Huang et al. EURASIP Journal on Audio, Speech, and Music Processing (2015) 2015:7
DOI 10.1186/s13636-014-0048-z

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

An investigation of supervector regression for forensic voice comparison on small data

Chee Cheun Huang[1,2*], Julien Epps[1,2] and Tharmarajah Thiruvaran[1]

Abstract

Automatic forensic voice comparison (FVC) systems employed in forensic casework have often relied on Gaussian Mixture Model - Universal Background Models (GMM-UBMs) for modelling with relatively little research into supervector-based approaches. This paper reports on a comparative study which investigates the effectiveness of multiple approaches operating on GMM mean supervectors, including support vector machines and various forms of regression. Firstly, we demonstrate a method by which supervector regression can be used to produce a forensic likelihood ratio. Then, three variants of solving the regression problem are considered, namely least squares and ℓ_1 and ℓ_2 norm minimization solutions. Comparative analysis of these techniques, combined with four different scoring methods, reveals that supervector regression can provide a substantial relative improvement in both validity (up to 75.3%) and reliability (up to 41.5%) over both Gaussian Mixture Model - Universal Background Models (GMM-UBMs) and Gaussian Mixture Model - Support Vector Machine (GMM-SVM) results when evaluated on two studio clean forensic speech databases. Under mismatched/noisy conditions, more modest relative improvements in both validity (up to 41.5%) and reliability (up to 12.1%) were obtained relative to GMM-SVM results. From a practical standpoint, the analysis also demonstrates that supervector regression can be more effective than GMM-UBM or GMM-SVM in obtaining a higher positive-valued likelihood ratio for same-speaker comparisons, thus improving the strength of evidence if the particular suspect on trial is indeed the offender. Based on these results, we recommend least squares as the better performing regression technique with gradient projection as another promising technique specifically for applications typical of forensic case work.

Keywords: Forensic voice comparison (FVC); Likelihood ratio; Reliability; Supervector regression; Validity

1 Introduction

Forensic voice comparison (FVC) systems have often employed Gaussian Mixture Model - Universal Background Models (GMM-UBMs) [1-3] for modelling in forensic casework, in which it is common that only a very small speech database is available for the entire system development. Other approaches, such as the supervector-based regression techniques prevalent in numerous face and speaker recognition studies [4-6], have received little attention in this context. This therefore motivates a comparative study of the effectiveness of other modelling approaches in FVC system performance.

The likelihood ratio is defined as the likelihood that the evidence would be observed if the same-origin hypothesis were true divided by the likelihood that the evidence would be observed if the different-origin hypothesis were true [7-11].

$$\mathrm{LR} = \frac{p(E|H_{\mathrm{so}})}{p(E|H_{\mathrm{do}})} \tag{1}$$

where LR is the likelihood ratio, H_{so} is the same-origin hypothesis and $H\mathrm{d}_{\mathrm{o}}$ is the different-origin hypothesis. E is the evidence or the observed property of a speech sample. $p(E|H_{\mathrm{so}})$ denotes conditional probability density of the evidence given same-origin hypothesis. An FVC system typically relies on statistical evaluation of input speech utterances that first involves training or modelling of the speaker identity based on an input speech utterance A and a subsequent testing of the trained model based on an input speech utterance B.

The initial output of the FVC system is defined as a score s. A higher valued score can be interpreted as

* Correspondence: cheecheunh@hotmail.com
[1]School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, NSW 2052, Australia
[2]National ICT Australia (NICTA), Australian Technology Park, Sydney, NSW 1430, Australia

providing greater support for the same-origin hypothesis over different-origin hypothesis whereas a lower valued score can be interpreted as providing greater support for the different-origin hypothesis over the same-origin hypothesis. This also implies that if identities A and B are from the same speaker, a higher score should be generated. Conversely, if identities A and B are from two different speakers, a lower score should be generated. The absolute value of a score cannot be directly interpreted as likelihood ratio; however, it can be seen as an intermediate step towards calculating a likelihood ratio, providing it can adequately account for both similarity and typicality [12]. Here, 'similarity' refers to the similarity of the pair of suspect-offender recordings, and 'typicality' refers to the typicality of the pair of suspect-offender recordings with respect to a model of the relevant population. The relevant population is the population to which the offender belongs and can typically be restricted to speakers of the same gender, language and dialect and similar age group as the offender on the basis of the offender recording and that these selection criteria would not be disputed by either the prosecution or the defence [9,11,13]. For more detailed discussions on score and likelihood ratio, readers may refer to [12,14].

In this paper, we present a study comparing methods for generating scores on the basis of the various modelling approaches. In particular, this paper presents a first study of score generation based on supervector domain regression for FVC on small data. Methods suitable for deriving a likelihood ratio from GMM supervectors [15,16] are considered herein, including pairwise nearest neighbour (PNN) and sparse regression techniques. Further, we investigate the applicability of these methods to small speech databases that are relevant to forensic case work, under clean, degraded and mismatched conditions.

2 Related work

2.1 Gaussian mixture model - universal background model FVC

The Gaussian Mixture Model - Universal Background Model (GMM-UBM) [1-3] is a prevalent speaker modelling technique used extensively in FVC and has become the primary method for modelling and likelihood ratio calculation in automatic FVC systems, see in particular [7,17,18]. In the context of FVC, data vectors representative of the voice recordings of speakers from the relevant population (i.e. background database) are used to the train the UBM (i.e. a UBM representing the different-origin hypothesis), while data vectors representative of the non-contemporaneous voice recordings from the suspect (i.e. suspect database) are used to perform MAP adaptation to form GMMs (i.e. models representing the same-origin hypothesis). The offender data vectors can then be evaluated against these two models (by taking the ratio of the two probability density values corresponding to the GMM and UBM models respectively at the offender value) to arrive at a likelihood ratio as illustrated in Figure 1. This GMM-UBM system, employed

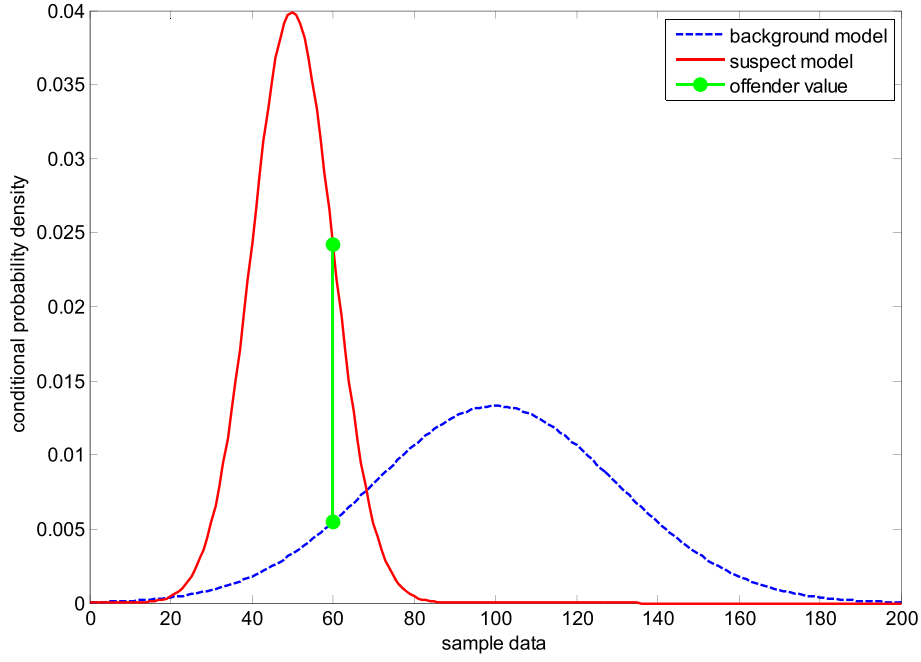

Figure 1 Simplified diagram showing the use of suspect model and background model. A simplified diagram showing the use of suspect model and background model in the calculation of forensic likelihood ratio by taking the ratio of the two probability density values corresponding to the GMM and UBM models at the offender sample value of 60 in this example.

in the current study as a baseline system, is depicted in Figure 2.

In the case of GMM-UBM modelling, likelihood ratio calculation is performed at the frame level initially with each frame of the offender recording producing a single likelihood ratio. Multiple likelihood ratios are therefore obtained in consideration of all frames within the offender recording at an utterance-based level. To combine these frame-level likelihood ratios, the mean of the natural log of these frame-level likelihood ratios is calculated, and the resulting value is referred to as a score. A subsequent score-to-likelihood-ratio transformation is performed by using logistic regression calibration [14,22,23]. Mathematically, this is shown in Equation 2 where for a given test utterance from the sample of questioned origin parameterized into a sequence of acoustic observations or feature vectors $\mathbf{X} = \{\mathbf{x}_1,\ldots,\mathbf{x}_T\}$, the score s of the test utterance is often expressed as

$$s = \frac{1}{T}\sum_{t=1}^{T}\log\left(\frac{p(\mathbf{x}_t|\lambda_{\mathrm{so}})}{p(\mathbf{x}_t|\lambda_{\mathrm{do}})}\right), \tag{2}$$

with λso and λ_{do} denote the probability density function parameters modelling the same-origin and different-origin hypotheses, respectively.

