Task-aware Deep Bottleneck Features for Spoken Language Identification

Bing Jiang¹, Yan Song¹, Si Wei², Ian McLoughlin¹, Li-Rong Dai¹
1 University of Science and Technology of China
2 iFlytek Research, Anhui USTC iFlytek Co., Ltd.

Introduction

Recently, Deep bottleneck features (DBF) have been applied for language identification (LID), and yield significant performance improvement over state-of-the-art methods.

However, the DBF extractor is trained using a large corpus of specific language which is not directly related to the LID task.

In this paper, we propose to tune the post-trained DNN parameters using an LID-specific training corpus, which may make the resulting DBF, termed a Discriminative DBF (D²BF), more discriminative and task-aware.

Discriminative Training for DBF

We aim to add language discriminative information into the corresponding extractor by tuning the parameters of the bottleneck layer.

Motivation

- The ability of the DBF for predicting phonemes or phoneme states is important. And the updated extractor should share these capabilities, but provide a compromise between phoneme and language information.
- The DBF is a non-linear transformation of the input feature. Tuning all parameters would be complex and unrealistic.
- The task is linear when confined to the parameters of the bottleneck layer.

Advantages of DBF

- Non-linear transformation
- Reduce the redundancy of input features
- Reflect the relevant class label information
- Effectively
- More robust to variations that are irrelevant to phonemes or phoneme states, including effects of different speakers, channels and background noise

Bottleneck Feature

DBF are generated from a DNN in which one of the internal layers has a small number of hidden units, termed the bottleneck layer. Typically, the DNN used to extract DBF is trained to predict the phonemes or phoneme states.

After the DNN is trained well, we remove the layers above the bottleneck layer and extract the corresponding DBF as follows

\[ y = f(x; 	heta) \]
\[ y_n = f(h_n) = W^t h_n \]

where

\[ y = [y_1, y_2, \ldots, y_R] \]
\[ y_n = [y_{n1}, y_{n2}, \ldots, y_{Rn}] \]

Advantages of DBF

- Non-linear transformation
- Reduce the redundancy of input features
- Reflect the relevant class label information
- Effectively
- More robust to variations that are irrelevant to phonemes or phoneme states, including effects of different speakers, channels and background noise

Experiments

The six most confused language from NIST LRE 2009 are selected to evaluate the performance: Dari, Farsi, Russian, Ukrainian, Hindi and Urdu. Three experiments are conducted as follows:

Exper1. A GMM-MMI using untuned DBF against the proposed GMM-UBM with D²BF.

Exper2. The GMM-UBM with D²BF against a GMM-MMI using the D²BF.

Exper3. A TV system with a 400-dimensional i-vector using the DBF against the proposed D²BF.

WB and OB denote the system with or without a gaussian score backend.