Machine learning based speaker gender classification using transformed features

Ahmed, Ahmed I.; Chiverton, John; Ndzi, David L.; Al-Faris, Mahmoud

Published in:
2021 International Conference on Communication & Information Technology (ICICT)

DOI:
10.1109/ICICT52195.2021.9568452

Published: 26/10/2021

Document Version
Peer reviewed version

Ahmed, A. I., Chiverton, J., Ndzi, D. L., & Al-Faris, M. (2021). Machine learning based speaker gender classification using transformed features. In *2021 International Conference on Communication & Information Technology (ICICT)* IEEE. https://doi.org/10.1109/ICICT52195.2021.9568452

Machine Learning based Speaker Gender Classification using Transformed Features

Ahmed Isam Ahmed
Informatics and Telecommunication Public Company
Ministry of Communications
Mosul, Iraq
ahmed.ahmed5@myport.ac.uk

John P. Chiverton
School of Energy and Electronic Engineering
University of Portsmouth
Portsmouth, United Kingdom
john.chiverton@port.ac.uk

David L. Ndzi
School of Computing, Engineering and Physical Sciences
University of the West of Scotland
Paisely, United Kingdom
david.ndzi@uws.ac.uk

Mahmoud M. Al-Faris
Informatics and Telecommunication Public Company
Ministry of Communications
Mosul, Iraq
mahmoud.alfaris1@myport.ac.uk

***Abstract*—Speech and image processing are fundamental components of artificial intelligence technology. Speech processing can be deployed to acquire unique features of a person's voice. These can then be used for speaker identification in addition to gender and age classification. This paper studies the effect of relative degree of correlation in speech features on gender classification. To this end, gender classification performance is evaluated using orthogonally transformed speech features. The performance is then compared to the case when speech features are used without transformation. Two machine learning approaches are used in the evaluation. One of them primarily depends on Gaussian Mixture Models (GMM) and the other one uses Support Vector Machine (SVM). The results show that less correlated speech features, obtained after the orthogonal transformation, provide better classification performance.**

Keywords—principal component analysis, speaker gender classification, machine learning

I. INTRODUCTION

One of the goals of speaker recognition is gender classification [1]. In the training of a gender classifier, speech samples need to be pooled together from a population of males and females to train models based on the two genders. Then, in the detection phase, speakers' genders can be classified. The acquired information is useful for a variety of applications ranging from security, health and safety, to behavioural economics [2].

Data of recordings collected at call centres can be used to obtain statistical information about clients' age and gender which can be used in marketing plans. The use of closed-circuit television cameras (CCTV) with integrated microphones is also widespread. It is convenient to take advantage of such capabilities to implement context-aware ambient systems to automatically gather and provide information on gender composition to complement video images. Microphones can also be more easily deployed than cameras. As in CCTV, privacy issues are mostly related to how the data is handled and used, including whether it could be used to identify individuals [3].

Although gender classification based on speech processing has been widely investigated, further enhancements can still be made [4]. The most successful speaker recognition (identification and verification) technologies perform better when the analysis is gender dependent. A recent study in gender classification, [5], showed how the performance of speaker recognition is negatively affected when the gender of the speaker to be recognised is mis-classified. Cepstral features extracted from the speech signal, especially mel-frequency cepstral coefficients (MFCC), are predominantly used for this purpose alone or combined with other features such as f_0, the fundamental frequency of voice. MFCC has proven to be a useful transformation of the speech signal suitable for the purpose of speech and speaker recognition. It is fundamentally based on the decomposition of the speech spectrum using a filterbank which is designed to mimic the function of the human auditory system [2].

In this work, we focus on improving the use of MFCC in gender classification. In [6], it was reported that there is a correlation between the coefficients of MFCC to some degree that depends on the order of the coefficients. The superiority of classification performance with correlated or uncorrelated features has been investigated in some disciplines. In some cases, uncorrelated features presented enhanced classification performance as in [7] for signal detection and in [8] for multimedia data classification. This has motivated the investigation of the effect of using less correlated features derived from MFCC in gender-classification. A common transformation of MFCC into less correlated sets of features is through Principal Component Analysis (PCA) [9].