It is also common among the automatic FVC community to adopt two-stage LR computation first proposed by Meuwly in 2001 [24] and subsequently used in many other studies [7,17,18,25]. In the current study, small databases of a few tens of speakers that are more relevant to forensic case work applications were employed (similarly to [26,27]), and therefore we adopted a simpler one-stage LR computation structure as depicted in Figure 2.

2.2 GMM mean supervector and support vector machine FVC

It is common in speaker recognition studies to employ a representation of a speaker using stacked d-dimensional mean vectors $\mathbf{m}_k$, $k \in \{1, .., K\}$ of a K-component adapted GMM into a Kd-dimensional Gaussian supervector $\mathbf{\Phi}$ [15]. Before stacking, the means are normalized with the factor $\sqrt{w_k}\Sigma_k^{-1/2}$, where w_k represents the kth Gaussian weight and Σ_k represents the diagonalised covariance of the kth mixture, to ensure a constant Kullback-Leibler (KL) distance between each of the supervectors [15,16], as seen in Equation 3. We will denote the GMM mean supervector derived from the offender recording as $\mathbf{\Phi}_{\mathrm{offender}}$, a GMM mean supervector derived from the suspect recording as $\mathbf{\Phi}_{\mathrm{suspect}}$ and GMM mean supervector derived from the background recording as $\mathbf{\Phi}_{\mathrm{UBM}}$.

$$\mathbf{\Phi} = \begin{bmatrix} \sqrt{w_1}\,\Sigma_1^{-\frac{1}{2}}\mathbf{m}_1 \\ \sqrt{w_2}\,\Sigma_2^{-\frac{1}{2}}\mathbf{m}_2 \\ \vdots \\ \sqrt{w_K}\,\Sigma_K^{-\frac{1}{2}}\mathbf{m}_K \end{bmatrix} \tag{3}$$

In speaker recognition, supervectors are typically applied as inputs to SVM [28,29] or joint factor analysis (JFA) [30,31]. The latter has been found to be very successful for modelling the inter-speaker variability and hence for compensating for channel or session effects in the high-dimensional GMM supervector space. The i-vector technique, a variant of JFA, performs channel compensation in a low-dimensional total variability space that is defined by factor analysis [32,33]. Both JFA and i-vector however are techniques which require

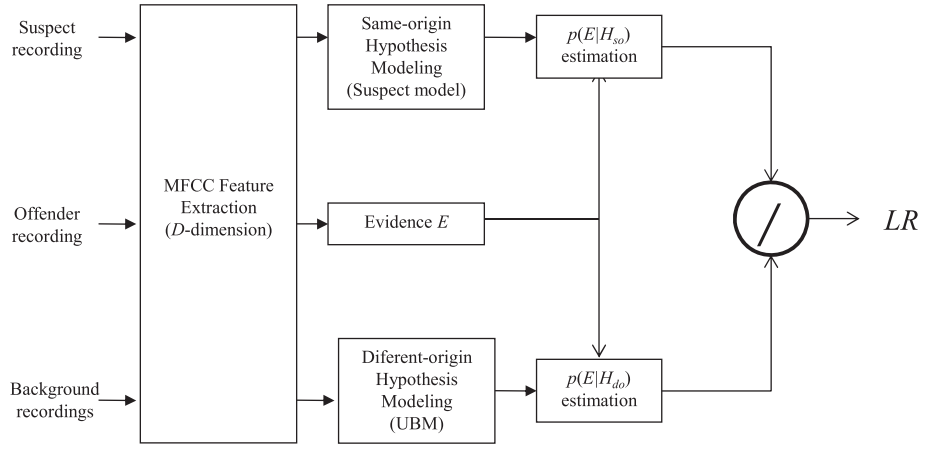

Figure 2 Conventional automatic FVC system based on Gaussian Mixture Model - Universal Background Models (GMM-UBMs), after [19-21].

independent large databases for training the appropriate inter-speaker variability models. By contrast, typical FVC studies have employed small databases, e.g. 68 male adult German speakers [26] or 27 male speakers of Australian English [27].

To demonstrate that JFA or *i*-vector techniques are ineffective for databases with a low number of speakers, in preliminary experiments, we attempted JFA and *i*-vector techniques based on [31] using the JFA cookbook (http://speech.fit.vutbr.cz/software/joint-factor-analysis-matlab-demo) developed by Ondrej Glembek at Brno University of Technology on our 60 female speaker forensic database. Out of the 60 speakers available, the background database was allocated 20 speakers, development and test databases were each allocated 10 speakers and the remaining 20 speakers were allocated for training the inter-speaker variability models associated with JFA or *i*-vector techniques. FVC results obtained were substantially poorer compared with FVC results from a GMM-UBM system based on the same 60 female speaker database, based on the same database allocation for background, development and test databases as the JFA or *i*-vector techniques. Other database arrangements were also investigated such as increasing the number of speakers assigned to train the inter-speaker variability models by the additional 20 speakers using the same 20 speakers from background database, and hence increasing the total number of speakers used for training the inter-speaker variability models to 40 speakers; however, results from JFA showed similar substantially poorer FVC results compared with results from the GMM-UBM system. This therefore implies that JFA or *i*-vector techniques will not perform well in FVC given that the inter-speaker variability models were based on such a low number of speakers.

Support vector machine (SVM) [28,34] is a discriminative classification technique that operates by defining a decision boundary between two classes separated by a hyperplane that maximizes the margin of separation between the two classes. In the context of FVC, data vectors representative of the voice recordings of speakers from the relevant population (i.e. background database) are used to form one class, while a data vector representative of a particular voice recording from the suspect (i.e. suspect database) is used to form the other class as illustrated in Figure 3. In the case of FVC, a binary decision is not sought but rather a forensic likelihood ratio indicating the strength of Evidence *E*. The likelihood ratio can be obtained by first computing the inner products of an offender vector with the support vectors in a higher dimensional kernel feature space. The resulting value from this computation of inner products on the basis of a kernel function is referred to as a score,

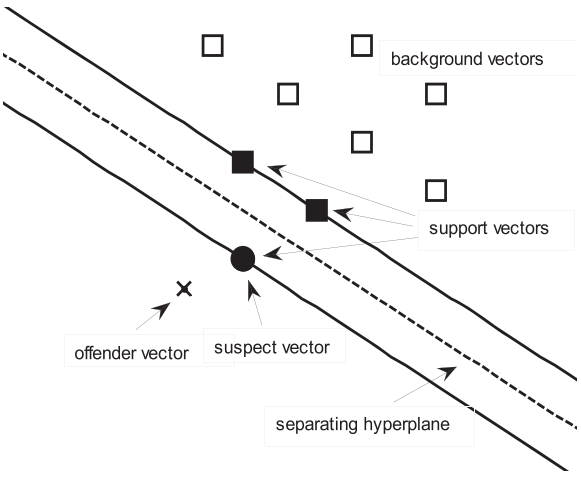

Figure 3 Simplified diagram showing the concept of support vector machine in the forensic likelihood ratio calculation. Circle: suspect vector, squares: background vectors, cross: offender vector.

and a subsequent score-to-likelihood-ratio transformation is performed by using logistic regression calibration [14,22,23]. The computed score is a similarity measure of the offender vector to the suspect vector while taking into account the typicality with respect to the background data vectors. It should be noted that although the SVM concept may not be completely forensically applicable, as potentially only a subset of background data vectors (i.e. the background support vectors) is utilized in the score computation rather than all background data, similarity (with respect to the suspect vector) and typicality (with respect to the background support vectors) are being appropriately evaluated via a relative distance measure to the offender vector in the score computation. A more positive valued score is obtained if the offender vector is closer to the suspect vector than the background support vectors, and a more negative valued score is obtained if the offender vector is closer to the background support vectors than the suspect vector.

In particular, consider an example in which an offender vector lies on the side of the linear separating hyperplane that contains the suspect vector and is in close proximity to the suspect vector, with all the background data vectors on the other side of the linear separating hyperplane. If we have the offender vector and all the background data vectors fixed in their positions while adjusting the suspect vector to move in the direction from the linear separating hyperplane that is further away from both the offender and background data vectors, then the offender vector now has a lower similarity (with respect to the suspect vector) while having the same typicality (with respect to the background support vectors), and this will result in a more negative-valued score. Conversely, if we have the

offender vector and the suspect vector fixed in their positions while adjusting all the background data vector to move in the direction from the linear separating hyperplane that is further away from both the offender and suspect vectors, then the offender vector now has the same similarity (with respect to the suspect vector), while having a lower typicality (with respect to the background support vectors), and this will result in a more positive-valued score. This approach is employed as an additional baseline system in our work, using supervectors as the data vectors.

