基於深度聲學模型其狀態精確度最大化之強健語音特徵擷取的初步研究

The preliminary study of robust speech feature extraction based on maximizing the accuracy of states in deep acoustic models

張立家　Li-chia Chang
國立暨南國際大學電機工程學系
Department of Electrical Engineering
National Chi Nan University
s108323518@mail1.ncnu.edu.tw

洪志偉　Jeih-weih Hung
國立暨南國際大學電機工程學系
Department of Electrical Engineering
National Chi Nan University
jwhung@ncnu.edu.tw

摘要

在本研究中，我們提出一種新穎的強健性語音特徵擷取技術，以增進雜訊干擾環境下的語音辨識效能。此新技術，利用語音辨識系統中後端的原聲學模型所提供的資訊，在不重新訓練聲學模型的前提下，藉由深度類神經網路架構，學習得到最大化聲學模型狀態之精確度對應之語音特徵，進而使此語音特徵擁有對雜訊的強健性，相較於其他改善聲學模型以達到雜訊強健性的技術，本研究所提出的新技術具有計算量小且訓練快的優點。

在初步實驗中，我們使用了 TIMIT 此中型語料庫來加以評估，實驗結果顯示所提之新語音特徵擷取法，相對於基礎實驗，能有效地降低各種雜訊種類與雜訊程度之環境下語音的音素錯誤率，凸顯此方法的效能及發展價值。

關鍵詞：雜訊強健性之語音特徵、語音辨識、深度學習

Abstract

In this study, we focus on developing a novel noise-robust speech feature extraction technique to achieve noise-robust speech recognition, which employs the information from the backend acoustic models. Without further retraining and adapting the backend acoustic models, we use deep neural networks to learn the front-end acoustic speech feature representation that can achieve the maximum state accuracy obtained from the original acoustic models. Compared with the robustness methods that retrain or adapt acoustic models, the presented method exhibits the advantages of lower computational complexity and faster training.

In the preliminary evaluation experiments conducted with the median-vocabulary TIMIT database and task, we show that the newly presented method achieves lower word error rates in recognition under various noise types and levels compared with the baseline results. Therefore, this method is quite promising and worth developing further.

Keywords: noise-robust speech feature, speech recognition, deep learning