USING CLUSTERING COMPARISON MEASURES FOR SPEAKER RECOGNITION

Jia Min Karen Kua1, 2, Julien Epps1, 2, Mohaddeseh Nosratighods1, Eliathamby Ambikairajah1, 2, Eric Choi2

1School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, NSW 2052, Australia
2ATP Research Laboratory, National ICT Australia (NICTA), Eveleigh 2015, Australia
jmkua@student.unsw.edu.au, j.epps@unsw.edu.au, hadis@unsw.edu.au, ambi@ee.unsw.edu.au, eric.choi@nicta.com.au

ABSTRACT

Recent results seem to cast some doubt over the assumption that improvements in fused recognition accuracy for speaker recognition systems based on different acoustic features are due mainly to the different origins of the features (e.g. magnitude, phase, modulation information). In this study, we utilize clustering comparison measures to investigate acoustic and speaker modelling aspects of the speaker recognition task separately and demonstrate that front-end diversity can be achieved purely through different ‘partitioning’ of the acoustic space. Further, features that exhibit good ‘stability’ with respect to repeated clustering are shown to also give good EER performance in speaker recognition. This has implications for feature choice, fusion of systems employing different features, and for UBM data selection. A method for the latter problem is presented that gives up to an 11% relative reduction in EER using only 20-30% of the usual UBM training data set.

Index Terms— speaker recognition, normalised mutual information, normalised information distance, UBM training

1. INTRODUCTION

In speaker recognition research, a ‘good’ feature is one that should give good acoustic discrimination for consistent GMM modelling and good speaker discrimination. In the UBM-GMM paradigm, it is clear that the UBM performs an implicit (soft) acoustic ‘partitioning’, while the MAP adaptation provides speaker discrimination, however the relative importance of the acoustic and speaker discrimination in the speaker recognition problem is not very well understood to date. Motivated by IBM’s phonetically inspired UBM [5] and Loquendo/Politecnico di Torino’s Phonetic GMM [6], we attempted to separate acoustic modelling (i.e. characterization of different phonetic content using Gaussian mixtures) and speaker modelling. However experiments conducted have so far been unsuccessful (unpublished) and suggest that acoustic and speaker modelling are difficult to decouple in the current GMM paradigm.

Many speaker recognition researchers, including our group, have been motivated to investigate features derived from different sources of information in speech (e.g. frequency, phase, modulation energy), with the assumption that systems built on these features will model different aspects of the speaker voices and will fuse well with mel-frequency cepstral coefficients (MFCCs) (of course there are multiple criteria for good features [7]). Our recent work, however, seems to indicate that front-end diversity can be achieved purely through different partitioning of the acoustic space, in the sense that a GMM can be considered to be a soft partitioning of the space [1]. In this paper, we utilize clustering comparison measures to shed light on the relative contributions of the acoustic and speaker modelling ‘stages’ and the benefits brought by fusing different acoustic features.

2. CLUSTERING COMPARISON MEASURES

In this paper, information-theoretic based measures for clustering comparison, in particular the Normalised Mutual Information (NMI) [8] and its complement, the Normalised Information Distance (NID) [9] are employed. Compared with other measures, they are normalized, with a range of [0, 1], have a strong mathematical foundation and the NID has the advantage of being a metric. The normalization property facilitates comparison across different features. NMI is a measurement of agreement between alternative data (hard) partitions that can be used even when considering partitions with different numbers of clusters. Let \( S \) be a set of \( N \) data points \( \{s_1, s_2, ..., s_N\} \). Given two clusterings of \( S \), namely \( U = \{U_1, U_2, ..., U_R\} \) with \( R \) clusters, and \( V = \{V_1, V_2, ..., V_C\} \) with \( C \) clusters:

\[
\bigcap_{i=1}^{R} U_i = \bigcap_{j=1}^{C} V_j = \phi, \bigcup_{i=1}^{R} U_i = \bigcup_{j=1}^{C} V_j = S,
\]

the information on cluster overlap between \( U \) and \( V \) can be summarized in the form of an \( R \times C \) contingency table

\[
M = \begin{bmatrix}
    n_{ij} \end{bmatrix}_{i=1...R, j=1...C}
\]

where \( n_{ij} \) denotes the number of data points.
Table 1 NID between UBM of fused systems on NIST 2004 female and system EER on NIST 2006 core condition (512-mixtures UBM)

