# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| 題目(和文) | |
| Title(English) | Leveraging Better Training Targets for Deep Neural Network Acoustic Models in Speech Recognition |
| 著者(和文) | PriceRyanWilliam |
| Author(English) | Ryan Price |
| 出典(和文) | 学位:博士（学術）,<br>学位授与機関:東京工業大学,<br>報告番号:甲第10361号,<br>授与年月日:2016年9月20日,<br>学位の種別:課程博士,<br>審査員:篠田 浩一,徳永 健伸,秋山 泰,村田 剛志,藤井 敦 |
| Citation(English) | Degree:,<br>Conferring organization: Tokyo Institute of Technology,<br>Report number:甲第10361号,<br>Conferred date:2016/9/20,<br>Degree Type:Course doctor,<br>Examiner:,,,, |
| 学位種別(和文) | 博士論文 |
| Category(English) | Doctoral Thesis |
| 種別(和文) | 論文要旨 |
| Type(English) | Summary |

（博士課程）
Doctoral Program

# 論 文 要 旨

THESIS SUMMARY

| | | | |
|---|---|---|---|
| 専攻：<br>Department of | Computer Science 専攻 | 申請学位（専攻分野）：<br>Academic Degree Requested | 博士 (Philosophy)<br>Doctor of |
| 学生氏名：<br>Student's Name | Ryan William Price | 指導教員（主）：<br>Academic Advisor(main) | Koichi Shinoda |
| | | 指導教員（副）：<br>Academic Advisor(sub) | |

要旨（英文 800 語程度）
Thesis Summary（approx.800 English Words ）


"Leveraging Better Training Targets for Deep Neural Network Acoustic Models in Speech Recognition" contains five chapters that are described below.

Chapter 1 ("Introduction") introduces Deep Neural Network (DNN) acoustic models as the state-of-the-art acoustic modeling method in automatic speech recognition (ASR). DNN utilize deep and wide hidden layers to provide very accurate acoustic modeling across many different ASR tasks. Despite their success, there are several remaining challenges in DNN acoustic modeling. This thesis focuses on the two challenges of making DNNs faster to evaluate and easier to deploy for online and embedded speech recognition tasks, and data sparsity during speaker adaptation. Soft target training is used to obtain DNNs which are faster-to-evaluate and easier to deploy for online and embedded speech recognition tasks. This is accomplished by using the predictions from an accurate but slow-to-evaluate DNN to provide training targets for a faster-to-evaluate network. To address the problem of data sparsity during adaptation, we propose a hierarchy of output layers which uses knowledge about the phonetic structure of DNN output classes to derive better training targets when adaptation data is very limited.

Chapter 2 ("Deep Neural Network Acoustic Modeling") begins with a formulation of the ASR problem and an overview of the previous acoustic modeling paradigm using Gaussian mixture models. After that we provide a fundamental description of feedforward neural networks before presenting the hybrid approach to acoustic modeling using DNN and hidden Markov models (HMM). Several standard training criteria for DNN acoustic models are described, as well as the process of obtaining the conventional hard training targets used with those training criteria.

Chapter 3 ("Soft Target Training for Fast and Accurate Models") reviews the concept of using a neural network to approximate the function learned by another neural network using soft targets instead of 0/1 labels for training. Soft targets are the vector of class conditional probabilities predicted by a neural network having a softmax activation function at the output layer. Several previous works have explored soft target training for training smaller neural networks or for condensing an ensemble of neural networks into a single neural network. Before reviewing previous work, we describe the use of soft target training for DNN acoustic models and highlight the important property of utilizing unlabeled training data. We propose using a DNN trained with speaker adaptive features which take multiple decoding passes to estimate in order to improve a single-pass, speaker independent DNN. Then we evaluate training a small DNN using the outputs from a much larger DNN and demonstrate the novel result that a small DNN acoustic model trained with soft targets can actually perform equal or better than a much larger DNN acoustic model when a large amount of untranscribed data is available for soft target training. After that, we propose using outputs from a convolutional neural network to train a feedforward neural network for voice activity detection, a new application for soft target training. Following the experiments, analysis is done on several aspects of soft targets, beginning with an interpretation of soft targets that is motivated by an examination of the posterior probabilities output by a large DNN acoustic model used for soft target training. Then the effectiveness of soft targets is compared to using hard targets generated by taking the class having the maximum prediction. After that, we investigate the rank and relative information content of the hidden layer weight matrices and discover that soft target trained networks tend to have matrices with fuller rank and are able to more fully leverage the capacity of large networks when learning the additional information encoded by the soft targets. We also perform some error analysis to compare the correct and incorrect predictions of a soft target trained DNN and the network used to train it. Finally, a quantitative analysis of the computational cost associated with evaluating our soft target trained networks at test time is performed.

Chapter 4 ("Speaker Adaptation of DNN Using a Hierarchy of Output Layers") turns to the problem of speaker adaptation of DNN with very limited amounts of supervised adaptation data. First the data scarcity problem is described in terms of the affect of adaptation using a small amount of data on DNN acoustic models which have very large output layers that correspond to context dependent HMM states. In many cases the adaptation data is so sparse that many of the HMM states modeled in the DNN output layer will not be observed during adaptation and, as a result, recognition performance is likely to

be degraded for those states. We propose a hierarchy of output layers to address this problem. In contrast to the soft target training approach in Chapter 3 which obtained training targets from an DNN which was very accurate for a given task but deemed too expensive-to-evaluate, in this chapter we rely on knowledge about the phonetic structure of DNN output classes to derive better training targets when adaptation data is very limited. We demonstrate the effectiveness of our approach using a mobile voice search task with between 5 to 75 short utterances for adaptation.

Chapter 5 ("Conclusions and Future Work") finishes this thesis with conclusions and future work.

備考 ： 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。

Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).