a stochastic algorithm for differentiable models which alternates the assignment step with batch updates in stochastic gradient descent. Consider the partial derivative of the objective in (6) with respect to the output of the m^{th} single model on sample x_i,

$$\frac{\partial \mathcal{L}_O}{\partial f_m\left(\boldsymbol{x}_i\right)} = \rho_{i,m}\frac{\partial \ell\left(\boldsymbol{s}_i, f_m\left(\boldsymbol{x}_i\right)\right)}{\partial f_m\left(\boldsymbol{x}_i\right)} \tag{7}$$

Notice that if model f_m has the minimum error for example $\boldsymbol{x}_i$, then $\rho_{i,m} = 1$, and the progress of gradient back-propagation is the same as training a single model; otherwise, the gradient is zero. This behavior applies to a straightforward optimization strategy for models trained by stochastic gradient descent(SGD). The structure of this algorithm is shown in the Fig. 1(b). Comparing with conventional ensemble method, this algorithm can be trained under a joint loss. For each batch, we pass the samples through the models and calculate losses from each submodel for each sample. During the backward propagation, according to the 'pick-one-model' constraint, the gradient of the loss for each sample is back-propagated only to the model that has the lowest error. The black dashed line of the Fig. 1(b) means no gradient backward propagation for the submodel that isn't selected, which decreases the computation cost.

Equation (6) can be generalized by loosening the constraints with 'pick-k-model', i.e. $\sum_{m=1}^{M}\rho_{i,m} = k$, where k is a robust parameter that controls data overlap between models. The forward propagation is unchanged and the loss gradient is back-propagated to the top k best models. If $k = M$, all models learn the same function, which is the same as the traditional ensemble method. And we analyze the effect of k in our experiments.

This approach called Stochastic Multiple Choice Learning (SMCL) is shown in algorithm 1. SMCL is applicable to a broad range of complicated neural networks by stochastic gradient descent. Unlike the iterative training method that [27] propose, ensemble models can be trained concurrently by SMCL. We perform calculating losses at the most straightforward frame level. Nevertheless, training ensemble models by SMCL at the fullsequence level is considerably more complex and will not be discussed in this paper.

Algorithm 1 SMCL method

Input: Dataset $\boldsymbol{x}_i \in \mathcal{X}$, $\boldsymbol{s}_i^* \in \mathcal{S}$, randomly initialized deep models $f_1\left(\boldsymbol{x}\right), \cdots, f_M\left(\boldsymbol{x}\right)$ parameterized by $\theta_1, \cdots, \theta_M$
Output: Ensemble of M learned models $f_1\left(\boldsymbol{x}\right), \cdots, f_M\left(\boldsymbol{x}\right)$

```
repeat
  mini-batches B
  for m = 1 to M do
    // forward propagation
    s_{m,1}, ..., s_{m,B} ← f_m(B)
    for i = 1 to |B| do
      // select k lowest error model per example
      m*_{1~k} ← arg min^{m∈[1,...,M]}_{1~k} ℓ(s*_i, s_{m,i})
      // backward propagation
      θ_{m*_{1~k}} = θ_{m*_{1~k}} − λ∂ℓ_{m*_{1~k},i}/∂θ_{m*_{1~k}}
    end for
  end for
until convergence
return f_1(x), ..., f_M(x)
```

IV. Experiments

A. Datasets and setup

We systemically evaluate the Stochastic Multiple Choice Learning for acoustic modeling on the TIMIT [30] and THCHS-30 Mandarin Chinese corpus [31], which is realized by the Kaldi speech recognition toolkit [32].

For TIMIT corpus, the standard 462-speaker training set is used and a separate validation set from 50 speakers' data is used for early stopping. Results are reported using the 24-speaker core test set. The THCHS-30 corpus is collected and transcribed by the center for speech and language technologies(CSLT) at Tsinghua University, which contains more than 30 hours speech, including 10893 utterances in the training set and 2496 utterances in the test set, respectively. And 90% of the training set is used for training acoustic models and the remaining for validation.

In following experiments, we take the 40-dimensional filterbanks as the input features, and each dimension of features is normalized to have zero mean and unit variance over the training set. To capture temporal information, 11 frames of context features were concatenated as the final inputs. The frame-level targets required in training comes from a well-trained Gaussian Mixture Model (GMM-HMM) system. After state clustering, the neural network for TIMIT and THCHS-30 have 1969 and 3480 targets respectively.

