MODELLING OUTPUT PROBABILITY DISTRIBUTIONS FOR ENHANCING SPEAKER RECOGNITION

Jason Pelecanos     Sridha Sridharan

Speech Research Lab, RCSAVT
Queensland University of Technology
GPO Box 2434, George St, Brisbane, Australia
j.pelecanos@qut.edu.au
s.sridharan@qut.edu.au

ABSTRACT

This paper discusses the use of a secondary likelihood classifier scheme for improving speaker recognition performance. The system models the output likelihoods of a typical Gaussian Mixture Model system across multiple speakers. The Output Probability Distributions (OPD) of the primary classifiers contain information on inter-speaker relationships, and are modelled by secondary classifiers to improve recognition accuracies. A comparison of the OPD system with the traditional likelihood ratio and maximum likelihood scoring schemes for verification and identification is performed. Fusion of traditional measures with OPDs is shown to enhance overall recognition performance.

Keywords: Speaker Recognition, Output Probability Distribution, Secondary Classification.

1. INTRODUCTION

Speaker Recognition is the process of recognising speakers by their voice. Speaker recognition can be separated into two main tasks: Speaker Identification and Verification. Closed-Set speaker identification has the requirement of nominating a speaker with the highest likelihood from a given set of speakers. In speaker verification, the claimed identity of the test speaker must be accepted or rejected. Decisions are usually derived from a speaker confidence measure. Successful speaker verification techniques use the likelihood ratio (for score normalisation) [1, 2] based on speaker and background speaker sets.

An alternative to score normalisation of this form, is the use of discriminative training techniques. The goal of discriminative training is to estimate model parameters that minimise classification errors in the training data. Speaker discriminative training techniques such as Neural Networks [3] and discriminative GMMs [4] have been trialed to perform speaker discrimination using a set of primary input features. In this paper we propose a new technique based on what is called Output Probability Distributions for enhancing existing likelihood score based speaker recognition systems.

2. OUTPUT PROBABILITY DISTRIBUTIONS

2.1 Overview

The term Output Probability Distributions (OPDs) [5], in the context of speaker recognition, refers to the distribution of the likelihood scores of the primary classifiers of a speaker set, when tested against speech from a given speaker. The key idea behind our technique is to use a secondary classifier to learn the OPD. ie; the output likelihoods of a primary classification scheme. When the secondary classifier is applied such that it learns the multi-dimensional likelihood distributions of multiple competing speakers, the system can encompass the inter-speaker relationships of the speaker set. In a speaker identification context, the competing speakers are all speakers in the database. For speaker verification, competing speakers are the background speaker models of the target speaker.

A single OPD vector is obtained from the likelihood scores generated by comparing a single feature vector against a series of speaker models. These models may include a target speaker model and a list of background models (for speaker verification), or a general selection of speaker models (for closed-set speaker identification).

OPD models are formed when trained primary models are tested with validation speech, and these likelihood vectors are used to train the secondary models. During the testing phase, the OPD vectors arising from test speech applied to the primary
classifiers, are applied to the secondary model to derive a likelihood measure. For speaker identification, the segment based likelihood measures derived from the secondary classifiers can be fused with the corresponding speaker scores obtained from the primary classifiers. For speaker verification, the likelihood ratio calculated from the background speaker modelling process is merged with the OPD system score estimates.

2.2 OPD Feature Vectors

In the experiments, the primary models for speaker recognition are Gaussian Mixture Models (GMMs) [6]. The likelihood of an observation, \( \tilde{X}_r \), given an \( N \) mixture GMM, \( \lambda \), can be determined:

\[
p(\tilde{X}_r | \lambda) = \sum_{j=1}^{N} p_j g(\tilde{X}_r, \bar{\mu}_j, \Sigma_j)
\]

A GMM comprises the additive contribution of \( N \) multi-dimensional Gaussians with mixture weights \( p_j \), means \( \bar{\mu}_j \), and covariances \( \Sigma_j \). Typically, only the diagonal components of the covariance matrices are used. Model estimation is performed using the Expectation-Maximisation algorithm.

For closed-set speaker ID consisting of \( M \) speakers, the OPD feature vector consists of a set of likelihood scores for each parameterised frame of speech, \( \tilde{X}_r \). This is obtained by comparing the feature vector \( \tilde{X}_r \), against each speaker’s model \( \lambda_1, \lambda_2, ..., \lambda_M \).

\[
O\tilde{P}D(\tilde{X}_r) = \begin{bmatrix} p(\tilde{X}_r | \lambda_1) \\ p(\tilde{X}_r | \lambda_2) \\ \vdots \\ p(\tilde{X}_r | \lambda_M) \end{bmatrix}
\]

(2.2)

For speaker verification, there is a target speaker model \( \lambda_t \), and \( B \) background speaker models \( \lambda_{b_1}, \lambda_{b_2}, ..., \lambda_{b_B} \). Given feature vector \( \tilde{X}_r \), with a target and background speaker models, the OPD features can be calculated as:

\[
O\tilde{P}D(\tilde{X}_r) = \begin{bmatrix} p(\tilde{X}_r | \lambda_t) \\ p(\tilde{X}_r | \lambda_{b_1}) \\ \vdots \\ p(\tilde{X}_r | \lambda_{b_B}) \end{bmatrix}
\]

(2.3)

The target speaker model test is the first entry in the OPD feature vector with the background speaker scores following.