Denoting $\mathbf{\Phi}_{\text{offender}}$ as the offender vector and $\mathbf{\Phi}_{\text{support},i}$ as the ith support vector, then the score generated on the basis of sequential minimal optimization algorithm from [35] can be computed (e.g. via the publicly available toolkit named LIBSVM [36]) as follows

$$s_{\text{SVM}}(\mathbf{\Phi}_{\text{offender}}) = \sum_{i=1}^{L} \alpha_i t_i \mathbf{\Phi}_{\text{offender}}{}^{t} \mathbf{\Phi}_{\text{support},i} + d \qquad (4)$$

Alternatively, the score may be generally expressed as

$$s_{\text{SVM}}(\mathbf{\Phi}_{\text{offender}}) = \sum_{i=1}^{L} \alpha_i t_i K\left(G_{\text{offender}}, G_{\text{support},i}\right) + d \qquad (5)$$

Here, $\sum_{i=1}^{L} a_i t_i = 0$ and $\alpha_i > 0$, $t_i \epsilon \{+1, -1\}$ are the ideal output values, L is the number of support vectors and both α_i and d are learned constants as defined in [28]. $G_{\text{support},i}$ is the GMM used to create the ith support vector $\mathbf{\Phi}_{\text{support},i}$. All parameters were obtained from training the SVM via an optimization process [35]. The kernel function $K(.,.)$ can be expressed as

$$K\left(G_{\text{offender}}, G_{\text{support},i}\right) = b(G_{\text{offender}})^{t} b\left(G_{\text{support},i}\right) \qquad (6)$$

where b is a mapping from input space (i.e. GMM model space) to a higher dimensional kernel feature space (i.e. GMM mean supervector space), that is $b(G_{\text{offender}}) = \mathbf{\Phi}_{\text{offender}}$.

2.3 Supervector-based regression techniques

Sparse representation of signals has been a major research interest in the area of statistical signal processing [37,38]. One of the significant discoveries in these studies revolves around the finding that if an optimal representation of a signal is sufficiently sparse when linearly represented with respect to a dictionary of base elements, it can be computed by convex optimization [38]. Although sparse representation can be used for solving a system of linear equations that are overdetermined as seen in [4], it has also shown promise for underdetermined systems as demonstrated in robust face recognition studies such as [39]. Mathematically, the sparse representation equation can be represented as $\mathbf{y} = \mathbf{A}\mathbf{x}$, in which the dictionary $\mathbf{A}$ is used to linearly represent signal $\mathbf{y}$ in a sparse manner.

There are numerous approaches to the solution for the $\mathbf{x}$ in the sparse representation equation. The sparse solution $\mathbf{x}$ will contain mostly zero entries, except those entries which correspond to the signal $\mathbf{y}$ are non-zero. The approach considered in this study is to treat the sparse representation equation as a regression-based problem, for which applicable techniques include least squares (LS) and ℓ_1 and ℓ_2 norm minimization [37,39,40]. A technique which uses a mixture of ℓ_1 penalty (lasso) and ℓ_2 penalty (ridge regression) on the basis of a tuning parameter known as the elastic net [41,42] is also considered. The discriminative nature of these sparse signal processing techniques has been exploited in numerous face and speaker recognition studies [4-6], which have employed dictionaries comprising GMM mean supervectors [4,43] or speaker factors [6] and achieved good experimental success. Other applications of sparse signal processing in the speaker recognition area include a study of GMM mean shifted supervectors using learned and discriminatively learned dictionaries [44] and a study employing feature vectors as the base elements in the dictionary [45].

After solving for the regression problem in the sparse representation equation, four scoring methods are considered in this paper. The first scoring method directly utilizes the first coefficient x_1 from the vector of coefficients $\mathbf{x}$ as a score, we henceforth named this scoring method as *direct parameter* $x(1)$. The remaining three scoring methods were based on ℓ_1 *norm ratio* and ℓ_2 *residual ratio* introduced in [5] and the *background normalized (BNorm)* ℓ_2 *residual* criterion introduced in [6]. The score generated from any of these methods can then be converted to likelihood ratio via logistic regression calibration [14,22,23].

3 Supervector regression FVC

3.1 Proposed paradigm

The overall supervector regression-based FVC system is shown in Figure 4.

To begin, we construct a dictionary $\mathbf{A}$ using the GMM mean supervectors $\mathbf{\Phi}_{\text{UBM}}$, derived from the MAP-adapted GMMs using all the recordings from the background database, which form the model of the relevant population. To facilitate offender-suspect pair comparison, an additional GMM mean supervector $\mathbf{\Phi}_{\text{suspect}}$ derived from one of the suspect recordings is prepended to the beginning of the dictionary, creating the first column of the dictionary. The supervector $\mathbf{\Phi}_{\text{offender}}$ derived from a particular offender recording is then

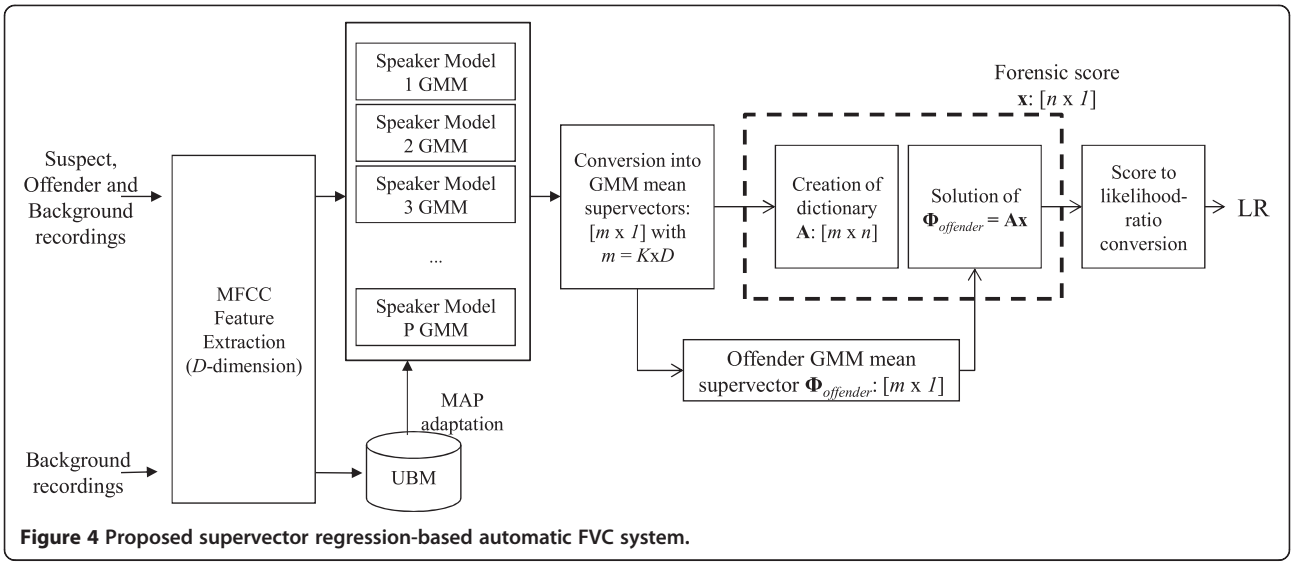

Figure 4 Proposed supervector regression-based automatic FVC system.

represented as a linear combination of this dictionary of size n, where n is the total number of $\mathbf{\Phi}_{\text{UBM}}$ supervectors plus one (since there is an additional $\mathbf{\Phi}_{\text{suspect}}$ supervector in the first column of the dictionary).

Mathematically, this can be expressed as $\mathbf{\Phi}_{\text{offender}} = \mathbf{Ax}$ per Equation 7 below with the m entries of the GMM mean supervector $\mathbf{\Phi}_{\text{offender}}$ represented as a linear combination of the dictionary, and $\mathbf{x}$ denoting the vector of n unknown coefficients that we wish to solve. For any offender-suspect pair comparison, we would like to use the dictionary $\mathbf{A}$ to linearly represent $\mathbf{\Phi}_{\text{offender}}$ in a sparse way [4-6,39,43]. After solving Equation 7 below, ideally if the first coefficient x_1 has a value of 1 while the other coefficients $x_i, (i = 2,\ldots,n)$ representing the background speaker supervectors $\mathbf{\Phi}_{\text{UBM}}$ have values of 0, then the offender supervector $\mathbf{\Phi}_{\text{offender}}$ and the suspect supervector $\mathbf{\Phi}_{\text{suspect}}$ (i.e. the first column of the dictionary) will have originated from the same speaker. Conversely, if x_1 has a value of 0 while the summation of $x_i, (i = 2,\ldots,n)$ has a value of 1, then the offender supervector $\mathbf{\Phi}_{\text{offender}}$ and the suspect supervector $\mathbf{\Phi}_{\text{suspect}}$ (i.e. the first column of the dictionary) should have originated from two different speakers. For each new offender-suspect pair comparison, the first column of the dictionary and $\mathbf{\Phi}_{\text{offender}}$ are replaced as needed.

$$\begin{pmatrix} \Phi_{\text{offender}}(1) \\ \Phi_{\text{offender}}(2) \\ \cdot \\ \Phi_{\text{offender}}(m) \end{pmatrix} = \begin{pmatrix} \Phi_{\text{suspect}}(1) & \Phi_{\text{UBM}}(1)_1 & \cdot & \Phi_{\text{UBM}}(1)_{n-1} \\ \Phi_{\text{suspect}}(2) & \Phi_{\text{UBM}}(2)_1 & \cdot & \Phi_{\text{UBM}}(2)_{n-1} \\ \cdot & \cdot & \cdot & \cdot \\ \Phi_{\text{suspect}}(m) & \Phi_{\text{UBM}}(m)_1 & \cdot & \Phi_{\text{UBM}}(m)_{n-1} \end{pmatrix} \times \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ x_n \end{pmatrix} \tag{7}$$

3.2 Sparse regression solution techniques

3.2.1 Least squares (LS)

As mentioned in the introduction, there are numerous approaches by which the unknown vector of coefficients $\mathbf{x}$ in Equation 7 can be solved. One is to treat the sparse representation problem in Equation 7 as a LS problem, i.e. minimizing the mean squared error of our estimate $\mathbf{x}$, with the familiar closed form solution.

$$\mathbf{x} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{\Phi}_{\text{offender}} \tag{8}$$

3.2.2 ℓ_1 and ℓ_2 norm minimization

In addition to the LS closed form solution, two well-established algorithms were considered in this study. In particular, the gradient projection algorithm as proposed in [40] was considered, which solves the convex unconstrained optimization problem as in Equation 9 with τ, a non-negative parameter empirically defined as 0.01 in this study.

$$\min_{\mathbf{x}} \frac{1}{2}\left\|\mathbf{\Phi}_{\text{offender}} - \mathbf{Ax}\right\|_2^2 + \tau\|\mathbf{x}\|_1 \tag{9}$$

Another algorithm, which utilizes efficient coordinate descent methods for fitting the entire lasso or elastic-net regularization path for linear regression, logistic regression and multinomial regression models, was proposed and detailed in [41]. It solves for the regression problem using ℓ_1 penalty (lasso), ℓ_2 penalty (ridge regression) or a mixture of the two (i.e. the elastic net) with a tuning parameter $0 \le \alpha \le 1$.

