Text-Independent Speaker Identification

Lidan TAN[1], **Ruiling XI**[2], **Feihong DONG**[3], **Yang QIAN**[4]

[1,2]Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China
[3]Shenzhen University of Technology, Southern University of Science and Technology, Shenzhen, China
[4]Shenzhen Institute of Advanced Technology, Chinese Academy of Science, Shenzhen, China

Abstract

Abstract Speaker identification is the task of classifying unknown segments of speech into a set of known classes that have been learned by the classifier. Text-independent speaker identification involves classifying the speaker based on their vocal characteristics alone, regardless of the content of their speech. The goal of this work is to train a model that performs speaker identification task on the given LibriSpeech-SI dataset. This work implements our own data pre-processing procedure on flac files as well as to use a hierarchical network of experts to perform the classification.

Keywords: *Neural Network; Speaker Identification*

1. Introduction

The task of speaker identification, or speaker recognition, is a multi-class classification problem that seeks to determine the identity of the speaker of a piece of audio [1]. Speaker identification broken down into text-dependent approaches and text-independent approaches based on whether the content of the speech is considered [1]. Speaker identification can be further subdivided into closed-set approaches and open-set approach. Namely, the closed-set approaches are to identify a speaker from a known set of speakers that the classifier is previously trained on, while the open-set approaches are to decide whether the speaker of the interested piece of audio belongs to the known speakers[2].

Audio pieces are segments of time series data that contain information about the pitch, accent, and speaking style of a person's voice [3]. However, these subjective properties are not easily understood by machine learning models. In addition, speech data may contain periods of silence or noise that are not useful for classification purposes. Therefore, pre-processing of speech signals is crucial for speaker identification.

The main contributions of this work are as follows:

To isolate the speaker's local acoustic features at each time point, the first step in the data preparation stage is to "slice" the data, or split the time series speech data into slices of uniform length. Then, we get rid of the quiet sections in the data slices since they don't add anything to the speaker identification process and don't tell us anything about the speaker's voice. Finally, we utilize the Fourier transform on each individual data slice to get the signal energy distribution at each time instant. We hypothesize that the variation in signal strength between speakers might provide important information about their personalities.

After feature extraction, this work implements a "network of experts" hierarchical structure to train on the extracted features [3]. The hierarchical structure consists of a "backbone" and "experts", which are feed-forward neural networks with the same structure. The data is first divided into several specialties by the backbone network, each specialty containing several mutually exclusive data classes. Then, the data is classified into corresponding classes by the corresponding specialty. Compared with using a single neural network to classify 250 classes, this hierarchical structure significantly improves the accuracy of classification.

2. Methodology

In this section, we will provide a detailed description of our technical methods. Section 2.1 will introduce the data preprocessing methods used in this work, as well as the dimensionality reduction and classification of the data during the preprocessing process. Section 2.2 will introduce the model architecture and corresponding training methods used in this work.

2.1. Data Pre-Processing

The data pre-processing contains slicing, Fourier transform operation, t-SNE (t-distributed stochastic neighbor embedding)[4] and k-Means cluster.

We firstly slice the total flac file into multiple slices, each slices contains 0.2s of information. In the Fourier transform operation as shown in Eq.1, we focus on the frequency information between 50 and 2000 Hz, as this scope contains normal human vocal frequency. Additionally, the silent slices should be removed as they cannot provide speaker information. In practice, if most of the power are in the bottom half frequencies, we regard this slice as silent slice.

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i x \xi} dx \tag{1}$$

After getting Fourier transformed information, which is 400 dimensional for each slice, we implement the t-SNE dimension reduction to get a 2D output. The t-SNE dimension reduction computes the probabilities $p_{i|j}$ that arre proportional to the similarity of objects x_i and x_j as shown in Eq.2, where N represents the number of original dimensions, and KL indicates the Kullback–Leibler divergence. The location of each point is determined by minimizing the KL value though gradient descent. The reduced results are used for k-Means cluster, since we need to get the specialty of each speaker first. The specialty result is shown as below.

$$\begin{aligned} p_{i|j} &= \frac{exp(-|x_i - x_j|^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-|x_i - x_k|^2/2\sigma_i^2)} \\ p_{ij} &= \frac{p_{i|j} + p_{j|i}}{2N} \\ KL(P \parallel Q) &= \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \end{aligned} \tag{2}$$

The t-SNE result contains 2 dimensions for each vectors, and they are used for further knn cluster. Firstly, we set the specialty number to 16, as $15^2 < 250 < 16^2$. The cluster result is shown in Figure 1, as the red dots indicate the centroid of each cluster.

Finally, in order to fully utilize the Fourier transform result, we feed the 400 dimension data into the neural network, both expert and backbone models.