PCA is also used for dimensionality reduction in different fields, e.g. computer vision [10]. Dimensionality reduction can bring another benefit to a classification system by decreasing the computational requirements as a result of reduced feature

dimensionality. This can be very helpful in applications where a centralized processing system could have been configured to handle speaker recognition from a large number of inputs. Furthermore, in [11], the effects of feature selection and dimensionality reduction techniques were compared in terms of classification performance. In particular, PCA was shown to provide improvements for some types of data and classifiers in that work.

We propose to use principal component analysis (PCA) for the analysis of the universal variance of speech features, hence, not just encompassing individual speaker's variations. It appears from the literature that this has not been previously addressed for gender classification of speakers. The analysis is accomplished by applying PCA to a large number of feature vectors obtained from many speakers to produce one projection matrix. This projection matrix is then used to perform orthogonal transformation of feature vectors of training and test samples. This is computationally efficient because PCA does not need to be performed subsequently for each training and test speech sample.

The rest of the paper is organised as follows: Section II presents a review of related research. Section III presents the computational complexity and challenges of implementing a gender classification+6 system, whilst Section IV outlines the framework used in this paper. The results are presented in Section V which is followed by the conclusions in Section VI.

II. Related Work

Principal Component Analysis (PCA) has been used for different purposes in speaker recognition [1], generally for feature transformation as in [12], and also in the process of speaker modelling as in [13]. For feature transformation, it is mostly applied in traditional speaker recognition modelling methods, when the modelling for each speaker's feature vectors is performed independently from other speakers as in [14]. An example of those modelling methods is using a GMM to fit the feature vectors of each utterance of interest separately which incurred slow performance. PCA was also conducted separately for each utterance features.

Later approaches, such as the one proposed in [15], produce global models of feature vectors of a population of speakers, namely, Gaussian Mixture Models-Universal Background Models (GMM-UBM). These are then adapted using an utterance's features to model that utterance (speaker). Another example of global modelling is factor analysis [16]. In this paper, we propose to conduct principal component analysis at a global scale that does not theoretically affect the processing speed of the relatively fast speaker recognition systems. In other words, we perform PCA on a large amount of feature vectors of many speakers and then use the resultant universal principal components to project the features of the training and test utterances into the PCA space. The additional load in the recognition system is very light, because it is a simple matrix multiplication of utterances' features by the most significant principal components.

The performance of the proposed methodology is evaluated on two common classification systems used in [4]. The first is score based classification of utterance feature vectors with two gender-dependent GMM-UBMs, one for each gender class. The second, is SVM boundary-based classification of the supervectors. The structure of these classification systems is explained in detail in Section IV. The performance of our system with GMM-UBM is superior to that of a recently presented work in [5], although different datasets are used, both systems use telephone speech.

As mentioned in the previous section, the relation between feature transformation using PCA and classification accuracy was studied in a different field in [11]. It can be observed from the literature, see e.g [12]–[14], that PCA transformation of speech features in gender classification specifically, has not been previously proposed and its effects have not been investigated. This is therefore addressed here in this work.

III. Computation Comlexity of the Classification System

This section illustrates the reduction in computational complexity of gender classification as a result of dimensionality reduction of speech features using PCA. The focus is on the main processes involved in the classification systems under study. These are GMM estimation and SVM training and classification. Let a set of reduced dimensionality feature vectors be defined by $\mathbf{Y} = \{\boldsymbol{y_1}, \boldsymbol{y_2}, \dots, \boldsymbol{y_T}\}$. The dimensionality of these feature vectors is d where $d < D$ and, D is the original dimensionality of the features before PCA. In this section, $\mathbf{Y}$ denote any set of feature vectors involved in the process of the classification systems' training and evaluation. In the following, we will describe the processing aspects where reduced dimensionality feature vectors can ease the computational load leading to decreased computation time.