The elastic net solves for the following regression problem [41]

$$\min_{(\beta_0,\mathbf{x})\in\mathbb{R}^{n+1}}\left[\frac{1}{2m}\sum_{i=1}^{m}\left(\mathbf{\Phi}_{\text{offender}}(i)-\beta_0-\mathbf{A}_i\mathbf{x}\right)^2+\lambda \mathrm{P}_\alpha(\mathbf{x})\right] \tag{10}$$

$$\mathrm{P}_\alpha(\mathbf{x})=(1-\alpha)\frac{1}{2}\|\mathbf{x}\|_2^2+\alpha\|\mathbf{x}\|_1 \tag{11}$$

Note that $\mathbf{A}_i$ represents the ith row of the dictionary $\mathbf{A}$ from Equation 7, defined as

$$\mathbf{A}_i=\left[\Phi_{\text{suspect}}(i)\ \ \Phi_{\text{UBM}}(i)_1\ \ \Phi_{\text{UBM}}(i)_2\ldots\ \Phi_{\text{UBM}}(i)_{n-1}\right] \tag{12}$$

The penalty defined in Equation 11 is a compromise between the ridge-regression penalty (α = 0) and the lasso penalty (α = 1), and α was varied between these two values in the current study to evaluate on its effect on FVC system performance. The penalty parameter λ in Equation 10 was empirically defined as 0.01 in all of our experiments.

3.3 Sparse regression scoring methods

The four scoring methods discussed in Section 2.3 were compared, namely the *direct parameter* $x(1)$ $s_{x(1)}$, ℓ_1 *norm ratio* and ℓ_2 *residual ratio* introduced in [5] and *background normalized (BNorm)* ℓ_2 *residual* criterion introduced in [6]:

$$s_{\ell_1\text{norm}}=\frac{\|\delta_1(\mathbf{x})\|_1}{\|\mathbf{x}\|_1} \tag{13}$$

$$s_{\ell_2\text{residual}}=\frac{\left\|\mathbf{\Phi}_{\text{offender}}-\mathbf{A}\left(\sum_{i=2}^{n}\delta_i(\mathbf{x})\right)\right\|_2}{\|\mathbf{\Phi}_{\text{offender}}-\mathbf{A}\delta_1(\mathbf{x})\|_2} \tag{14}$$

$$s_{\text{Bnorm }\ell_2}=\frac{-\|\mathbf{\Phi}_{\text{offender}}-\mathbf{A}\delta_1(\mathbf{x})\|_2-\frac{1}{n-1}\sum_{j=2}^{n}\phi_j}{\sqrt{\frac{1}{n-2}\sum_{i=2}^{n}\left(\phi_i-\frac{1}{n-1}\sum_{j=2}^{n}\phi_j\right)^2}} \tag{15}$$

$$\phi_{j,j=2:n}=-\left\|\mathbf{\Phi}_{\text{offender}}-\mathbf{A}\delta_j(\mathbf{x})\right\|_2 \tag{16}$$

where

$$\delta_i(\mathbf{x})=\begin{cases}x(j) & ,\text{if } i=j\\ 0 & ,\text{if } i\neq j\end{cases} \tag{17}$$

3.4 Pairwise nearest neighbour (PNN)

The pairwise nearest neighbour (PNN) technique is a simple and well-known mathematical procedure that employs a distance metric based on a calculation of distances between all pairs of input data. In this study, for each pair of offender-suspect comparison, the Euclidean distance between the offender supervector $\mathbf{\Phi}_{\text{offender}}$ and the first column of the dictionary $\mathbf{A}$ which is the suspect supervector $\mathbf{\Phi}_{\text{suspect}}$ was used as the denominator in the score calculation, denoted as d_{suspect}. Similarly, the Euclidean distances between the offender supervector $\mathbf{\Phi}_{\text{offender}}$ and the second to the last columns of the dictionary $\mathbf{A}$ were determined. Three cases for evaluating these Euclidean distances between offender and background speaker supervectors are investigated in this study: namely, we find the minimum, mean and maximum of these Euclidean distances and they are subsequently used as the numerator in the score calculation, i.e.

$$s_{\text{PNN\_min}}=\frac{d_{\text{UBM, min}}}{d_{\text{suspect}}} \tag{18}$$

$$s_{\text{PNN\_mean}}=\frac{d_{\text{UBM,mean}}}{d_{\text{suspect}}} \tag{19}$$

$$s_{\text{PNN\_max}}=\frac{d_{\text{UBM, max}}}{d_{\text{suspect}}} \tag{20}$$

4 Methodology

4.1 Database of 60 female speakers of standard Chinese

The first database used in the present study is available from http://databases.forensic-voice-comparison.net/. It consists of voice recordings of 60 female speakers of Standard Chinese (i.e. Mandarin/Putonghua). Each speaker was recorded twice with each recording approximately 10 min long. All speakers were first-language speakers of Standard Chinese from Northeastern China, aged between 23 and 45. The nature of the speech was spontaneous, collected from an information exchange task over a telephone. Recordings were made at 44.1 kHz 16-bit using flat-frequency response lapel microphones (Sennheiser MKE 2 P-C) together with the use of an external soundcard (Roland® UA-25 EX) under studio-clean conditions. Furthermore, each of the approximately 10-min length recording was also post-processed using SoundLabeller [46] to remove of silence segments such that only the speech-active segments of about 2 to 5 min are used in our experiments. For more details on the database, readers may refer to section 6.3.1 of [47].

The data collection protocol for this database used in our study was an attempt to produce data highly typical of forensic case work conditions, refer to [48] for details of the protocol. For more details on database selection for FVC systems, readers may refer to discussions in [9,13] and section 2.4.1 and section 2.11 in [47].

A degraded mobile-to-landline version of the high-quality recordings of these 60 female speakers was also created by transmitting these high-quality recordings through a typical mobile-landline telephone transmission system. For implementation details of the setup of these

degraded versions of recordings, readers may refer to [49]. The effect of the mobile-telephone system on the original speech signal is often associated with data compression which can vary from moment to moment and can result in different transmission rates in the range of 4.75 to 12.20 kbits/s. The mobile-telephone system has a bandpass filter with a lower limit of 100 Hz to an upper limit of up to 3.6 kHz, whereas a landline-telephone system has a bandpass filter in the range of 300 to 3,400 Hz together with lossless compression and decompression algorithms with 64 kbits/s transmission rate [50].

A mismatched condition was further created by using a combination of high-quality studio-clean and degraded mobile-landline recordings in the forensic voice comparison system. In particular, the mismatch setting that was employed involves having background database and suspect data from both development and test databases comprised of studio-clean recordings and offender data from both development and test databases comprised of degraded mobile-landline recordings. This mismatched setting is more relevant for forensic applications since in a typical forensic casework, it is very likely and forensically realistic to have an offender voice recording collected from a telephone intercept which has an inherently degraded speech recording caused by the telephone-transmission system. Moreover, it is also very likely and forensically realistic to have suspect recordings collected at a different quality in comparison with the offender recordings as suspect recordings are typically being recorded in a controlled environment (such as a police interview) and with high-quality direct microphone.

The background, development and test databases were allocated evenly in three partitions, with each having 20 speakers. In particular, the initial 20 speakers (identification numbers: 01 to 04, 09 to 20, 22, 25, 26, 28), the next 20 speakers (29 to 48) and the last 20 speakers (49 to 68) were used for the background, development and test databases. Moreover, in the current study, a cross validation experiment was also performed by permuting the composition of the background, development and test databases, creating six permutations.

4.2 Database of 90 male speakers of standard Chinese

Another database of voice recordings of male speakers of Standard Chinese (i.e. Mandarin/Putonghua) was evaluated in this study. The male speaker database has a total of 90 speakers, with each speaker having two recordings. Apart from the gender difference, all other aspects of the male recordings such as the nature of the recordings, duration and recording conditions, equipment used for collection and post-processing of the recordings were exactly the same as those for the female recordings. For details of the data collection protocol, refer to [48].

The background, development and test databases were allocated evenly in three partitions, with each having 30 speakers. In particular, the initial 30 speakers (identification numbers: 01 to 30), the next 30 speakers (31 to 60) and the last 30 speakers (61 to 90) were used for the background, development and test databases. Similarly to the female database, a cross validation experiment was also performed by permuting the composition of the background, development and test databases, creating six permutations.