B. Models

In all experiments, Deep Neural Networks (DNNs) are used for acoustic modeling. We don't use any explicit methods to make the ensemble models diverse during training except for using different random seeds for the DNNs weight initializations, which is sufficient to demonstrate the SMCL training method. All DNNs are first initialized with layer-wise discriminative pretraining [33]. The initial weight values were drawn uniformly from the interval [-0.02, 0.02]. Nonlinear

activation function used is Rectifier Linear Unit (ReLU) [34]. The learning rate is reduced by half with an initial value of 1.0e-04 when the performance of validation set becomes worse.

The baseline system is classical ensemble in which each model is trained independently with different weight initializations. The combined frame posterior probabilities are calculated according to (4) and we set $\alpha_m = \frac{1}{M}$, where M is the ensemble size. Then the results are passed up to the HMMs as observation likelihoods, and used with standard decoding. We will refer to these as classical methods.

The proposed acoustic model is trained with the CE criterion at the first few epochs, getting the better initial points, then fine-tuned concurrently using SMCL training method. "6L-1024H" means a DNN which has 6 hidden layers with 1024 neurons in each layer. At the test time, we set α according to (8) for weighted average,

$$\alpha_m = e^{P(y_m)} / \sum_{m'=1}^{M} e^{P(y'_m)} \tag{8}$$

where $P(y_m)$ represents the prediction accuracy of the m_{th} model maximum output in the development set and the denominator is the normalization factor.

C. Results

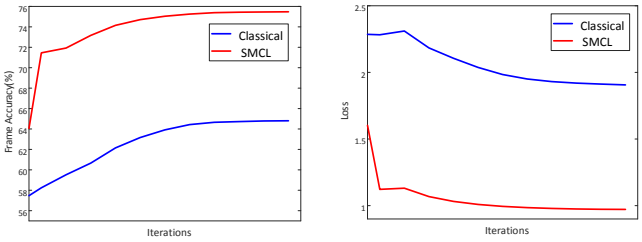
TABLE I
PER (%) ON TIMIT. 'SINGLE' HAS ONE DNN. 'ENSEM-N' MEANS AN ENSEMBLE OF N DNNS. 'NP' IS THE NUMBER OF PARAMETERS. '6L-1024H' DENOTES THE MODEL HAS 6 HIDDEN LAYERS WITH 1024 NEURONS IN EACH LAYER. 'CLASSICAL' AND 'SMCL' REPRESENTS TRAINING BY CLASSICAL ENSEMBLE LEARNING AND SMCL RESPECTIVELY.

Model	Architecture	NP	Classical	SMCL
Single	6L-1024H	7.72M	19.97	19.97
Ensem-2	6L-1024H	15.44M	18.50	17.51
Ensem-3	6L-1024H	23.16M	18.10	17.08
Ensem-4	6L-1024H	30.88M	17.33	16.97
Ensem-4	6L-512H	7.72M	18.24	17.31

TABLE II
WER (%) ON THCHS-30. 'SINGLE' HAS ONE DNN. 'ENSEM-N' MEANS AN ENSEMBLE OF N DNNS.

Model	Architecture	NP	Classical	SMCL
Single	6L-1024H	10.17M	23.56	23.56
Ensem-2	6L-1024H	20.34M	23.27	23.08
Ensem-3	6L-1024H	30.51M	23.06	22.80
Ensem-4	6L-1024H	40.68M	22.68	22.59
Ensem-4	6L-512H	10.17M	23.18	22.89

We compare the classical and SMCL ensembles in different ensemble sizes and the experiment results on the TIMIT task are shown in Table I and THCHS-30 in Table II. With the ensemble size increasing, both classical and SMCL ensembles obtain performance improvement, which can demonstrate that initializing each DNN with a different random seed is still able to provide combination gains. And SMCL training method for acoustic modeling outperforms classical ensemble learning at different ensemble size. Note that 4 DNNs ensemble of '6L-512H' has the same number of model parameters as the single model of '6L-1024H', and obtains relative 6-8% performance improvement. The acoustic model trained by SMCL can achieve performance improvement by mining the underlying distribution of speech data rather than relying on increasing the model parameters.



(a) Frame Accuracy (b) Loss During Training

Fig. 2. Frame accuracy and loss during training for SMCL and classical ensembles on TIMIT. 'Classical' and 'SMCL' represents training by classical ensemble learning and SMCL respectively.