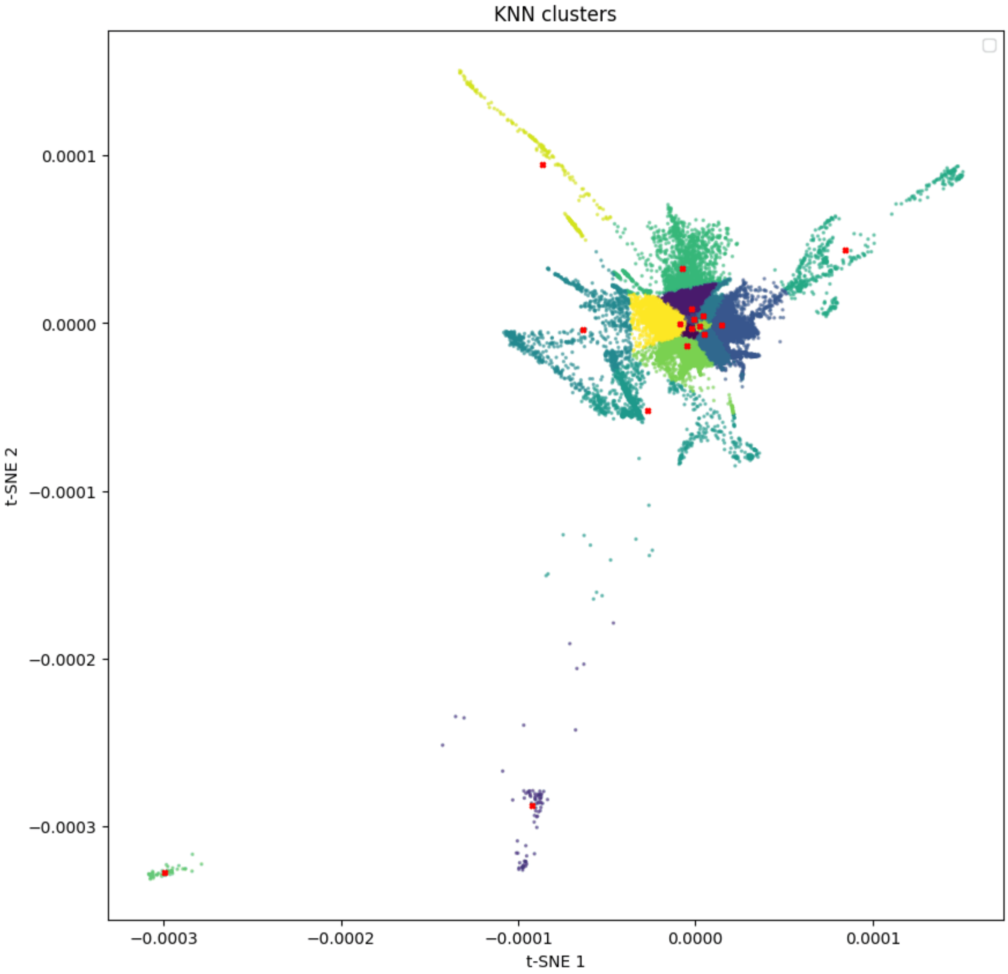

Figure 1. The visualization of knn cluster result, based on t-SNE output. The data belongs to the same cluster shares the same color, and the red dots indicates the 16 different centroids.

2.2. Network of Experts

The architecture of the network adopts the Network of Experts framework proposed by Ahmed in 2016 for large-scale image classification problems [3]. Its main idea is to classify a large number of categories into a smaller number of specialties, and then classify them in each specialty, thereby decomposing the complex large-scale classification problem into smaller multiple classification problems to improve classification performance. The architecture of the experts network is shown in Figure 2.

The Network of Experts for Large-Scale Image Categorization is a hierarchical architecture designed to classify large numbers of categories by dividing them into smaller groups called specialties, and then classifying them within each specialty. In this methodology, the data is first processed through dimensionality reduction using t-SNE and clustering with k-Means to create 16 specialties. Each of the 250 speaker classes is assigned to one and only one specialty.

A backbone model and specialist models are trained using the same architecture, which is a simple feed-forward neural network with 7 fully connected layers and 400 nodes per layer, except for the output layer, which has a number of nodes equal to the number of categories in the classification target. The fully connected layers are activated by the tanh function.

The training process for this architecture is as follows: first, the backbone model is trained using the spectral features of the data slices as features and the specialty labels as labels. Next, each specialist is trained using the spectral features of the data slices within its specialty as features and the class labels as labels. Once the backbone and specialists have reached a certain level of performance, they are connected and fine-tuned with all of the spectral features of the data slices. During the fine-tuning process, the backbone model first classifies all of the data slices. Then, the final specialty prediction for each sample (i.e., each person's recording) is determined by a majority vote of the classifications of all of the slices within the sample. The predicted specialist then classifies all of the slices within the sample and the final class label is determined by a majority vote of the specialist's classifications. Finally, the entire network is fine-tuned to improve the performance of both the backbone and specialists.