4.3 Forensic voice comparison system configuration

All automatic FVC systems used in the present study were built based on all speech-active segments within each recording of the 60 female speakers or 90 male speakers.

The baseline automatic FVC system, based on Gaussian Mixture Model - Universal Background Model (GMM-UBM) [1-3], had $K = 512$ mixture components. All automatic FVC systems employed $d = 32$ dimensional mel-frequency cepstral coefficients (MFCCs) [51-53] (16 static coefficients and 16 delta coefficients [54]) extracted from 20-ms frames overlapping by 10 ms with a 20-ms Hamming window [51]. Feature normalization was performed via cumulative distribution mapping [55], and no channel or session compensation technique was applied.

For the regression-based techniques, the supervector had dimension $m = K \times d = 16{,}384$. There were 40 recordings (two recordings per speaker with 20 UBM speakers) assigned for UBM training for the 60 female speaker database, whereas there were 58 recordings (two recordings per speaker with 30 UBM speakers, excluding session 2 of both speakers 85 and 86 as they were lower quality recordings mis-recorded at a sampling frequency of 11.025 kHz) assigned for UBM training for the 90 male speaker database. Each of these recordings was subsequently adapted from the trained UBM to derive a conventional adapted GMM and then converted to GMM mean supervector. The dictionary therefore had a total of $n = 41$ supervectors (one $\mathbf{\Phi}_{\text{suspect}}$ supervector and 40 $\mathbf{\Phi}_{\text{UBM}}$ supervectors) for the 60 female speaker database and $n = 59$ supervectors (one $\mathbf{\Phi}_{\text{suspect}}$ supervector and 58 $\mathbf{\Phi}_{\text{UBM}}$ supervectors) for the 90 male speaker database.

To solve the regression problems as detailed in Section 3.2.2, many variants of the state-of-the-art solvers for the sparse regression problem in Equation 7 are available publicly. In particular, we implemented the publicly available gradient projection for sparse reconstruction (GPSR) solver (http://www.lx.it.pt/~mtf/GPSR/) for the gradient projection algorithm as detailed in [40] and we

implemented the publicly available Glmnet solver (http://www-stat.stanford.edu/~tibs/glmnet-matlab/) for the coordinate descent algorithm as detailed in [40].

4.4 Score to likelihood ratio conversion

For conversion of a score to an interpretable likelihood ratio via an affine transform, logistic regression calibration with equal priors can be used [14,22,23]. Same-origin and different-origin scores, s_{dev}, from the development database are used to train the calibration weights, i.e. the intercept and regression coefficient of the logistic regression model, and subsequently these calibration weights can then be used to calibrate scores from the test database. The pooled procedure for calculating the calibration weights was adopted (refer to [19] for details) in this paper. For a detailed tutorial on logistic regression calculation in converting a score to an interpretable likelihood ratio, refer to [12].

4.5 Evaluation metrics

The validity and reliability (i.e. accuracy and precision) of the forensic voice comparison systems employed in the current paper were evaluated using the log-likelihood-ratio cost, C_{llr} (mean procedure [19]) as proposed by Brümmer [14], and 95% credible interval (CI) as proposed by Morrison et. al. [9,20,56,57], denoted as 95% CI (parametric procedure and with orders of magnitude expressed in log base ten). The log-likelihood-ratio cost has been applied in numerous FVC studies as seen in [23,58-60]. It should be noted that for all the above metrics, smaller values indicate better performance.

Tippett plots, which provide a graphical representation of the cumulative distribution function of log-likelihood ratios for same-origin and different-origin hypotheses [9,61], were also used in current study.

5 Results and discussion

5.1 Regression and scoring methods

The pooled values across the six permutations for C_{llr} mean and 95% CI based on the different sparse representation regression solutions and scoring methods evaluated on the 60 female speaker database and 90 male speaker database under studio clean conditions are given in Figure 5 (top and bottom row, respectively).

Examining the results from Figure 5, when considering systems that performed well irrespective of database composition, there were two systems (as highlighted in dashed red circle in Figure 5a,d) that performed consistently better in comparison with the baseline GMM-UBM and SVM systems: gradient projection with $s_{x(1)}$ and LS with $s_{x(1)}$. Further, there were two systems (as highlighted in dashed red circle in Figure 5e,h) that showed equally promising results with only slight degradation in validity in comparison with the baseline SVM system when tested with the 90 male speaker database: gradient projection with $s_{\ell_1\text{norm}}$ and LS with $s_{\ell_1\text{norm}}$.

The relative improvements in pooled results of C_{llr} mean and 95% CI across the six permutations for these four best systems: gradient projection, $s_{x(1)}$, LS, $s_{x(1)}$, gradient projection, $s_{\ell_1\text{norm}}$ and LS, $s_{\ell_1\text{norm}}$ over the baseline GMM-UBM and SVM systems were substantial, evaluated on both the 60 female speaker and 90 male speaker databases under studio-clean conditions as tabulated in Table 1. In particular, the four systems showed improvements over the GMM-UBM baseline system in the order of 45% to 80% improvement in terms of the C_{llr}-mean metric and in the order of 15% to 45% improvement in terms of the 95% CI metric. Similarly, the four systems also showed substantial improvements over the SVM baseline system in the order of 10% to 75% improvement in terms of the C_{llr}-mean metric albeit a slight degradation of about 19% for the gradient projection, $s_{\ell_1\text{norm}}$ and LS, $s_{\ell_1\text{norm}}$ systems when evaluated on the 90 male speaker database and in the order of 5% to 30% improvement in terms of the 95% CI metric.

One possible explanation for the good performance achieved by the familiar least squares regression technique in comparison with the state-of-the-art sparse regression techniques such as ℓ_1 and ℓ_2 norm minimization that have featured among speaker recognition studies is that the discriminative nature of sparse regression techniques implies an indirect manipulation of the weights in the entries of the regression solution to ensure sparseness of the solution in such a way that this manipulation could be disrupting the original or intrinsic weightings of the individual speakers. As an example, for the case of ℓ_1 norm minimization, the entries of the regression solution were forced to contain mostly zero entries, that is the technique forces the weights of the speakers from background set who are least similar to the test speaker to zero, and thus ignores the contribution of those speakers in the likelihood computation. In other words, ℓ_1 norm minimization ignores those speakers from the background set who are least similar in comparison to the test speaker by forcing their weights to zero to ensure sparseness. This effectively reduces the number of background speakers and only the most similar speakers to the test speaker are used for typicality evaluation in the likelihood ratio calculation; and subsequently, an undesirable tighter restriction on the test conditions for typicality is imposed and hence the system performance will be poorer in this respect. However, in the least squares regression case, all the speakers in the background set are included in the likelihood ratio calculation and their weights are not being altered directly to ensure