Both of the classification systems deployed require the estimation of GMMs. Each Gaussian mixture is defined by its mean (μ_i), covariance matrix ($\mathbf{\Sigma}_i$) and weight (ω_i), where i is the mixture's index. The estimation of these parameters for a GMM is achieved by using maximum likelihood (ML) estimation. The idea is to determine the parameters that maximise the likelihood of the GMM given a set of feature vectors, let it be Y. The parameter estimates using ML are accomplished through expectation-maximisation (EM) since a direct maximisation is not possible due to non-linearity of the relevant ML function [15]. The parameters of the GMM model, λ, are randomly initiated and then, using EM, a new model $\bar{\lambda}$ is estimated from the previous model such that $p(\mathbf{Y}|\bar{\lambda}) \geq p(\mathbf{Y}|\lambda)$. This processes is iterated until a convergence threshold is reached. Each EM iteration requires re-estimation of the model parameters using the following formulas:

$$\bar{\omega}_i = \frac{1}{T}\sum_{t=1}^{T} p(i|\,\vec{y}_i, \lambda), \tag{1}$$

$$\bar{\mu}_i = \frac{\sum_{t=1}^{T} p(i|\,\vec{y}_i, \lambda)\,\vec{y}_i}{\sum_{t=1}^{T} p(i|\,\vec{y}_i, \lambda)}, \tag{2}$$

$$\bar{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i|\vec{y}_i,\lambda)\bar{y}_i^2}{\sum_{t=1}^{T} p(i|\vec{y}_i,\lambda)} - \bar{\mu}_i^2,$$

(3)

where $p(i|\vec{y}_i,\lambda)$ is the *a posteriori* probability of the Gaussian mixture *i* given by

$$p(i|\vec{y}_i,\lambda) = \frac{\omega_i b_i(\vec{y}_t)}{\sum_{k=1}^{M} \omega_k b_k(\vec{y}_t)}.$$

(4)

M is the order of the Gaussian mixture model (number of mixtures). In each iteration of the EM algorithm, the re-estimation of the GMM parameters requires the calculation of $b_i(\vec{y}_t)$ which is the density of the i^{th} Gaussian model component and it is a *d*-variate Gaussian function expressed as

$$b_i(\vec{y}_t) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{y}-\vec{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\vec{y}-\vec{\mu}_i)\right\},$$

(5)

where *d* is the dimensionality of feature vector $\vec{y}$ and it is the length of the mean vector $\vec{\mu}_i$ and the size of the covariance matrix $\boldsymbol{\Sigma}_i$. Using PCA, the dimensionality is reduced from *D* to *d*, hence the density of each Gaussian mixture component is a d-variate, $d < D$, Gaussian function expressed in (5).

A. Classification using GMM-UBM

Feature vectors from speakers of the same gender are used to estimate gender-dependent GMM-UBM. In the classification process, the log-likelihood $\mathcal{L}$ is calculated for a given set of test feature vectors given class GMM-UBMs. Assuming that reduced dimensionality set of feature vectors **Y** would now represent a test utterance, the log-likelihood of **Y** given a gender-dependent GMM-UBM Λ is given by

$$\mathcal{L} = \sum_{t=1}^{T} \log p(\vec{y}_t|\Lambda).$$

(6)

The number of components (order) of Λ is *M* , hence

$$p(\vec{y}|\Lambda) = \sum_{i=1}^{M} \omega_i b_i(\vec{y}).$$

(7)

The calculation of $\mathcal{L}$ also requires the computation of $b_i(\vec{y}_t)$ given in (5). Hence, the computation cost of classification is also reduced as a result of the reduced dimensionality of feature vectors.

B. Classification using SVM

For classification in SVM, a gender-independent GMM-UBM is first estimated using feature vectors from a group of speakers (males and females). Then for each speaker, the parameters of the GMM-UBM are adapted using the feature vectors extracted from the speaker speech which leads to speaker-dependent GMM. A form of Bayesian adaptation known as maximum a posteriori estimation is used. The adaptation process is not as extensive as the estimation of GMMs because it does not involve an iterate expectation-maximisation steps.

For a set of feature vectors, the GMM-UBM adaptation starts with the calculation of (4) for each Gaussian component which involves the computation of (5). Then the outcome of (4) is used with the same set of feature vectors to calculate sufficient statistics of the weight, mean and variance parameters. Dimensionality reduction reduces the complexity of adaptation. A complete description of the process can be found in [15].