The prediction process of the architecture is similar to the fine tuning process, only omitting the process of back-propagation and adjusting parameters.

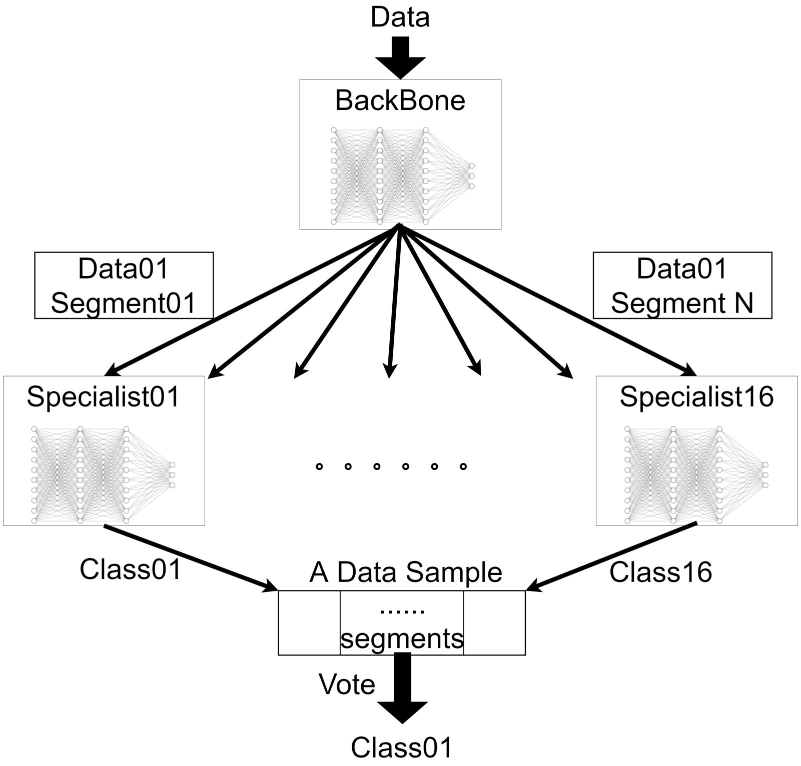

Figure 2. The visualization of the Network of Experts architecture.

3. Result

This work implements all data preprocessing, data dimensionality reduction, data classification, neural network model implementation, dataset encapsulation, training of the standalone backbone network and specialist network end-to-end, training of the complete network of experts architecture end-to-end, validation, and prediction tasks.

Table 1. The classification performance based on each specialty.

specialty No.	0	1	3	9	12	14	15
No. of speakers	26	62	7	1	1	151	2
Accuracy	70.72%	54.80%	80.57%	100.00%	100.00%	53.52%	100.00%

Among them, if the standalone backbone network is used to directly predict the class label of the 250-class speech data, its accuracy is 50.14%. Using the backbone network to classify speech data into 16 specialties, its accuracy is 84.61%. Among the 16 specialties, 9 specialties do not contain any classes. In the 7 specialties containing different numbers of categories, the prediction accuracies are 70.72%, 54.80%, 80.57%, 100.00%, 100.00%, 53.52%, and 100.00%, respectively, as shown in Table. 1. On the 0.2-second data slices, the complete network of experts architecture has a prediction accuracy of 0.6058%. If voting is used with all slices of an audio data, the prediction accuracy will be higher, but due to time constraints, we were not able to validate the experiment with all slices of an audio data.

4. Conclusion

Based on the experimental findings presented in Section 4, this study successfully completed the speaker identity classification job by using the spectral properties of voice data slices as training features. The approach enhanced the accuracy of speaker identification by transforming the 250-class classification issue into numerous smaller classification problems using the network of experts architecture.

The division of labor for this work is as follows:

Lidan TAN (12232423) is responsible for reviewing, selecting the model architecture, specifying the division of labor, implementing the model, integrating the training and testing code, running the model's training and testing, and writing the experimental report.

Ruiling XI (12232438) is responsible for data preprocessing, data dimensionality reduction, data clustering, the selection of dimensionality reduction and clustering methods, and writing the experimental report.

Feihong DONG (12233220) is responsible for implementing the training and testing code for the model.

Qian YANG (12233305) is responsible for implementing the training and testing code for the model.

References

[1] R. Jahangir, Y. W. Teh, H. F. Nweke, G. Mujtaba, M. A. Al-Garadi, and I. Ali, "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges," *Expert Systems with Applications*, vol. 171, p. 114591, 2021.

[2] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE signal processing magazine*, vol. 11, no. 4, pp. 18–32, 1994.

[3] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[4] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.