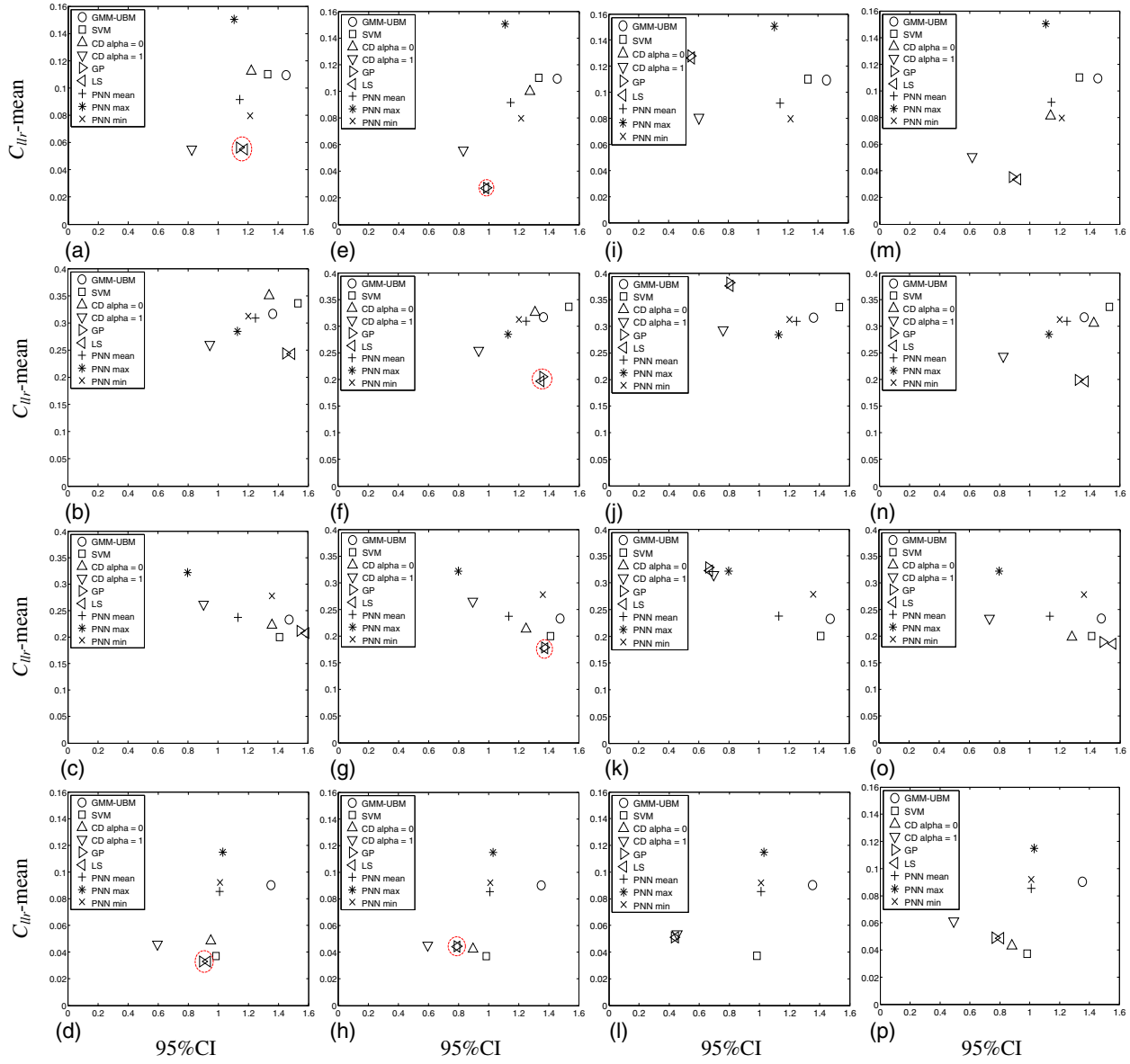

Figure 5 Pooled results for validity and reliability across the six permutations. Pooled results for validity (C_{llr} mean) and reliability (95% CI) across the six permutations for the various sparse regression solutions utilizing the $s_{x(1)}$ (i.e. **a-d**), $s_{\ell 1}$ (i.e. **e-h**), $s_{\ell 2}$ (i.e. **i-l**), and $s_{\text{Bnorm}\ell 2}$ (i.e. **m-p**) scoring methods, evaluated on the 60 female speaker database under studio-clean (top row), degraded mobile to landline (second row from top) and mismatched (third row from top) conditions and 90 male speaker database under studio-clean (bottom row) conditions. The various configurations are denoted as follows: circle: GMM-UBM, square: s_{SVM}, upward-pointing triangle: coordinate descent (CD) $\alpha = 0$, downward-pointing triangle: coordinate descent (CD) $\alpha = 1$, right-pointing triangle: gradient projection (GP), left-pointing triangle: LS, plus: $s_{\text{PNN\_mean}}$, asterisk: $s_{\text{PNN\_max}}$, cross: $s_{\text{PNN\_min}}$. Note that the GMM-UBM, s_{SVM}, $s_{\text{PNN\_mean}}$, $s_{\text{PNN\_max}}$ and $s_{\text{PNN\_min}}$ configurations do not change across the plots in the same row, as these do not use sparse representation regression methods.

sparseness. This means that in the least squares regression case, all the entries of the regression solution preserve the original speaker weightings, and all speakers in the background set in this case are used for typicality evaluation in the likelihood ratio calculation; and hence, there is no restriction in terms of test conditions for typicality in comparison with the ℓ_1 norm minimization case. The performance of the system for the least square regression case therefore should be better since we are evaluating the system with no constraint on typicality. This effect was also empirically verified by varying the non-negative parameter τ of the gradient projection for sparse reconstruction (GPSR) solver (i.e. in Equation 9) to have values greater than 0.01 that is defined in this paper. Experimental results from this setup showed an inverse relationship in that as τ was increased, the performance of the FVC system was seen to become poorer. This result was in agreement with our previous discussion in that: if the

Table 1 Improvements over the GMM-UBM and SVM baseline systems in relative percentage terms

		60 female speaker database		90 male speaker database	
		Validity (C_{llr} mean)	Reliability (95% CI)	Validity (C_{llr} mean)	Reliability (95% CI)
GMM-UBM	Gradient projection, $s_{x(1)}$	48.5%	21.5%	63.5%	33.8%
	LS, $s_{x(1)}$	49.7%	19.1%	63.6%	31.3%
	Gradient projection, $s_{\ell_1\text{norm}}$	75.0%	32.3%	50.5%	41.5%
	LS, $s_{\ell_1\text{norm}}$	75.1%	32.7%	50.8%	41.5%
s_{SVM}	Gradient projection, $s_{x(1)}$	48.9%	14.1%	11.6%	9.1%
	LS, $s_{x(1)}$	50.1%	11.5%	11.8%	5.7%
	Gradient projection, $s_{\ell_1\text{norm}}$	75.1%	26.0%	−19.9%	19.7%
	LS, $s_{\ell_1\text{norm}}$	75.3%	26.4%	−19.1%	19.7%

For the pooled results of validity (C_{llr} mean) and reliability (95% CI) across the six permutations for the four best systems: gradient projection, $s_{x(1)}$; LS, $s_{x(1)}$; gradient projection, $s_{\ell 1\text{norm}}$; and LS, $s_{\ell 1\text{norm}}$ evaluated on both databases under *studio-clean conditions*.

parameter τ was increased, we put more emphasis on the ℓ_1 norm minimization which then leads to a poorer FVC performance; and if parameter τ was decreased, we put more emphasis on the least square regression technique which then leads to a better FVC performance.

5.2 Degraded and mismatched conditions

The corresponding results for the degraded mobile to landline and mismatched conditions are displayed in the second and third rows from the top in Figure 5. These results for degraded and mismatched conditions were evaluated on the female speaker database solely as no degraded version of the male database was available.

Considering the results from Figure 5, there were two systems (highlighted by the dashed red circle in Figure 5f,g) that performed consistently better under degraded and mismatched conditions in comparison with the baseline GMM-UBM and SVM systems: gradient projection with $s_{\ell_1\text{norm}}$ and LS with $s_{\ell_1\text{norm}}$. In particular, under both degraded mobile and landline and mismatched conditions evaluated, the two systems showed substantial improvements over the GMM-UBM baseline system (as tabulated in Table 2) in the order of 20% to 40% in terms of the C_{llr}-mean metric and in the order of 0% to 7% in terms of the 95% CI metric. Similarly, the two systems also showed substantial improvements over the SVM baseline system in the order of 10% to 45% in terms of the C_{llr}-mean metric and in the order of 2% to 13% in terms of the 95% CI metric. The other two systems that performed well under studio-clean conditions, gradient projection with $s_{x(1)}$ and LS with $s_{x(1)}$, were observed to perform more poorly under degraded (Figure 5b) and mismatched (Figure 5c) conditions relative to the two systems: gradient projection with $s_{\ell_1\text{norm}}$ and LS with $s_{\ell_1\text{norm}}$.

The experiments also demonstrated that under mismatched conditions, the relative performance of the SVM baseline system can be considerably better than the GMM-UBM baseline system than for the case of under degraded conditions. This was depicted in the third row from top in Figure 5 which shows the SVM baseline system performing much better than GMM-UBM system under mismatched conditions. An implication of this is that the SVM baseline system may be more resilient to undesirable channel artefacts such as recording noise and in particular could be a more robust system than a GMM-UBM system under the scenario where there is a mismatch in recording conditions of the suspect and offender recordings.

5.3 Tippett plot results

The actual LR distributions when H_{so} (blue lines) and H_{do} (red lines) are respectively true across the six permutations

Table 2 Improvements over the GMM-UBM and SVM baseline systems

		60 female speaker database (degraded conditions)		60 female speaker database (mismatched conditions)	
		Validity (C_{llr} mean)	Reliability (95% CI)	Validity (C_{llr} mean)	Reliability (95% CI)
GMM-UBM	Gradient projection, $s_{\ell_1\text{norm}}$	35.1%	0.3%	23.1%	7.0%
	LS, $s_{\ell_1\text{norm}}$	37.9%	1.1%	24.0%	6.8%
s_{SVM}	Gradient projection, $s_{\ell_1\text{norm}}$	38.9%	11.4%	10.6%	2.8%
	LS, $s_{\ell_1\text{norm}}$	41.5%	12.1%	11.6%	2.5%

In terms of percentage for the pooled results of validity (C_{llr} mean) and reliability (95% CI) across the six permutations for the two systems: gradient projection, $s_{\ell 1\text{norm}}$; and LS, $s_{\ell 1\text{norm}}$ evaluated on 60 female speaker database under *degraded and mismatched conditions*.

based on the two baseline systems: GMM-UBM and SVM and the two best systems: gradient projection with $s_{\ell_1\text{norm}}$ and LS with $s_{\ell_1\text{norm}}$, evaluated on the 60 female speaker database under studio-clean (top row), degraded mobile to landline (second row from top) and mismatched (third row from top) conditions and 90 male speaker database under studio-clean (bottom row) conditions are given in Figure 6.

Comparing the studio clean (first row in Figure 6) recording conditions with noisy conditions (second and third rows from top in Figure 6) on the female speaker database, the overall trend observed for all the four systems is that Tippett plots become narrower in separation between cumulative distribution plots for H_{so} (blue lines) and H_{do} (red lines), indicating poorer performance with poorer recording conditions as expected.

