Automatic Cross-Biometric Footstep Database Labelling using Speaker Recognition

Ruben Vera-Rodriguez[1], John S.D. Mason[1] and Nicholas W.D. Evans[1,2]

[1] Speech and Image Research Group, Swansea University,
Singleton Park, Swansea, SA2 8PP, UK
[2] Institut Eurécom, 2229 route des Crêtes, 06904 Sophia-Antipolis, France
(r.vera-rodriguez.405831, j.s.d.mason)@swansea.ac.uk, nicholas.evans@eurecom.fr

Abstract. The often daunting task of collecting and manually labelling biometric databases can be a barrier to research. This is especially true for a new or non-established biometric such as footsteps. The availability of very large data sets often plays a role in the research of complex modelling and normalisation algorithms and so an automatic, semi-unsupervised approach to reduce the cost of manual labelling is potentially of immense value.
This paper proposes a novel, iterative and adaptive approach to the automatic labelling of what is thought to be the first large scale footstep database (more than 10,000 examples across 127 persons). The procedure involves the simultaneous collection of a spoken, speaker-dependent password which is used to label the footstep data automatically via a pre-trained speaker recognition system. Subsets of labels are manually checked by listening to the particular password utterance, or viewing the associated talking face; both are recorded with the same time stamp as the footstep sequence.
Experiments to assess the resulting label accuracy, based on manually labelled subsets, suggest that the accuracy of the automatic labelling is better than 0.1%, and thus sufficient to assess a biometric such as footsteps, which is anticipated to have a much higher error rate.

Key words: Automatic database labelling, speaker verification, score normalisation, footstep biometric, multimodal biometrics.

1 Introduction

When developing a new biometric one of the first considerations entails the collection of a representative dataset of meaningful size. Data collection is notoriously expensive and problematic but instrumental to the success of a new project and confidence in the results. Many fundamental questions need to be addressed. Among them are: the number of samples; the number of clients; the enrolment and labelling procedures. To get these wrong would devalue the database and any results derived from it.

Doddington's 'rule of 30' [1] gives some guidance regarding the number of samples, the expected error rate and the confidence in the result. He states that

'to be 90 percent confident that the true error rate is within +/- 30 percent of the observed error rate, there must be at least 30 errors'. Thus if we expect a relatively higher error rate we may be satisfied with a smaller database than if we expect a relatively lower error rate; there is a trade-off between database size and expected error rate. Of course if we are researching a new biometric then we cannot know the expected error rate; it is likely that one of the fundamental goals of the research is to establish precisely this. We might opt to conduct an initial trial on a small dataset to help us decide upon the required database size, however to extract the best value the database should be sufficient for both today's and tomorrow's research. Advances in biometrics research often come from very large databases designed to facilitate the learning of complex modelling and normalisation strategies which may not have been possible on smaller datasets. Thus in order to prepare for the research of tomorrow it is in any case always advantageous to collect as large a database as possible within economic and practical constraints.

The financing of database collection can however be difficult to obtain especially when the research involves a new biometric for which a baseline error rate does not exist and we cannot reliably predict the potential of the biometric under investigation. It is sometimes possible to reduce the cost of collection through automated collection systems and this can go someway to help the labelling of the collected data, namely the assignment of ownership to each collected sample.

In the collection system described here there are two modes, supervised and unsupervised. The initial enrolment of each person participating in the database collection (donor) is supervised and hence the allocation of an identity label to this enrolment data is also supervised. Subsequently, and for the large majority of the data collection, the process is unsupervised. Thus a strategy is necessary to assign the correct donor identity to each of the recorded signals. This paper describes such a strategy using a combination of automation plus human cross-checking. The automation itself uses a biometric approach based on person specific spoken utterances captured at the same time as each of the footstep signals is captured and recorded.