For each training and test speaker's feature vectors, a GMM is produced by adaptation of the GMM-UBM. The means of each speaker's GMM are stacked together to form supervectors used for training and classification in SVM. The length (dimensionality) of the supervector is equal to the dimension of feature vectors *D* by the number of Gaussian components *M*, and it is reduced to *d* by *M* as a result of the reduced dimensionality feature vectors.

The simplest kernel function of SVM is the linear kernel function which is used in this paper. This kernel function computes an inner product of two vectors (which are the supervectors in our case) in the input space. Let $\mathbf{s}_m$ and $\mathbf{s}_f$ denote samples of training supervectors. The training of SVM involves the computation of the linear kernel function which is defined as

$$K_L(\mathbf{s}_m, \mathbf{s}_f) = \langle \mathbf{s}_m, \mathbf{s}_f \rangle = \mathbf{s}_m^T \mathbf{s}_f .$$

(8)

Let $\hat{D}$ be the size of the supervectors based on original feature vectors dimensionality, thus the computation of (8) is expressed as

$$\mathbf{s}_m^T \mathbf{s}_f = \sum_{l=1}^{\hat{D}} \mathbf{s}_{m,l}^T \mathbf{s}_{f,l} .$$

(9)

Based on reduced feature vectors, the size of the supervectors is reduced to $\hat{d}$, where $\hat{d} < \hat{D}$. Hence, the space of the inner product of (9) is reduced

$$\mathbf{s}_m^T \mathbf{s}_f = \sum_{l=1}^{\hat{d}} \mathbf{s}_{m,l}^T \mathbf{s}_{f,l} .$$

(10)

Let $\hat{\mathbf{s}}$ represent reduced size supervectors, the SVM models both gender classes by sums of kernel functions

$$f(\hat{\mathbf{s}}) = \sum_{j=1}^{n} \alpha_j t_j K_L(\hat{\mathbf{s}}, \hat{\mathbf{s}}_j) + b,$$

(11)

where $\hat{\mathbf{s}}$, is a supervector in the input space *S*, $\hat{\mathbf{s}}_j$, are the support vectors obtained by an optimisation process [2]. $\sum_{j=1}^{n} \alpha_j t_j = 0$, where t_j are the true outputs, -1 or 1, and $j > 0$ are real-value

coefficients called Lagrange multipliers. The term b is a bias. The classification decision is based on the value of $f(\hat{\mathbf{s}})$ compared to a threshold. We notice from (11) that the cost of classification in SVM is also reduced.

IV. EXPERIMENTAL SETUP

The speech features used in this work are Mel-Frequency Cepstral Coefficients (MFCC). 12 MFCC coefficients are calculated from Hamming windowed speech frames of a size of 25ms with 60% overlap (10ms shift). These coefficients are appended to their first and second derivatives to make 36 dimensional feature vectors. Afterwards, the mean and variance of the feature vectors of each utterance (test and training) are normalised. The dataset used is the 2002 NIST Speaker Recognition Evaluation telephone data, [17]. In the training phase, we have used utterances of 134 female and 134 male speakers. In the evaluation, we have used the available test samples, 1728 for females and 1215 for males.

For principal component analysis, feature vectors of all the training data (of both genders) are pooled together, then their mean and variance are normalised. Afterwards, the principal components are determined using Singular Value Decomposition (SVD) [1]. We call the resulted principal components the universal projection matrix.

In the GMM-UBM baseline system, two gender-dependent GMM-UBMs are trained with the training data of each class separately. For each class, the relevant training data is combined together in one matrix and the expectation maximisation algorithm is used to fit each gender-dependent GMMUBMs with 128 components each. The log-likelihood is then calculated for the feature vectors of each test utterance using both of the gender-dependent GMM-UBMs. The classification decision is made based on the greater log-likelihood value with either GMM-UBMs. With PCA, the mean and variance normalised feature vectors of the test and training data are projected using the universal projection matrix. Afterwards, the same modelling and evaluation procedure of the baseline system is followed.

The baseline system of the supervector classification using SVM can be described as follows. Feature vectors of the training data of both genders are pooled together and expectation maximisation is applied to produce a gender-independent GMM-UBM of 128 components. For each test and training utterance's feature vectors, a GMM is estimated by adapting the means of the gender-independent GMM-UBM with those feature vectors using *maximum-a-posteriori* estimation [15].

