Title	Speech Recognition Enhanced by Lightly-supervised and Semi-supervised Acoustic Model Training(Abstract_要旨)
Author(s)	Li, Sheng
Citation	Kyoto University (京都大学)
Issue Date	2016-03-23
URL	https://doi.org/10.14989/doctor.k19849
Right	許諾条件により本文は2017-03-01に公開
Type	Thesis or Dissertation
Textversion	ETD

（ 続紙 1 ）

京都大学	博士（情報学）	氏名	李 勝 (Li Sheng)
論文題目	Speech Recognition Enhanced by Lightly-supervised and Semi-supervised Acoustic Model Training （音響モデルの準教師付き及び半教師付き学習による音声認識）		

（論文内容の要旨）

Automatic transcription of lectures is one of the promising applications of automatic speech recognition (ASR), since captions to the lectures are needed not only for hearing-impaired persons but also for non-native viewers and elderly people. ASR is also useful for indexing the content. This work addresses effective acoustic model training targeted on Chinese spoken lectures. ASR of lectures has been investigated for almost a decade in many institutions world-wide, but there are still technically challenging issues for the system to reach a practical level. The biggest challenge is the limitation of training data.

In this work, a relatively small-sized database for Chinese spoken lectures with faithful transcripts is first compiled, but it is not sufficient for supervised training. On the other hand, there is huge amount of audio and video data of lectures with closed caption texts or without any related texts, which should be exploited to increase the training data. This thesis presents a progressive framework for acoustic model training by effectively incorporating speech data without faithful transcripts. Since the automatically generated label with a seed model will have a low accuracy, lightly-supervised training is introduced by leveraging closed caption texts. Then, semi-supervised training is reasonably adopted by incorporating unlabelled data. A novel discriminative approach is proposed to select reliable data in this framework. A dedicated set of classifiers are designed to select or verify the hypothesis from multiple ASR systems or the closed caption text.

Chapter 1 introduces the background, the problem, and the approaches addressed in the thesis. In Chapter 2, a review of speech recognition and deep neural network (DNN)-based acoustic model training is presented, and then the basic concept of lightly-supervised and semi-supervised training in the machine learning paradigm is introduced with related work.

Chapter 3 describes the corpus and the baseline system. For a comprehensive study on ASR of spontaneous Chinese, a corpus of Chinese Lecture Room (CCLR) is compiled. An overview of this corpus and some linguistic analysis are presented. Then, a baseline ASR system is developed based on GMM (Gaussian Mixture Model) and DNN using this corpus.

In Chapter 4, the proposed lightly-supervised acoustic model training with discriminative data selection from closed caption texts is explained. In the proposed method, a sequence of the closed caption text and that of the ASR hypothesis by the baseline system are aligned. Then, a set of dedicated

classifiers based on CRF (Conditional Random Fields) is designed and trained to select the correct one among them or reject both. It is demonstrated that the classifiers can effectively filter the usable data for acoustic model training without tuning any threshold parameters. A significant improvement in the ASR accuracy is achieved from the baseline system (3% absolute) and also in comparison with the conventional method of lightly-supervised training based on simple matching and confidence measure score (CMS).

In Chapter 5, the proposed semi-supervised acoustic model training with discriminative data selection from multiple ASR systems' hypotheses is described. In the proposed method, ASR hypotheses are obtained from complementary GMM and DNN based ASR systems. Then, a set of CRF-based classifiers are trained to select the better hypothesis and verify the selected data. The combined hypothesis for acoustic model training shows higher quality compared with the conventional system combination method (ROVER). Moreover, compared with the conventional data selection based on CMS, the method is demonstrated more effective for filtering usable data. A significant improvement in the ASR accuracy is achieved over the baseline system (1.5% absolute) and in comparison with the models trained with the conventional system combination and data selection methods.

Chapter 6 concludes the thesis with a brief outlook of future work.

注)論文内容の要旨と論文審査の結果の要旨は１頁を３８字×３６行で作成し、合わせて、３,０００字を標準とすること。
論文内容の要旨を英語で記入する場合は、４００～１,１００wordsで作成し
審査結果の要旨は日本語５００～２,０００字程度で作成すること。

（続紙 ２ ）

（論文審査の結果の要旨）

本論文は、講演の自動音声認識の音響モデルの高精度化のために、従来の教師付き学習に必要な書き起こしデータを大規模に用意するのが困難であるという問題に対して、字幕テキストを活用する準教師付き学習、及び書き起こしがないデータを活用する半教師付き学習を定式化・実現した研究をまとめたもので、主な成果は以下の通りである。

1. これまで研究開発が行われていなかった中国語の講演音声のコーパスを設計・構築し、約百講演にアノテーション付与した上で、ディープニューラルネットワーク(DNN)に基づくベースライン音声認識システムを実現した。

2. 字幕テキストを活用する準教師付き学習の新たな方法を提案・実現した。これは、字幕テキストと音声認識結果を対応付けした上で、不一致がある場合にどちらが正しいかを自動的に選別する識別器を構成・適用するものである。従来の単純な文字マッチングに基づく手法と比べてより多くの有用なデータを得ることができ、従来の音声認識結果の信頼度に基づく手法と比べてしきい値の調整をする必要がない。評価実験の結果、これらの手法より高い認識率の改善を得ることができ、ベースライン音声認識からは３ポイント改善した。

3. 書き起こしがないデータを活用する半教師付き学習の新たな方法を提案・実現した。これは、複数の音声認識システムの結果（仮説）を対応付けした上で、それらを組み合わせて選別する識別器を構成・適用するものである。複数の仮説から選択を行う識別器と、仮説が正しいか検証を行う識別器を各々構成し、段階的に適用することで、認識結果仮説の改善と選別を実現する。従来の多数決に基づく仮説組合せ手法と比べて仮説の精度が改善し、従来の音声認識結果の信頼度に基づく手法と比べて選別の精度が改善することを確認した。評価実験の結果、これらの手法より高い認識率の改善を得ることができ、前記の音声認識からさらに１．５ポイントの認識率の改善を実現した。

以上のように本論文は、大規模に集積が進む講演音声アーカイブを効率的に活用することで、音声認識の高精度化を実現する方法を提案したもので、学術上・実用上寄与するところが少なくない。よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、平成28年 2月24日に論文とそれに関連した内容に関する口頭試問を行った結果、合格と認めた。

注)論文審査の結果の要旨の結句には、学位論文の審査についての認定を明記すること。更に、試問の結果の要旨（例えば「平成　年　月　日論文内容とそれに関連した口頭試問を行った結果合格と認めた。」）を付け加えること。

Webでの即日公開を希望しない場合は、以下に公開可能とする日付を記入すること。
要旨公開可能日：　　年　　月　　日以降