2 Concept of Automatic Labelling

The cost of human resources can represent a potential barrier to research: we might have difficulty in justifying and financing the collection and manual labelling of a large database unless the commercial potential of a new biometric is proven. However, we might not be able to demonstrate the true potential of the biometric without a large database and we have something of a chicken and egg situation. Over recent years we have been investigating a relatively new and little researched biometric and in this paper we describe our approach to collect a large database without the full cost of manual labelling. The idea is to collect a multimodal biometric database, where the primary mode of interest is footsteps and the secondary modes are speech, talking face and gait. Each set (comprising footstep, speech, talking face and gait sequence) is assumed to be

consistent, coming from just one person, by a time stamp assigned at the time of capture from a single clock. Of course, in the absence of supervision, more than one person could mischievously combine to give anomalous data sets but this is thought to be a small risk under the given conditions. The secondary modes, particulary speech and talking face, are included specifically to assist in labelling the database in order to subsequently carry out biometric research based on footsteps. This labelling process is a combination of enrolment, automation and human cross-checking. This paper describes the automation process based on (acoustic) speaker recognition with the goal of accurately labelling the database. Speech is chosen for convenience and is also one of very low proven error rates [2]. Speech also perhaps reduces the need for sequestering when compared to other biometrics with lower EER, like fingerprints or iris. The assignment of a speaker specific PIN gives a text-dependent characteristic and importantly provides a means of human based cross-checking of the labels. The PIN is allocated at the (supervised) enrolment stage.

Naturally, even when the speaker recognition has proven low error rates, there will nonetheless be concerns over the labelling accuracy and its unpredictable repercussions. This though, has to be seen in the context of (i) international evaluation campaigns which have been shown to contain labelling anomalies, and (ii) the trade-off between a large database (with a small number of labelling errors) which provides the richness required for the development of complex modelling and normalisation algorithms and a smaller database with (possibly) no labelling errors. In any case for a larger database perhaps inevitably with some labelling anomalies, we suppose that (i) the potential of the biometric can be assessed and (ii) we have greater confidence in the results than we would otherwise have for a smaller database. Even the sceptic has to accept that it would be unwise to suppose that a database does not have labelling anomalies and their potential for occurrence is a function of the size of the database. In the approach described in this paper we accept that labelling errors are possible, and this paper describes the efforts to minimize the number of errors using a range of strategies.

Speech is used for the automatic labelling as shown in Figure 1. There is a large data set captured in an unsupervised mode and hence unlabelled. This is introduced into an speaker recognition system, which is trained on manually labelled data as ground truth. Then a decision based on the speech signal is taken to obtain new labels. These new labels apply not only to the speech but also to the other contemporaneous signals in the data set, namely the footstep signals together with the talking face and gait image sequences. We refer to one example of these four signals as a *set* with the primary interest here being the footsteps and the speech. The talking face has benefit in the manual labelling and cross-checking for anomalies of a set.

At the time of carrying out the following experiments the database was comprised of a total of 11,537 sets. The collection of the database was unsupervised, apart from the initial enrolment session of each person in the system, where normally around 10 sets were manually labelled. Apart from the enrolment data

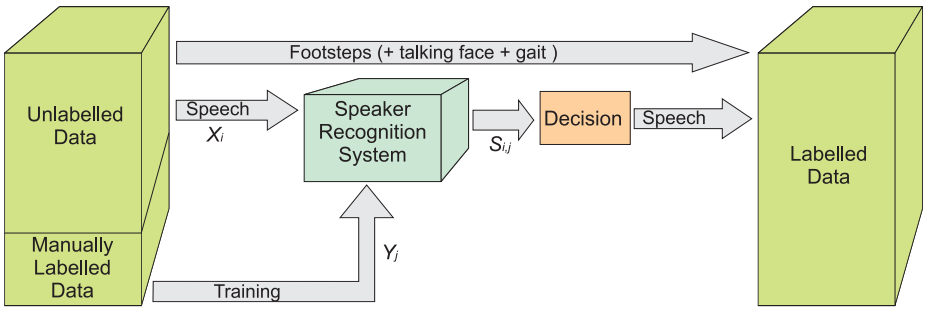

Fig. 1. Speech-based automatic labelling system. The speaker recognition system trained on labelled data receives unlabelled data. The decision process considers all combination of scores from the data on the left and systematically labels the most likely sets, passing them across to the right repeating the process until all data is labelled or discarded.

(1,123 sets in total from 127 clients), more data was manually labelled (1,385 sets in total). This labelling exercise took part during the collection stage itself and proved the enormity of the task, amply demonstrating the need for automation.