Comparing the performance of the four systems on any one particular recording condition, however, reveals that the two systems based on gradient projection with $s_{\ell_1\text{norm}}$

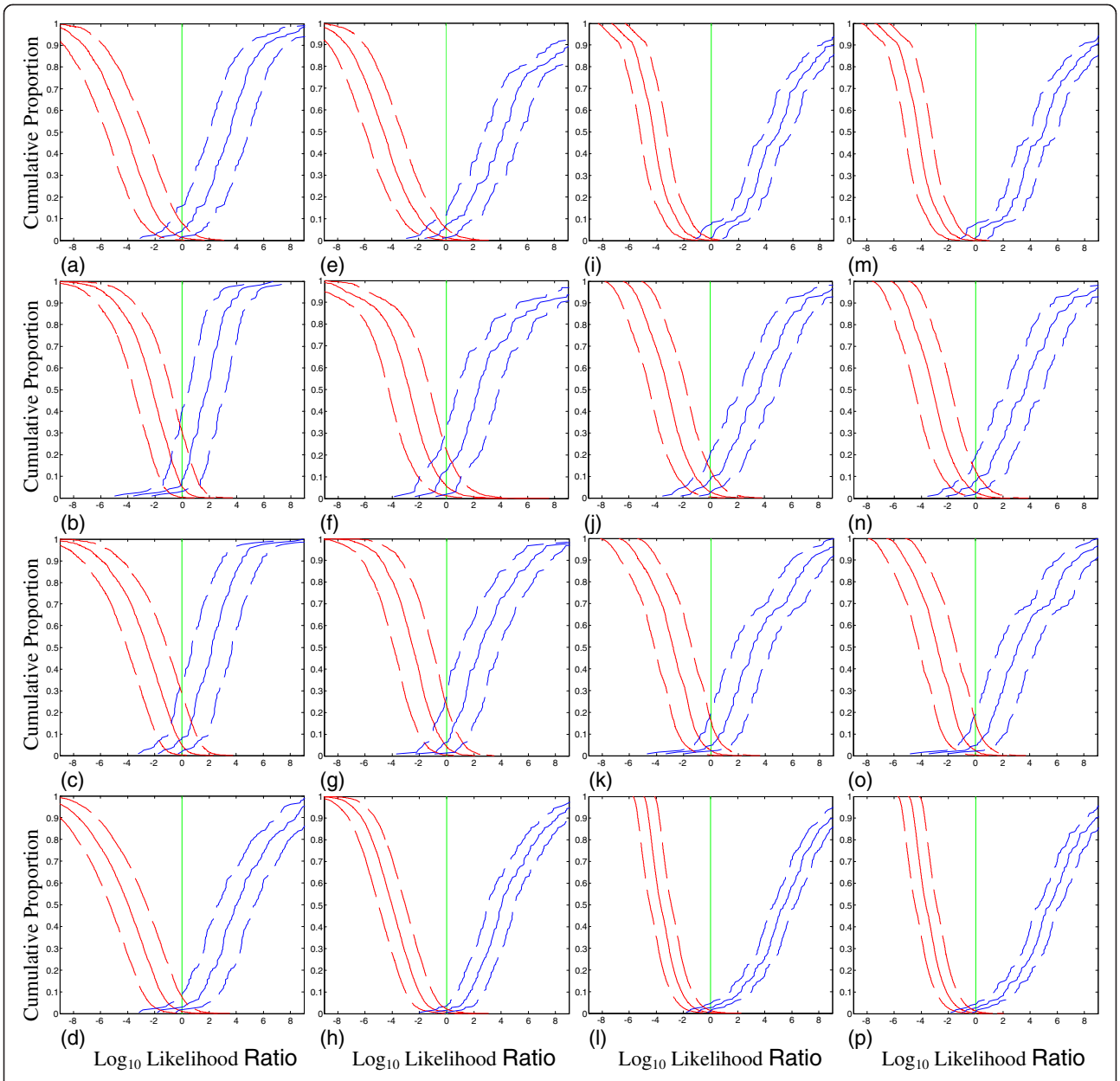

Figure 6 Tippett plots showing the actual LR distributions. Tippett plots showing the actual LR distributions when H_{so} (blue lines) and H_{do} (red lines) are respectively true for different systems evaluated on the 60 female speaker database under studio-clean (top row), degraded mobile to landline (second row from top) and mismatched (third row from top) conditions and 90 male speaker database under studio-clean (bottom row) conditions. The solid lines represent the C_{llr} values, and dashed lines to the left and right of the solid lines show the estimates of the 95% CI values. GMM-UBM: **(a-d)**. SVM: **(e-h)**. Gradient projection, $s_{\ell 1\text{norm}}$: **(i-l)**. LS, $s_{\ell 1\text{norm}}$: **(m-p)**.

and LS with $s_{\ell_1\text{norm}}$ respectively perform much better than baseline systems with much wider separation between the blue and red solid lines (e.g. compare Figure 6m with Figure 6a) for the cumulative distribution plots for H_{so} (blue lines) and H_{do} (red lines) indicating better system accuracy and much narrower dashed lines to the left and right of the solid lines indicating better precision or reliability based on the 95% CI values.

This improvement was observed to primarily stem from the same-origin comparisons as the cumulative distribution plot for H_{so} (blue lines) moved further to the right favourably and not from the different-origin comparisons as the cumulative distribution plot for H_{do} (red lines) moved slightly to the left unfavourably (this can be seen, for example, in Figure 6m, the solid red line reaches full saturation point at $\log_{10}$(LR) of −6, whereas the solid red line in Figure 6a reaches it at $\log_{10}$(LR) of approximately −9 and the more extreme case in Figure 6p compared with Figure 6d).

This implies that scores generated from same-origin comparisons are substantially higher valued if based on the supervector regression method than conventional GMM-UBM or SVM methods. These results therefore give a clear indication to the strength of supervector regression method in that it is able to generate much stronger same-origin comparison scores based on regression than the conventional GMM-UBM approach based on a ratio of probability densities.

From a practical standpoint, this translates to the supervector regression method giving a much higher valued likelihood ratio (i.e. a much stronger strength of evidence) by generating a more accurate and greater support for the same-origin hypothesis than the different-origin hypothesis if the particular suspect on trial is indeed the offender.

The process of implementing the supervector regression method in an actual court case could be as follows. In practice, forensic scientists may only have one offender recording to compare against several suspect recordings. To generate meaningful likelihood ratios from this one offender recording with other suspect recordings, they perform database selection to collect a database of homogenous nature (as described in Section 4.1) suitable for the particular court trial and split the database into background, development and test databases for FVC system development. It is at this stage of system development that forensic scientists can choose the modelling stage to be based on the supervector regression method rather than the conventional GMM-UBM for better system accuracy and precision. After the FVC system has been properly calibrated (i.e. using logistic regression calibration from scores from development database as discussed in Section 4.4) and evaluated on the test database to be performing well based on the collected homogenous database, this FVC system is then ready to test on the actual unique offender recording with the suspect recording to generate a likelihood ratio to be presented in court as strength of evidence.

6 Conclusion

This paper has investigated the use of supervector regression methods in automatic FVC systems, for the specific database conditions that are relevant to forensic case work applications. In comparison with GMM-UBM- and SVM-based forensic-voice-comparison systems, supervector regression techniques consistently resulted in a large improvement in both validity and reliability. Among the many techniques considered in this study, the best was from the familiar least squares regression technique, combined with the ℓ_1 *norm ratio* scoring method. On both male and female databases under studio-clean conditions, substantial improvements from the least squares configuration relative to GMM-UBM baseline were observed. Similar substantial improvements were observed from the least squares configuration relative to SVM baseline with only a slight degradation in validity over the SVM baseline in one condition tested; that of the 90 male speaker database. Evaluation under degraded mobile to landline and mismatched conditions again demonstrated that LS with $s_{\ell_1\text{norm}}$ performed well and gave consistent gains in both validity and reliability over the GMM-UBM and SVM baselines. From the practical viewpoint, supervector regression was demonstrated to be capable of generating improved strength of evidence by providing a more accurate and greater support for the same-origin hypothesis than the different-origin hypothesis if the suspect on trial is the true offender in a court case as compared with GMM-UBM or SVM systems. As future work, other speech databases that are relevant to forensic applications could be tested to validate our experimental observations.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The authors received financial support from the Australian Research Council, through Linkage Project LP100200142, and from NICTA, which is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence programme. Thanks to Geoffrey Stewart Morrison, Ewald Enzinger, Eliathamby Ambikairajah and Jia Min Karen Kua for comments on an earlier draft of the paper and Cuiling Zhang for provision of the databases.