2.3 OPD Modelling and Classification

Once the OPD vectors are obtained, a secondary Gaussian Mixture Model is trained on these vectors. The testing or classification process requires the OPD features to be generated from the primary classifier and then tested against the secondary GMM scheme to arrive at a series of likelihood scores in a similar fashion to that of the GMM primary classifier. These primary and secondary speech utterance scores should be time normalised. Thus, the time-normalised utterance likelihood for \( T \) observations becomes:

\[
\log p(X | \lambda) = \frac{1}{T} \sum_{i=1}^{T} \log p(\tilde{X}_i | \lambda)
\]

(2.4)

In speaker identification, the OPD segment scores are fused with the primary classifier likelihood scores. Speaker verification OPD merit figures are fused with the likelihood ratio estimates from the primary classifier.

2.4 OPD Inter-Speaker Shape Attributes

Shown in Figure 1, are examples of the OPD mean features for two different test speech utterances for speaker identification. The statistics for two speakers are indicated. It is observed that there are unique attributes for the set of average speaker scores particular to each speaker.

![Figure 1. Speaker Identification mean OPD Scores for Speakers one (1) and two (2)](image-url)
Relatively large likelihoods for the first and second speaker positions of the corresponding speaker’s OPDs are expected. By the same token, the non-target speaker models would be generally lower in primary classifier likelihood score. Secondary classification based on OPDs utilise shape information to check that characteristics of the primary classifier speaker scores are present in a given speech segment.

For speaker verification, we found that the shape of the scores for a target model and corresponding background speakers were relatively consistent across target speech segments. When a suitable background speaker set comprising maximally-spaced-close and maximally-spaced-far speakers [6] is selected, the nature of OPDs can be exploited. Certain background speakers will trigger as a result of a close impostor match, and the OPD classifier is able to discern this situation. The OPD scheme is able to determine the existence of non-characteristic attributes in a test utterance. The OPD method can be used to evaluate the suitability of the likelihood scores and reject any utterances that do not conform to the type of conditions that occurred for the training speech.

3. SYSTEM IMPLEMENTATION

The speaker verification system structure used in the experiments is detailed in Figure 2. This system can be adapted for speaker identification. A notable difference is that for verification, background speaker models are used instead of other potential target speaker models.

The speaker verification system uses a primary classifier likelihood ratio test. The log-likelihood ratio, given a target speaker model and a set of B background speaker models is calculated as follows:

$$\log(LR) = \log p(X | \lambda_t) - \log \left\{ \frac{1}{B} \sum_{b=1}^{B} p(X | \lambda_b) \right\} \quad (3.1)$$

The fusion process is a simple linear opinion pool fusion of the scores provided by the likelihood ratio and the OPD system. The linear opinion pool fusion score $S$, with weighting $\alpha$, is calculated for two likelihood sources $LR$ and $OPD$ as shown:

$$S = \alpha (LR) + (1 - \alpha) (OPD), \quad 0 \leq \alpha \leq 1 \quad (3.2)$$

4. EXPERIMENTS

Separate experiments were performed for the speaker identification and verification scenarios.

The small-set speaker identification experiment was tested on an excerpt of 19 speakers from the King Wideband Speech Corpus. The first session was used for primary model development, the second for the OPD model, and the remaining 8 sessions were integrated for testing clusters of four second speech segments. The standard benchmark system used the first two sessions for model training, and eight sessions for testing.

The verification experiment was performed on the NIST 96 Speaker Recognition Database. Two of the one minute same-session training recordings and the three second test utterances were utilised for performance evaluation. For the OPD system, the two one minute sessions were each split into primary classifier training and testing speech. The benchmark system utilised both single minute sessions for GMM training. Background speaker models were developed from the evaluation non-speaker/impostor set, and target speaker models were extracted from the development database.

The parameterisation scheme used 15 dimensional MFCCs with pre-emphasis and liftering. No channel or handset compensation techniques were implemented for this evaluation.
5. RESULTS

The results obtained for the OPD fused system in contrast with the standard recognition systems showed improvement. The stand-alone OPD technique did not surpass the overall performance of the standard GMM systems. Speaker identification performance using basic OPDs achieved an accuracy of 65.3% compared to a base system performance of 76.8%. However, the fused system achieved a 79.6% performance, i.e., an absolute improvement of around 3%. The benefit of using this method for speaker identification is limited to small speaker set scenarios.

When the basic OPD system was used for Speaker verification, results improved over the basic likelihood ratio system for a low false alarm probability only. However, the fused system was able to utilise the strengths of both systems to improve the Detection Error Trade-off (DET) curve verification statistics (see Figure 3).

6. CONCLUSION

Both speaker identification and verification systems can be improved with the fusion of speaker likelihood shape information provided by an OPD. It is also evident for speaker verification, that secondary classification based on OPDs outperforms standard primary classification using Gaussian Mixture Models, for a low false alarm rate. Fusion of these two systems can incorporate the benefits of both systems to improve the DET curve operating range characteristics.

For speaker identification, the incorporation of OPD scores indicates the benefit of utilising certain shape information provided by implementing a secondary classifier to learn the primary classifier likelihood scores. An interesting aspect of using a secondary classifier in this fashion is its’ ability to evaluate and identify abnormal test sessions and environments that may affect the practical performance of these biometric systems. The multi-speaker likelihood shape criteria is useful for identifying such situations.

7. FUTURE WORK

Our research has indicated improved speaker recognition performance by the use of a target and non-speaker likelihoods embedded in the OPD. Future investigations will examine other transformations for OPDs to achieve improved performance across a larger range of the DET curve.

8. ACKNOWLEDGEMENTS

This work was supported by a research contract from Australian Defence Science and Technology Organisation (DSTO).

9. REFERENCES