In the speaker recognition system, each X_i test data, left of Figure 1, being $i = 1....9{,}029$, is tested against the Y_j models, being $j = 1....127$, created from the manually labelled data. This could be seen as a form of 1-in-N identification if there was confidence that X_i definitely came from one of the Y_j persons in the enrolled set. Alternatively, the task may be viewed as a specific case of verification applying acceptance to the most likely pairing across all X, Y combinations. The benefit of this interpretation is (at least) two-fold, both being of critical importance. The first covers the case when X_i is outside of set Y. The second is in terms of score normalization. Before any test-to-model score is assessed it is normalized using standard techniques well established in the speaker recognition world [8–10].

The key point is that, the assignment of any data set is prioritized in an order of confidence. This means that the most likely assignments take place first. Also once X_i is assigned, to say Y_j, then it is possible to re-train the model for person Y_j in an adaptive manner, potentially improving the model as a representation of person Y_j. Of course such adaptation can be dangerous in the case of false acceptance. We address this issue in our experimental procedures (Sect. 5).

3 Speaker Recognition System

The speaker recognition system shown in Figure 1 is based on a linear frequency cepstral coefficient (LFCC) front-end and a Gaussian mixture model (GMM) system [4], using SPro[3] and ALIZE[4] open source toolkits. The GMM system

[3] http://gforge.inria.fr/projects/spro/
[4] http://www.lia.univ-avignon.fr/heberges/ALIZE/

is close to the description in [5]. The signal is characterised by 33 coefficients including 16 LFCC, their first derivative coefficients and the energy derivative (16LFCC + 16Δ + ΔE).

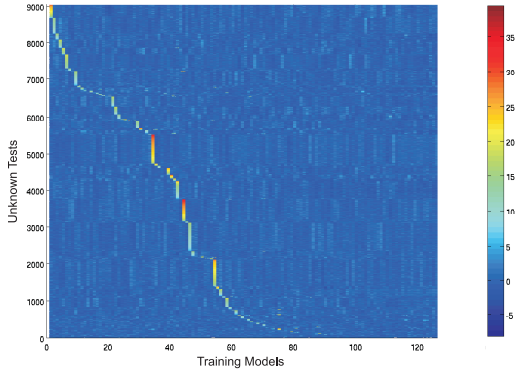

Fig. 2. Representation of the score matrix after using the speaker verification system. Test signal scores against trained models with manually labelled data (enrolment plus human labelling).

The first experiment relates to 9,029 unlabelled sets tested against each of the 127 models trained on the manually labelled sets (2,508). The recogniser scores $S_{i,j}$ can form a score matrix as represented in Figure 2 with scores for the 9,029 unlabelled sets plotted against the 127 models. The colour scale in the figure shows the range of the scores. The 127 models are sorted according to the amount of training data, with the largest on the left to models with the smallest amount of training data on the right. The correlation with the amount of training data and high scores is distinct, with almost half of the 127 models having very few high scores. The 9,029 test signals have been sorted to have high scores together per model, this way we could have an idea of the number of tests that belong to each model. As Figure 2 shows, there are not many high scores for models on the right; this suggests that there are not many test signals that belong to those models. It is clear that score normalisation is essential in order to remove the influence of the amount of data in the training models.

4 Score Normalisation

Score normalisation is widely used in biometrics, for example it is a key factor when fusing different biometric modalities, and here in the context of speaker recognition it is required to balance inherent test data and model variation. The most popular methods for score normalisation are T-norm [8, 9] and Z-norm [10].

4.1 Test Normalisation

First, a test normalisation is applied to the scores from each of the Y models for a given test utterance (the rows of the score matrix in Figure 2). This normalisation attempts to align test scores by using scores from impostor models. In the case where a particular unknown test belongs to one of the models, only one of the scores can be a true score while the rest correspond to impostor scores. Of course, in the present context the one particular model is unknown, hence here all 127 scores are used with the one potentially true score assumed to be swamped by the remaining 126 scores. A general equation for the test normalisation is given by:

$$STn_{i,j} = (S_{i,j} - \mu_i)/\sigma_i \tag{1}$$

where $STn_{i,j}$ are the normalised scores and μ_i and σ_i the mean and standard deviation of the impostor scores respectively for each test. Figure 3 (a) shows an example of a test signal. The plot shows scores against the 127 models corresponding to a horizontal trajectory across Fig 2), and indicating a very high likelihood of ownership of the given test signal by model number 105. The profile after the test normalisation can be observed in Figure 3 (a) with red dashed line. As observed, the effect of this normalisation makes a wider range for the scores and also gives a common zero threshold. Figure 3 (b) shows the same but for one of a bad test example where there is no score clearly higher than the rest.

