NSYSU+CHT 團隊於 2020 遠場語者驗證比賽之語者驗證系統

NSYSU+CHT Speaker Verification System for Far-Field Speaker Verification Challenge 2020

張育嘉 Yu-Jia Zhang, 陳嘉平 Chia-Ping Chen
國立中山大學資訊工程學系
Department of Computer Science and Engineering
National Sun Yat-sen University
M083040025@student.nsysu.edu.tw, cpchen@mail.cse.nsysu.edu.tw,
蕭善文 Shan-Wen Hsiao, 詹博丞 Bo-Cheng Chan, 呂仲理 Chung-li Lu
中華電信研究院
Chunghwa Telecom Laboratories, Taoyuan, Taiwan
swhsiao@cht.com.tw, cbc@cht.com.tw, chungli@cht.com.tw

摘要

在本論文中，我們描述了 NSYSU+CHT 團隊在 2020 遠場語者驗證比賽 (2020 Far-field Speaker Verification Challenge, FFSVC 2020) 中所實作的系統。單一系統採用基於嵌入的語者識別系統。該系統的前端特徵提取器是結合了時延神經網路，與卷積神經網路模組兩者的優點，稱為時延殘差神經網路的架構。在池化層，我們實驗了不同方式：統計池化層和 GhostVLAD。而後端的評分器則採用機率線性判別分析，我們訓練跟調適機率線性判別分析用以不同系統的融合。我們分別參加了 FFSVC 2020 採單一麥克風陣列資料的文本相關（任務一）與文本無關（任務二）的語者驗證任務。我們提出的系統在任務一上取得 minDCF 0.7703，EER 9.94%，在任務二上則是 minDCF 0.8762，EER 10.31%。

Abstract

In this paper, we describe the system Team NSYSU+CHT has implemented for the 2020 Far-field Speaker Verification Challenge (FFSVC 2020). The single systems are embedding-based neural speaker recognition systems. The front-end feature extractor is a neural network architecture based on TDNN and CNN modules, called TDResNet, which combines the advantages of both TDNN and CNN. In the pooling layer, we experimented with different methods such as statistics pooling and GhostVLAD. The

back-end is a PLDA scorer. Here we evaluate PLDA training/adaptation and use it for system fusion. We participate in the text-dependent(Task 1) and text-independent(Task 2) speaker verification tasks on single microphone array data of FFSVC 2020. The best performance we have achieved with the proposed methods are minDCF 0.7703, EER 9.94% on Task 1, and minDCF 0.8762, EER 10.31% on Task 2.

關鍵詞：遠場語者驗證、時延神經網路、卷積神經網路、時延殘差神經網路、GhostVLAD

Keywords：Speaker Verification, TDNN, CNN, TDResNet, GhostVLAD