In order to further make a comparison of training process between SMCL and classical ensemble learning, we analyze the loss and frame accuracy during training in the TIMIT recognition task. Fig. 2(a) shows that the ensembles trained by SMCL result in higher frame accuracy than the classical ensembles. The ensembles trained by SMCL optimize for the joint cross-entropy losses directly, not only arriving at lower loss solutions but also reducing errors more quickly, which is shown in Fig. 2(b). Compared with the improvement obtained in the frame accuracy and loss, the PER/WER improvement is relatively smaller. The acoustic model is the submodule in the ASR task and it is the key for the further speech recognition performance improvement to exploite potentialities of the acoustic models trained by SMCL in the following decoding.

We select 10 classes randomly from 1969 neural network outputs of TIMIT and explore class-wise distribution of the validation set frames assigned to the lowest error model. When the acoustic model is trained by SMCL, the label-space clustering is shown in Fig. 3(a). Most models become experts at classifying certain classes and these assignments don't rely on manual adjustments or pre-initialization in any case, which automatically appear after training by SMCL. In contrast, Fig. 3(d) shows that the class assignments for a standard ensemble are nearly uniform. The level of class division can demonstrate the hypothesis that the phones may be spontaneously aggregated into several clusters due to the intuitive perceptual properties of speech. The phonetics knowledge may help us make better class clustering but it's uncertain whether the manual adjustments could reveal the underlying distribution of the speech data. For the generalization of (6) by changing the constraints to "pick-k-model", we also research the effect of k. By varying k between 1 and the number of

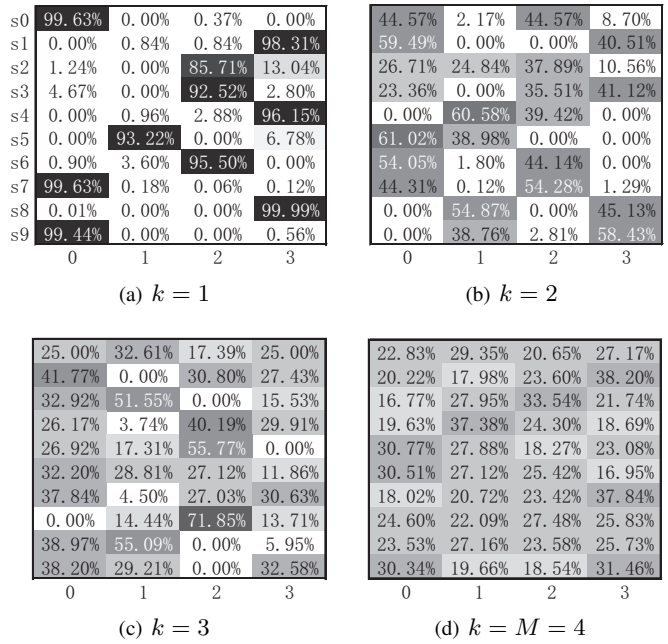


Fig. 3. 's0-s9' represent 10 different tied states that are selected randomly from 1969 neural network outputs on TIMIT experiment. k is a parameter that controls the number of models each example can be assigned to during training. (a-d) show the assignment of validation set examples for various k between 1 and the number of ensemble size M.

ensemble size M, the models transition from minimizing the joint loss at $k = 1$ to a standard ensemble learning at $k = M$. Fig. 3 shows these results and model trained by SMCL with $k = 1$ gets the highest frame accuracy.

V. Conclusions

In this paper, we propose a novel training schedule for DNN-based ensemble acoustic model. In the proposed training schedule, all submodels are jointly trained to cooperatively optimize the loss objective by a Stochastic Multiple Choice Learning approach. It results in that different submodels have specialty capacities for modeling senones with different properties. Systematic experiments show that the proposed model is competitive with the dominated DNN-based acoustic models. The future work will perform Stochastic Multiple Choice Learning for acoustic modeling at the full-sequence level.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (No. 61573357, No. 61503382, No. 61403370, No. 61273267, No. 91120303).

References

[1] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "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] George E Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[3] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, and Gerald Penn, "Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4277–4280.

[4] Tara N Sainath, Abdel-rahman Mohamed, Brian Kingsbury, and Bhuvana Ramabhadran, "Deep convolutional neural networks for lvcsr," in *Acoustics, speech and signal processing (ICASSP), 2013 IEEE international conference on*. IEEE, 2013, pp. 8614–8618.

[5] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, 2013, pp. 6645–6649.

[6] Alex Graves, Navdeep Jaitly, and Abdel Rahman Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *Automatic Speech Recognition and Understanding*, 2013, pp. 273–278.

[7] Zhang Y, Chen G, Yu D, et al. Highway long short-term memory RNNS for distant speech recognition[J]. 2015:5755-5759.