4.2 Model Normalisation

Second, model normalisation is applied which attempts to align between-speaker differences by producing statistical parameters for each model to align the scores to zero. Figure 3 (c) shows an example of the score distribution for a good model. As the figure shows there are some high scores on the right (zoom of the distribution on the top of Figure 3 (c)) that are likely to correspond to true scores for this model, and a large number of low scores on the left, which would be likely to belong to other models or not within the 127 set. To carry out a normalisation similar to Z-norm, it is necessary to obtain the distribution for the out of class data to calculate its mean and standard deviation. Therefore, the score distribution was approximated by two gaussian distributions, using the one with lower mean and higher weight as the distribution of the out of class data. A general equation for the model normalisation is given by:

$$SZTn_{i,j} = (STn_{i,j} - \mu_j)/\sigma_j \tag{2}$$

where $SZTn_{i,j}$ are the normalised scores and μ_j and σ_j are the mean and standard deviation of the out of class scores respectively for each model. Figure 3 (e) shows the effect that this normalisation makes to the distribution, i.e. aligning the out of class distribution to zero. Figure 3 (d) shows an example of a bad model, one with either small amount of training data, or few high scores.

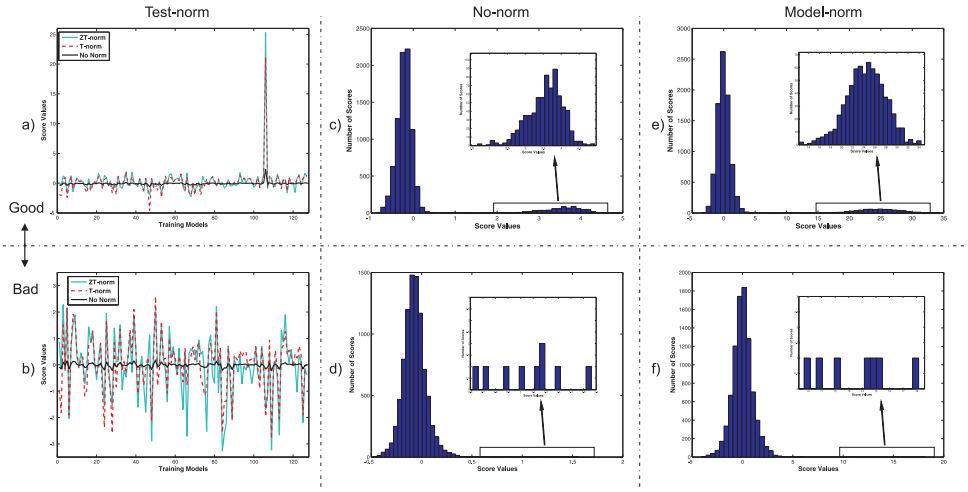

Fig. 3. (a)Example of scores for a good test utterance against 127 models. Black solid profile before normalisation. Red dashed profile after T-norm. Solid light blue profile after ZT-norm. (b) Same as (a) but for a bad test utterance. (c)Example of score distribution for a good model and 9,029 test sets before normalisation. (d) Same as (b) but for a bad model. (e) Same model as (c) but after ZT-norm. (f) Same model as (d) but after ZT-norm.

The effect of the model normalisation can be observed in Figure 3 (f). This normalisation is called ZT-norm and in [11] a reduction of 20% EER is reported when compared with standard Z-Norm.

5 Automatic Labelling, Implementation and Results

This section describes the iterative process followed to carry out the labelling of the database. Figure 4 shows the structure of the database, which is a ragged array with different amounts of data per person and per labelling class: enrolment, manually labelled, automatically labelled (by speaker recognition) and unlabelled. The diagram represents a state during the iterative recognition and label allocation process. The enrolment session is the only one that is square as there are consistently 10 data sets per model. Then, there is another extra set of manually labelled data, followed by the automatically labelled set and finally the unlabelled set. Unlabelled data on the top is iteratively moved down as it becomes labelled by an iterative process. The automatic labelling process may be summarised by the following steps:

- Test all unlabelled data against all models (obtaining scores as per Figure 2).
- ZT-norm the scores.
- For each unassigned set find the most likely model.
- Sort all unassigned sets.