Received: 19 June 2014 Accepted: 23 December 2014
Published online: 24 February 2015

References

1. DA Reynolds, RC Rose, Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE. Trans. Audio Speech Lang. Process **3**(1), 72–83 (1995)
2. DA Reynolds, TF Quatieri, RB Dunn, Speaker verification using adapted Gaussian mixture models. Digit. Signal Process **10**(1–3), 19–41 (2000)

3. DA Reynolds, Speaker identification and verification using Gaussian mixture speaker models. Speech Comm. **17**(1–2), 91–108 (1995)
4. JMK Kua, E Ambikairajah, J Epps, R Togneri, *Speaker Verification Using Sparse Representation Classification, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011
5. M Li, S Narayanan, *Robust Talking Face Video Verification Using Joint Factor Analysis and Sparse Representation on GMM Mean Shifted Supervectors, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011
6. M Li, X Zhang, Y Yan, S Narayanan, Speaker verification using sparse representations on total variability i-vectors, in *Proc. Interspeech, Florence, Italy*, 2011, pp. 2729–2732
7. J Gonzalez-Rodriguez, A Drygajlo, D Ramos-Castro, M Garcia-Gomar, J Ortega-Garcia, Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. Comput. Speech Lang. **20**(2–3), 331–355 (2006)
8. GS Morrison, Forensic voice comparison and the paradigm shift. Sci. Justice **49**(4), 298–308 (2009)
9. GS Morrison, Forensic Voice Comparison, in *Expert Evidence (Ch. 99)*, ed. by I Freckelton, H Selby (Thomson Reuters, Sydney, Australia, 2010)
10. P Rose, Technical forensic speaker recognition: evaluation, types and testing of evidence. Comput. Speech Lang. **20**(2–3), 159–191 (2006)
11. P Rose, *Forensic speaker identification* (Taylor & Francis, London, 2002)
12. GS Morrison, Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. Aust. J. Forensic Sci. **45**(2), 173–197 (2012)
13. GS Morrison, F Ochoa, T Thiruvaran, Database selection for forensic voice comparison, in *Proc. of Odyssey 2012: The Language and Speaker Recognition Workshop, Singapore International Speech Communication Association, 62–77*, 2012
14. N Brümmer, J du Preez, Application-independent evaluation of speaker detection. Comput. Speech Lang. **20**(2–3), 230–275 (2006)
15. WM Campbell, DE Sturim, DA Reynolds, Support vector machines using GMM supervectors for speaker verification. IEEE. Signal Process Lett. **13**(5), 308–311 (2006)
16. T Kinnunen, H Li, An overview of text-independent speaker recognition: from features to supervectors. Speech Comm. **52**(1), 12–40 (2010)
17. J Gonzalez-Rodriguez, D Garcia-Romero, M García-Gomar, D Ramos-Castro, J Ortega-Garcia, *Robust Likelihood Ratio Estimation in Bayesian Forensic Speaker Recognition, in Eighth European Conference on Speech Communication and Technology*, 2003
18. D Meuwly, A Drygajlo, *Forensic Speaker Recognition Based on a Bayesian Framework and Gaussian Mixture Modelling (GMM), in A Speaker Odyssey - The Speaker Recognition Workshop*, 2001
19. GS Morrison, T Thiruvaran, J Epps, *An Issue in the Calculation of Logistic-Regression Calibration and Fusion Weights for Forensic Voice Comparison, in Proc. Of the 13th Australasian International Conference on Speech Science and Technology*, 2010
20. GS Morrison, T Thiruvaran, J Epps, *Estimating the precision of the likelihood-ratio output of a forensic-voice-comparison system, in Proc. of Odyssey 2010: The Language and Speaker Recognition Workshop* (Brno, Czech Republic, 2010)
21. C Zhang, GS Morrison, T Thiruvaran, *Forensic voice comparison using Chinese /iau/, in Proc. of the 17th International Congress of Phonetic Sciences, Hong Kong, China, 2280–2283*, 2011
22. G Morrison, Robust version of train_llr_fusion. m from Niko Brümmer's FoCal Toolbox (2009), http://geoff-morrison.net/#TrainFus. Software release 2009-07-02
23. GS Morrison, Y Kinoshita, *Automatic-Type Calibration of Traditionally Derived Likelihood Ratios: Forensic Analysis of Australian English/o/Formant Trajectories, in Proceedings of Interspeech*, 2008
24. D Meuwly, *Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique, Université de Lausanne*, 2001
25. A Drygajlo, Automatic Speaker Recognition for Forensic Case Assessment and Interpretation, in *Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism*, ed. by A Neustein, HA Patil (New York, Springer, 2011), pp. 21–39
26. T Becker, M Jessen, C Grigoras, Forensic speaker verification using formant features and Gaussian mixture models, in Proc. Interspeech, (Brisbane, Queensland, Australia, 2008), p. 1505–1508
27. GS Morrison, A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model-universal background model (GMM-UBM). Speech Comm. **53**(2), 242–256 (2011)
28. WM Campbell, JP Campbell, DA Reynolds, E Singer, PA Torres-Carrasquillo, Support vector machines for speaker and language recognition. Comput. Speech Lang. **20**(2–3), 210–229 (2006)
29. N Dehak, P Kenny, R Dehak, O Glembek, P Dumouchel, L Burget, V Hubeika, F Castaldo, *Support Vector Machines and Joint Factor Analysis for Speaker Verification, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009
30. P Kenny, G Boulianne, P Ouellet, P Dumouchel, Joint factor analysis versus eigenchannels in speaker recognition. IEEE Trans Audio Speech Lang Process **15**(4), 1435–1447 (2007)
31. P Kenny, P Ouellet, N Dehak, V Gupta, P Dumouchel, A study of interspeaker variability in speaker verification. IEEE Trans Audio Speech Lang Process **16**(5), 980–988 (2008)
32. S Cumani, N Brümmer, L Burget, P Laface, *Fast Discriminative Speaker Verification in the i-Vector Space, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011
33. N Dehak, PJ Kenny, R Dehak, P Dumouchel, P Ouellet, Front-end factor analysis for speaker verification. IEEE Trans Audio Speech Lang Process **19**(4), 788–798 (2011)
34. WM Campbell, *Generalized Linear Discriminant Sequence Kernels for Speaker Recognition, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002
35. RE Fan, PH Chen, CJ Lin, Working Set selection using second order information for training support vector machines. J. Mach. Learn. Res. **6**, 1889–1918 (2005)
36. CC Chang, CJ Lin, LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(3), 27 (2011)
37. EJ Candès, *Compressive sampling, in Proc. of the International Congress of Mathematicians: Madrid, August 22–30, 2006: invited lectures*, 2006
38. DL Donoho, For most large underdetermined systems of linear equations the minimal l1-norm Solution is also the Sparsest Solution. Commun. Pure Appl. Math. **59**(6), 797–829 (2006)
39. J Wright, AY Yang, A Ganesh, SS Sastry, Y Ma, Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell **31**(2), 210–227 (2009)
40. MAT Figueiredo, RD Nowak, SJ Wright, Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. IEEE J Sel Top Signal Process **1**(4), 586–597 (2007)
41. J Friedman, T Hastie, R Tibshirani, Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33**(1), 1–22 (2010)
42. D Kanevsky, TN Sainath, B Ramabhadran, D Nahamoo, An analysis of sparseness and regularization in exemplar-based methods for speech classification, in Proc. Interspeech, (Makuhari, Chiba, Japan, 2010), p. 2842–2845
43. I Naseem, R Togneri, M Bennamoun, *Sparse Representation for Speaker Identification, in Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*, 2010
44. BC Haris, R Sinha, *Sparse Representation Over Learned and Discriminatively Learned Dictionaries for Speaker Verification, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012
45. V Boominathan, KSR Murty, *Speaker Recognition via Sparse Representations Using Orthogonal Matching Pursuit, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012
46. G Morrison, SoundLabeller: Ergonomically designed software for marking and labelling portions of sound files (2010), http://geoff-morrison.net. Release 2010-11-18
47. CC Huang, *Automatic Phonetic-Unit Selection and Modelling Techniques for Forensic Voice Comparison* (The University of New South Wales, Doctor of Philosophy, 2013)
48. GS Morrison, P Rose, C Zhang, Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. Aust J Forensic Sci **44**(2), 155–167 (2012)
49. CC Huang, J Epps, C Zhang, *An Investigation of Automatic Phonetic-Unit Selection for Forensic Voice Comparison, in Proceedings of the 14th Australasian International Conference on Speech Science and Technology, Sydney, Australia, 129–132*, 2012
50. BJ Guillemin, C Watson, Impact of the GSM mobile phone network on the speech signal - some preliminary findings. Int J Speech Lang Law **15**(2), 193–218 (2008)

51. JR Deller, JG Proakis, JHL Hansen, *Discrete-Time Processing of Speech Signals* (Macmillan Publishing Company, New York, 1993)
52. S Davis, P Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process **28**(4), 357–366 (1980)
53. X Huang, A Acero, HW Hon, *Spoken Language Processing: a Guide to Theory, Algorithm, and System Development* (Prentice Hall, New Jersey, 2001)
54. S Furui, Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Transactions on Acoustics, Speech and. Signal Process. **34**(1), 52–59 (1986)
55. J Pelecanos, S Sridharan, *Feature warping for robust speaker verification, in Proc. of Speaker Odyssey* (The Speaker Recognition Workshop, Crete, Greece, 2001), pp. 213–218
56. GS Morrison, Measuring the validity and reliability of forensic likelihood-ratio systems. Sci. Justice **51**(3), 91–98 (2011)
57. GS Morrison, C Zhang, P Rose, An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. Forensic Sci. Int. **208**(1–3), 59–65 (2011)
58. J Gonzalez-Rodriguez, P Rose, D Ramos, DT Toledano, J Ortega-Garcia, Emulating DNA: rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. IEEE Trans Audio Speech Lang Process **15**(7), 2104–2115 (2007)
59. GS Morrison, Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. J. Acoust. Soc. Am. **125**(4), 2387–2397 (2009)
60. D Ramos-Castro, *Forensic Evaluation of the Evidence Using Automatic Speaker Recognition Systems* (Universidad autónoma de Madrid, Madrid, 2007)
61. D Ramos-Castro, J Gonzalez-Rodriguez, J Ortega-Garcia, *Likelihood Ratio Calibration in a Transparent and Testable Forensic Speaker Recognition Framework, in IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, 2006