[8] Yuanyuan Zhao, Shuang Xu, and Bo Xu, "Multidimensional residual learning based on recurrent neural networks for acoustic modeling," *Interspeech 2016*, pp. 3419–3423, 2016.

[9] Anders Gustavsson, Anders Magnuson, Bjorn Blomberg, Magnus Andersson, Jonas Halfvarson, and Curt Tysk, "On the difficulty of training recurrent neural networks," *Computer Science*, vol. 52, no. 3, pp. 337–345, 2012.

[10] Jonathan G Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*. IEEE, 1997, pp. 347–354.

[11] L. Deng, J.C. Platt. Ensemble deep learning for speech recognition[J]. Proc Interspeech, 2014.

[12] Xiong W, Droppo J, Huang X, et al. The Microsoft 2016 Conversational Speech Recognition System[J]. 2016.

[13] Saon G, Sercu T, Rennie S, et al. The IBM 2016 English Conversational Telephone Speech Recognition System[J]. 2016.

[14] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig, "Achieving human parity in conversational speech recognition," *arXiv preprint arXiv:1610.05256*, 2016.

[15] Kirchhoff, Katrin, and Jeff A. Bilmes. "Combination and joint training of acoustic classifiers for speech recognition." ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW). 2000

[16] Yoav Freund and Robert E Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 2010.

[17] Tumer K, Ghosh J. Error Correlation and Error Reduction in Ensemble Classifiers[J]. Connection Science, 2015, 8(3-4):385-404.

[18] Yong Liu and Xin Yao, "Ensemble learning via negative correlation," *Neural Networks*, vol. 12, no. 10, pp. 1399–1404, 1999.

[19] Yoav Freund and Robert E Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *European conference on computational learning theory*. Springer, 1995, pp. 23–37.

[20] Ivan Medennikov, Alexey Prudnikov, and Alexander Zatvornitskiy, "Improving english conversational telephone speech recognition," in *INTERSPEECH*, 2016, pp. 2–6.

[21] Jacobs R. Adaptive Mixture of Local Experts[J]. Neural Computation, 1991, 3(1):78-88.

[22] Eigen D, Ranzato M, Sutskever I. Learning Factored Representations in a Deep Mixture of Experts[J]. Computer Science, 2014.

[23] Jun Du, Yan-Hui Tu, Lei Sun, Feng Ma, Hai-Kun Wang, Jia Pan, Cong Liu, Jing-Dong Chen, and Chin-Hui Lee, "The ustc–iflytek system for chime-4 challenge," *Proc. CHiME*, pp. 36–38, 2016.

[24] Kirillov A, Shlezinger D, Vetrov D P, et al. M-Best-Diverse Labelings for Submodular Energies and Beyond[J]. IEEE Transactions on Medical Imaging, 2015, 32(8):1504-1514.

[25] Dhruv Batra, Payman Yadollahpour, Abner Guzman-Rivera, and Gregory Shakhnarovich, "Diverse m-best solutions in markov random fields," in *European Conference on Computer Vision*, 2012, pp. 1–16.

[26] Kirillov A, Schlesinger D, Vetrov D, et al. M-Best-Diverse Labelings for Submodular Energies and Beyond[C]// NIPS. 2015.
[27] A. Guzman-Rivera, D. Batra, and P. Kohli, "Multiple choice learning: Learning to produce multiple structured outputs," *Advances in Neural Information Processing Systems*, vol. 3, pp. 1799–1807, 2012.
[28] Debadeepta Dey, Varun Ramakrishna, Martial Hebert, and J. Andrew Bagnell, "Predicting multiple structured visual interpretations," in *IEEE International Conference on Computer Vision*, 2015, pp. 2947–2955.
[29] Lee S, Purushwalkam S, Cogswell M, et al. Stochastic Multiple Choice Learning for Training Diverse Deep Ensembles[J]. 2016.
[30] J. S Garofolo, L. F Lamel, W. M Fisher, J. G Fiscus, and D. S Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *Nasa Sti/recon Technical Report N*, vol. 93, 1993.
[31] Dong Wang and Xuewei Zhang, "Thchs-30: A free chinese speech corpus," *arXiv preprint arXiv:1512.01882*, 2015.
[32] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.
[33] Frank Seide, Gang Li, Xie Chen, and Dong Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 24–29.
[34] Xavier Glorot, Antoine Bordes, and Yoshua Bengio, "Deep sparse rectifier neural networks.," in *Aistats*, 2011, vol. 15, p. 275.