– Label the best set(s).
– Update model(s) using newly labelled data.

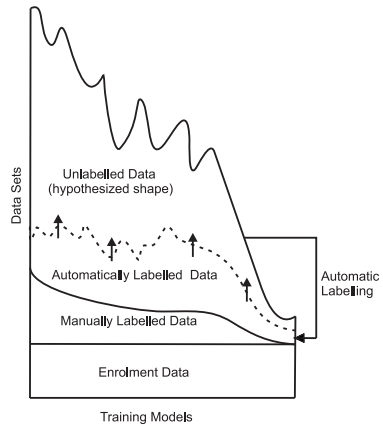

Fig. 4. Database structure. Enrolment session data on the bottom, then manually labelled data, assigned data and unlabelled data on the top. In each iteration the best sets of the unlabelled data are assigned to the respective models according to the criteria defined.

This process is repeated until the label confidence threshold is reached shown as the knee point in Figure 5 (a). As indicated above, the assignment of the unknown sets to the models is an iterative process. In each iteration unlabelled data is tested against 127 models producing a score matrix as in Figure 2. Then a ZT-norm is carried out as described in Section 4. At each iteration, data with the highest overall score statistics (best sets) are assigned to their respective models, and then are used to re-train the models in an adaptive manner. To prioritize the tests to be assigned to the models in an order of confidence, the tests are sorted considering statistics of the peak score and variance of the remaining 126 scores for each test (rows of Figure 2). In this way, tests with less confidence remain until the last iterations. This strategy is independent to the number of models, which in a dynamic database can be variable.

At each iteration it is possible to estimate a figure of the relative merit of the assignment. Figure 5 (a) shows the EER for the tests assigned at each iteration. The left part of the figure is a region with high confidence due to the sorting of the tests assigned (tests have similar profiles as example in Figure 3 (a)). The middle part of Figure 5 (a) is a region with less confidence, and the right part is a region of very low confidence suitable for manual labelling or rejection, where examples such as those in Figure 3 (b) can be found. Figure 5 (a) shows how the EER stays in a range of 0.06% - 0.1% until the last 400 tests are assigned. Then the error increases to 0.58%. This suggests that either this data are different in some way to data previously assigned either due to within class variation

or alternatively the data comes from outside of the set of the 127 speakers. Therefore, this data in this region and to the right should be labelled manually or rejected. Using the above strategy, 8,867 out of the 9,029 unlabelled sets were each labelled to one of the 127 models.

In order to assess the labelling accuracy, a further speaker recognition experiment was conducted using manually labelled sets. This represents a meaningful evaluation of the speaker recognition system and hence a reasonable assessment of the automatic labelling of the database. The results are shown in Figure 5 (b). The profiles relate to two systems trained on a manually labelled set of data common to both, namely a smaller set of further manually labelled data and a much larger set (8,867) of automatically labelled data. most reassuringly, the error rate for the later are smaller than for the manually labelled set. This might be attributed to differences in the gender ratio: in the manually labelled set a balance across gender was sought; in the automatically labelled set there are proved to be more males than females given that is well known that speaker recognition error rates tend to be higher for females. This could account for the difference in the two profiles.

Finally, having lower confidence levels, data in the vicinity of the knee point in Figure 5 (a) could be manually labelled. Even so, the numbers requiring manual labelling have been massively reduced by the automatic procedure described with a predicted error rate in the region of 0.1% well within limits for study of footsteps as a biometric, the ultimate goal of this work.