<table>
<thead>
<tr>
<th>FEATURES</th>
<th>EER (%)</th>
<th>FEATURE-FUSED MFCC NID (UBM)</th>
<th>EER FUSED WITH MFCC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>6.66</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MFCCDropCep1 [1]</td>
<td>6.90</td>
<td>0.47</td>
<td>6.31</td>
</tr>
<tr>
<td>LogLSGD [4]</td>
<td>8.12</td>
<td>0.65</td>
<td>6.04</td>
</tr>
<tr>
<td>SCF [3]</td>
<td>7.82</td>
<td>0.73</td>
<td>5.87</td>
</tr>
<tr>
<td>PLP</td>
<td>7.45</td>
<td>0.71</td>
<td>5.87</td>
</tr>
<tr>
<td>LPCC</td>
<td>6.74</td>
<td>0.78</td>
<td>5.72</td>
</tr>
</tbody>
</table>

In this section, we investigate if different features have similar or different clustering in the acoustic space modelling, for systems fused with MFCC based systems. The features are MFCC, Linear Prediction Cepstrum Coefficient (LPCC), perceptual linear prediction coefficients (PLP), spectral centroid frequency (SCF) [4], log compressed least square group delay (LogLSGD) [3] and dropping the first cepstral coefficient (C1) of MFCC (MFCCDropeCep1) [1]. The first three features were chosen as they are the typical features of NIST SRE consortium submissions, followed by phase-based, frequency-based and MFCC variant features.

The evaluation data in this paper were the core test condition (1con4w-1con4w) of the NIST 2006 SRE where 51 448 trials were tested, which included 3612 true trials and 47 836 false trials. The background data consisted of 3079 speech utterances from the NIST 2004 SRE, which covers a number of speakers (1849 female and 1230 male).

The front-end of the recognition system included an energy-based speech detector which was applied to discard silence and noise frames. The back-end of all systems evaluated on the 1con4w-1con4w database was based on the GMM-SVM technique, with GMM supervector kernels. The Nuisance Attribute Projection (NAP) [10] training data included approximately 10 000 speech utterances from the NIST 2004 and 2005 corpora. The training data in the NIST 2004 and NIST 2005 corpora were used for training cohort models in ZNorm and Tnorm score normalization respectively.

As shown in Table 1, we observed that the ‘further’ (more different) the pairs of UBMs (higher NID), consistently the greater the reduction in terms of EER for the fused system. This supports the hypothesis that diversity can be achieved purely through different partitioning of the acoustic space as speculated in [1]. In our previous studies, an ensemble of different variants of MFCCs were proposed [1] and showed that the fusion of suboptimal systems based on features comprising essentially the same information as MFCCs does commonly outperform an individual MFCC based system. This prompts a re-evaluation of what types of features might be considered complementary.

4. INVESTIGATION OF ACOUSTIC MODELLING

The aim of this experiment was to assess the “stability” of the clusters (GMMs) with respect to different features, that is, the robustness of the putative clusters to sampling variability. One hypothesis for the poorer performance of alternative features is that their ability to describe the acoustic space of a speaker is less stable/reliable than that of MFCC [1]. This experiment will be conducted through the use of resampling, proposed in [2, 11], to assess the stability of the clustering results with respect to sampling variability by simulating permutations of the original data set. The basic assumption of this method is intuitively simple: if the data represent a sample of items drawn from distinct sub-populations, and if we were to observe a different sample drawn from the same sub-populations, the induced cluster composition should not be radically different. Therefore, the more the attained clusters are robust to sampling variability, the more we can be confident that these clusters represent the real underlying structure.
Following studies in [2, 11], a similar experiment was conducted on the NIST 2004 female database (1849 speakers), with the NID as a measure of agreement between alternative data. $k = 10$ permutations of the original data were simulated according to the feature stability assessment procedure summarized in Figure 1. Shown in Table 2 are the average NID and EER for various features. It can be observed from the results that better feature stability, or smaller average distance between the putative clusters (lower average NID) corresponds with a lower EER, demonstrating the importance of stability in acoustic modelling, and helping explain in clustering terms why MFCC usually outperforms alternative features. As with section 3, we find that the acoustic modelling has a strong relationship with the EER - in both sections the NID is computed only on the UBM and speaker modelling is not a part of these NID calculations.