This results in an adapted GMM for each training and test utterance (speaker). Then, the means of these GMMs are variance normalised and are stacked together to form supervectors with dimension of 4608 (dimensionality of the feature vectors by the number of GMM components). From the training utterances, 134 supervectors for each gender are used to train the SVM with a linear kernel. PCA is incorporated into this system in the same way as described for classification using GMM-UBM (previous paragraph). The only difference is that after projection to the principal components, the variance of the resultant features is also normalised for each test and training utterance. This normalisation is found to be important for the SVM to perform with reasonable accuracy. We explain this behaviour by that the classification decision of SVM is based on a boundary (hypersurface).

V. EVALUATION RESULTS

The performance of the GMM-UBM classification system is superior to that of SVM whether PCA is involved or not. From Figures 1 and 2, it can be seen that at short utterances, the performance using SVM deteriorates compared to that of GMM-UBM which steadily degrades as a natural result of decreased utterance length. In the same figures, we find that PCA has presented marginal improvement over all the addressed range of utterance lengths, 1s to 10s, in both systems.

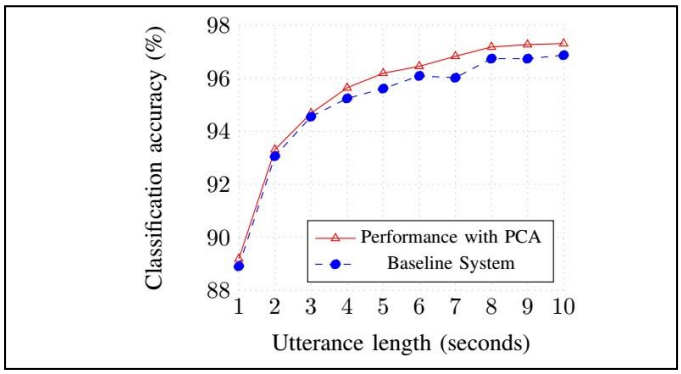

Fig. 1. Classification accuracy at various utterance lengths in the GMM-UBM based system.

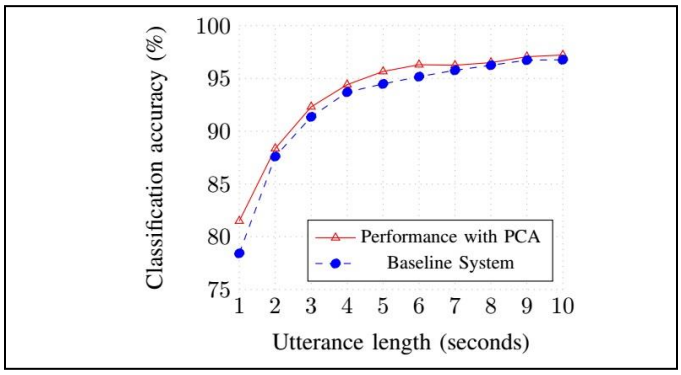

Fig. 2. Classification accuracy at various utterance lengths in the SVM based system.

The purpose of addressing the performance for very short utterances (e.g. 1s) is to show that dimensionality reduction presented by PCA does not have a negative effect on classification performance for short utterances. On the contrary, it produced an improvement of 3% with an SVM at an utterance length of 1s. We have also expressed (in figures) the relation between the classification accuracy and the percentage of the data variance captured by the number of principal components chosen for projection of the original features. The total number of the principal components is equivalent to the number of original features which is 36. The components with the higher eigenvalues express higher variance of the original MFCC features. We have studied the effect of variance captured by the selected PCA components which is found to be in the range of 72.29% to 94.77%. This applies to the highest eigenvalue

components which are in the range of the 17^{th} to 30^{th} eigenvectors of the projection matrix.