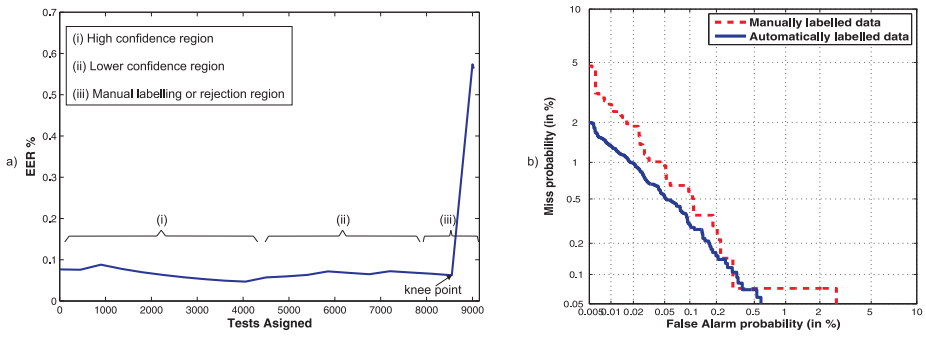

Fig. 5. (a) EER against test assigned in the iterations. (b) DET curve estimating the expected error of the speaker recognition system. Manually labelled data in dashed profile, and automatically labelled data in solid profile.

6 Conclusions

This paper describes an automatic system to label a database designed to assess footsteps as a biometric. The novel contribution is the way in which the data

has been collected and labelled. A total of four modes were collected simultaneously namely footsteps, speech, talking face and gait. All four modes within a set were linked by the same time stamp. Of principal importance are the footsteps, followed by the speech, the later included for labelling both manually and automatically. The large majority of the 9,000 plus signals have been labelled automatically using speech as a biometric. This has significantly reduced the manual effort and therefore cost of creating the database. Most large scale collections such as this one, are likely to have some form of data anomalies. Here we estimate the labelling errors to be less than 1% and thus, sufficient for the objectives of studying footsteps as a biometric. Finally it is clear that this system could be applied to the collection of other large scale biometric databases, where supervision and labelling is likely to prove expensive.

References

1. Doddington, G.R., Przybocki, M.A., Martin, A.F., Reynolds, D.A.: The NIST speaker recognition evaluation: Overview methodology, systems, results, perspective. Speech Communication. vol. 31, 225–254 (2000)
2. Toledano, D. T., Esteve-Elizande, C., Gonzalez-Rodriguez, J., Fernandez-Pozo, R., Hernandez-Gomez L.: Phoneme and Sub-Phoneme T-Normalization for Text-Dependent Speaker Recognition. In: Proc. IEEE Speaker and Language Recognition Workshop (Odyssey) (2008)
3. Vera-Rodriguez, R., Lewis, R.P., Mason, J.S.D., Evans, N.W.D.: A Large Scale Footsteps Database for Biometric Sutdies Created using Cross-Biometrics for Labelling. To appear in: 10th IEEE International Conference on Control, Automation, Robotics and Vision, ICARCV, Vietnam (2008)
4. Reynolds, D.A., Quatieri, T.F., Dunn, R.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing. vol. 10, no. 1-3, 19–41 (2000)
5. Bonastre, J-F., Scheffer, N., Fredouille, C., Matrouf, D.: NIST04 Speaker Recognition Evaluation Campaign: New LIA Speaker Detection Platform Based on ALIZE Toolkit. In: Proc. NIST SRE04 Workshop, Spain (2004)
6. Fauve, B., Evans, N. W. D., Mason, J. S. D.: Improving the Performance of Text-Independent Short Duration GMM and SVM Based Speaker Verification. In: Proc. Odyssey: the Speaker and Language Recognition Workshop, (2008)
7. Fauve, B., Bredin, H., Karam, W., Verdet, F., Mayoue, A., Chollet, G., Hennebert, J., Lewis, R., Mason, J., Mokbel, C., Petrovska, D.: Some Results From The Biosecure Talking Face Evaluation Campaign. In: Proc. ICASSP, (2008)
8. Auckenthaler, R., Carey, M.J., Lloyd-Thomas, H.: Score normalisation for text-independent speaker verification system. Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 Speaker Recognition Workshop. vol. 10, no. 1-3, 42–54 (2000)
9. Navrati, J., Ramaswamy, G.N.: The awe and mystery of T-Norm. In: Proc. Eurospeech, 2009–2012, Geneva (2003)
10. Li, K.P., Porter, J.E.: Normalizations and selection of speech segments for speaker Recognition Scoring. In: Proc. ICASSP, 595–598, (1988)
11. Zhang, S., Zheng, R., Xu, A.: A Comparative Study of Feature and Score Normalization for Speaker Verification. In: Proc. IEEE Odyssey Speaker and Language Recognition Workshop, 531–538, (2005)