5. INVESTIGATION OF SPEAKER MODELLING

In Section 3, we have shown that different features result in different partitioning of the acoustic space. In this section, we investigate the extent to which different features give different clusterings after MAP adaptation, with respect to the UBM. This methodology is necessarily different because MAP adaptation of the UBM is done on a per-utterance basis, and it is not possible to associate an EER with each NID between the UBM and utterance-adapted UBM. Instead, we look at the adaptation of the UBM for a single speaker (1849 utterances) for two very differently clustered features (MFCC and SCF), expecting to see that adapted models exhibit more deviation in one feature domain than the other (varying between different features), if we employ the assumption that different features carry different speaker-specific information. Figure 2 shows the NID comparison between MFCC and SCF with an average NID of 0.5002 for MFCC and an average NID of 0.5361 for SCF. Surprisingly, we observed that all pairwise feature sets shown in Table 1 have almost equal amounts of adaptation in both feature domains (not shown due to space constraints), with respect to the UBM. This might suggest that different features do not differ very much in how they model individual speakers, or that speaker modelling may not be as important as acoustic modelling.

6. UBM DATA SELECTION USING CLUSTERING COMPARISON

Based on the results of the foregoing sections, it seems that acoustic modelling is a good area to focus investigation in speaker recognition, and further that feature stability has an important role to play. The consensus clustering based approach in Figure 1 suggests itself for application to data selection.

A common assumption in UBM training for speaker recognition is that the more utterances used, the better the system performance. However according to [12], Reynolds et al. mentioned that a small amount of data is sufficient for reasonable system. Recently a few groups have investigated this further: Hasan et al. proposed a novel feature subsampling method for selecting UBM feature vector frames [13] and Omar et al. looked into alternative techniques for training UBM [5], in which they also demonstrate the significance of the UBM in terms of entire speaker recognition system performance.

In this section, we utilize the feature stability assessment procedure in Section 4 for choosing utterances that would contribute to a stable UBM (procedure as outlined in Figure 3), and evaluate these on the NIST 2006 database. The features used are 14 MFCCs + 14 $\Delta$ extracted from 20 ms

Table 2 NID and EER on NIST2004 female

<table>
<thead>
<tr>
<th>FEATURES</th>
<th>AVERAGE NID</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>0.44</td>
<td>6.66</td>
</tr>
<tr>
<td>LPCC</td>
<td>0.45</td>
<td>6.74</td>
</tr>
<tr>
<td>MFCCDropCep</td>
<td>0.47</td>
<td>6.90</td>
</tr>
<tr>
<td>PLP</td>
<td>0.47</td>
<td>7.45</td>
</tr>
<tr>
<td>SCF</td>
<td>0.56</td>
<td>7.82</td>
</tr>
<tr>
<td>LogLSGD</td>
<td>0.57</td>
<td>8.12</td>
</tr>
</tbody>
</table>

Figure 1 Feature stability assessment procedure (after [2])

Figure 2 NID between clustering of UBM and speaker model for MFCC and SCF where each ‘+’ sign correspond to one speaker (1849 utterances)
Hamming-windowed frames, overlapped by 10 ms, using a 26 mel-scaled triangular filterbank. Figure 4 shows the EER versus x% UBM training data for random UBM utterances selection and our proposed UBM utterances selection technique. We can observe in both cases that using fewer utterances for training UBM can result in better system performance. This could be due to the fact that introducing more utterances to a "stable" UBM may simply increase the variability within the UBM which might not be desirable in terms of acoustic space modelling. Using the proposed UBM data selection algorithm, we achieve a 4% and 11% relative reduction in EER using only 20% and 30% of the usual female and male UBM training data set respectively (~370 utterances for each gender); a better performance improvement as compared with [13] and random utterance selection. In addition, this corresponds to a reduction in UBM development time of up to 3 times (once the data have been selected) as compared with using all utterances. For further investigation, data selection employing segment rather than utterance likelihood will be considered.

7. CONCLUSION

In this paper, we utilize the clustering comparison measures, in particular the Normalised Information Distance, to test the hypothesis that front-end diversity is achieved largely through different partitioning of the acoustic space. A novel utterance selection algorithm for training a “stable” UBM is presented and evaluated on the NIST 2006 database. Results show that using NID-based resampling to select utterances during UBM training can improve speaker recognition performance despite employing a smaller set of training data.

8. REFERENCES