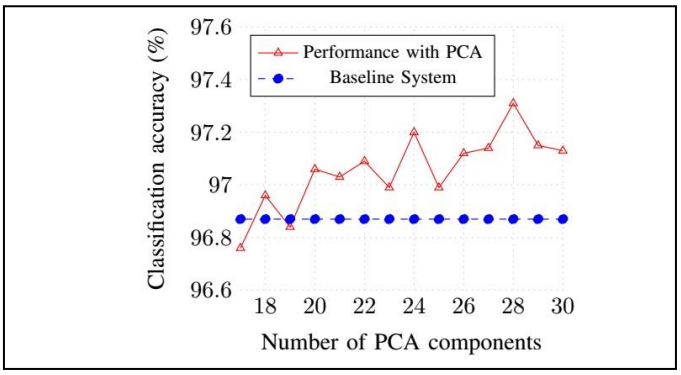

Fig. 3. Performance of gender classification with the GMM-UBM based system in relation to the number of PCA components used.

Figures 3 and 4 illustrate this effect at an utterance length of 10s. The maximum improvement is produced at 28 principal components which is equal to using 92.40% of the variance of the original MFCC features with a dimensionality reduction of 22%. This amount of reduction means, for example, that the 4608 dimensional supervectors used in the training and test in SVM are reduced to 3548.

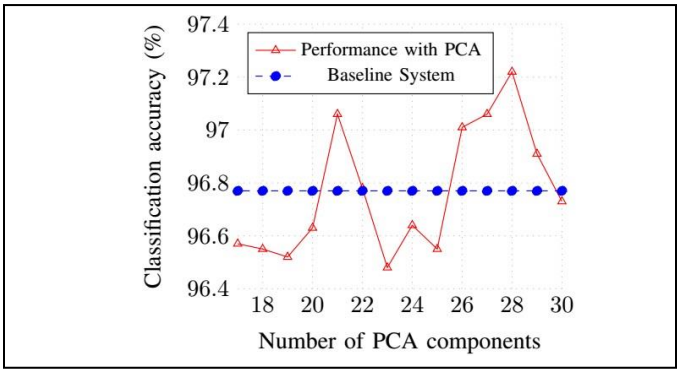

Fig. 4. Performance of gender classification with the SVM based system in relation to the number of PCA components used.

TABLE I. COMPUTATION TIME IN RELATION TO FEATURE DIMENSIONALITY FOR GMM ESTIMATION (GMM-UBM ESTIMATION FOR 128 COMPONENTS USING 588654 FEATURE VECTORS) AND LOG-LIKELIHOOD SCORING WITH GMM-UBM (FOR TEST UTTERAQNCES FEATURE VECTORS WITH SUM OF 2943000). THE MAXIMUM TIME REDUCTION AT 18 DIMESNIONS IS 20% FOR GMM-UBM ESTIMATION AND 19% FOR LOG-LIKELIHOOD CALCULATION, RESPECTIVELY.

Feature Dimensionality	GMM Estimation	Log-Likelihood Scoring
36	78.120 s	5.958 s
27	71.013 s	5.174 s
18	63.119 s	4.720 s

TABLE II. COMPUTATION TIME IN RELATION TO FEATURE DIMENSIONALITY FOR GMM ADAPTATION AND SVM TRAINING AND CLASSIFICATION. GMM-UBM IS ADAPTED FOR EACH TRAINING AND TEST UTTERANCE (THE SUM OF FEATURE VECTORS IS 5329549). SVM WAS TRAINED USING 268 SUPERVECTORS AND IS USED TO CALSSIFY 2943 SUPERVECTORS. THE MAXIMUM TIME REDUCATION AT FEATURE DIMENSIONALITY OF 18 IS 15% FOR GMM-UBM ADAPTATION AND IT IS 47% AND 44% FOR SVM TRAINING AND CLASSIFICATION, RESPECTIVELY.

Feature Dimensionality	GMM Adaptation	SVM Training	SVM Classification
36	451 s	0.141 s	5.132 s
27	423 s	0.110 s	4.192 s
18	384 s	0.075 s	2.886 s

In favour of the application of gender classification systems, we notice from figures 3 and 4 that a dimensionality reduction of 50% can be achieved by using only 18 principal components, whilst maintaining a comparable performance to that of the baseline systems. This means faster performance which can be very useful in applications that need to deal with large data.

The reduction in computation time is also investigated. Tables I and II show the computation time (in seconds) of the processes that follows feature extraction and dimensionality reduction using PCA. Three feature dimensions are investigated which is the original dimension of 36, and two cases of reduced feature dimensionality using PCA which are 27 and 18. Table I includes the GMM estimation times taken for both classification systems. The same table also shows the time taken by log-likelihood calculation between the test feature vectors and the GMM-UBM for classification in GMM-UBM system. Table II shows the time taken by the adaptation of the GMM-UBM to test and training feature vectors prior to SVM training and classification. It also shows the time taken by the training and classification of SVM.

VI. DISCUSSION AND CONCLUSIONS

Incorporating PCA in gender classification systems is shown here to be useful and efficient. It is found to provide a maximum improvement of 3% (Figure 2). The consequence of adding the PCA step in the system is negligible as it only requires a simple matrix multiplication of the training or test data by the PCA projection matrix. Furthermore, this matrix would have already been produced before engaging the system. On top of the improvements shown, the processing time of gender classification systems investigated is decreased by a maximum of 47% as a result of dimensionality reduction provided by PCA. The work could be extended to more complex gender classification systems such as i-vector based systems [18]. Conventional PCA, which is used in this work, may not account for potential anomalies in the data such as outliers. A future work can investigate alternative approaches to PCA in gender classification, for example, weighted PCA, see [9]. The approach taken here has the advantage of being straightforward, but the work in [9] would enable weighting the feature vectors to reduce the effect of outliers on PCA analysis.

REFERENCES

[1] H. Beigi, *Fundamentals of speaker recognition*. Springer Science & Business Media, 2011.

[2] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer Science & Business Media, 2007.

[3] The-Information-Commissioner-Office, "Processing personal data fairly and lawfully," Website, 2017. [Online]. Available: https//ico.org.uk/fororganisations/guide-to-dataprotection/princicpal-1-fair-and-lawful/

[4] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, vol. 27, no. 1, pp. 151–167, 2013.

[5] A. Kanervisto, V. Vestman, M. Sahidullah, V. Hautamaki, and T. Kinnunen, "Effects of gender information in text-independent and textdependent speaker verification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 5360–5364.

[6] J. M. Huerta, "Speech recognition in mobile environments," Carnegie Mellon University, Pittsburgh, PA 15213, Tech. Rep., 2000.

[7] B. Ayyub and M. Gupta, *Uncertainty Analysis in Engineering and Sciences: Fuzzy Logic, Statistics, and Neural Network Approach*, ser. International Series in Intelligent Technologies. Springer US, 2012.

[8] J. Kludas, E. Bruno, and S. Marchand-Maillet, "Information fusion in multimedia information retrieval." in *Adaptive Multimedia Retrieval*, Springer. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 147–159.

[9] A. I. Ahmed, J. P. Chiverton, D. L. Ndzi, and V. M. Becerra, "Speaker recognition using PCA-based feature transformation," *Speech Communication*, vol. 110, pp. 33 – 46, 2019.

[10] M. Al-Faris, J. Chiverton, D. Ndzi, and A. I. Ahmed, "A review on computer vision-based methods for human action recognition," *Journal of Imaging*, vol. 6, no. 6, 2020. [Online]. Available: https://www.mdpi.com/2313-433X/6/6/46

[11] A. Janecek, W. Gansterer, M. Demel, and G. Ecker, "On the relationship between feature selection and classification accuracy," in *New Challenges for Feature Selection in Data Mining and Knowledge Discovery*, 2008, pp. 90–105.

[12] Y. Zhou and L. Shang, "Speaker recognition based on principal component analysis and probabilistic neural network," *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence*, vol. 6839, pp. 708–715, 2012.

[13] C. Seo, K. Y. Lee, and J. Lee, "GMM based on local PCA for speaker identification," *Electronics Letters*, vol. 37, no. 24, pp. 1486–1488, 2001.

[14] X. Jing, J. Ma, J. Zhao, and H. Yang, "Speaker recognition based on principal component analysis of LPCC and MFCC," in *Signal Processing, Communications and Computing (ICSPCC)*, 2014 *IEEE International Conference on.* IEEE, 2014, pp. 403–408.

[15] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[16] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Frontend factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[17] A. Martin and P. Mark, "2002 NIST speaker recognition evaluation LDC2004s04," *Web Download*, 2004.

[18] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.