Improved Language-Independent Speaker Identification in a Non-contemporaneous Setup

Smarajit Bose, Amita Pal, Anish Mukherjee, and Debasmita Das

***Abstract*—One of the most effective approaches available in the literature for Automatic Speaker Identification is based on Gaussian Mixture Models (GMMs) with Mel Frequency Cepstral Coefficients (MFCCs) as features (Reynolds (1995). The use of GMMs for modeling speaker identity is motivated by the interpretation that the Gaussian components represent some general speaker-dependent spectral shapes, and the capability of mixtures to model arbitrary densities. In an earlier work, the authors have presented and demonstrated empirically (using the benchmark speech corpus NTIMIT) how combining two different well-known set of features (MFCCs and Perceptual Linear Predictive Coefficients (PLPCs)) and using ensemble classifiers in conjunction with the Principal Component Transformation (PCT) and some robust statistical estimation techniques, enhances significantly the performance of the baseline MFCC-GMM speaker recognition system. In this work, the authors demonstrate that this approach, besides being statistically robust, is also significantly more robust than the baseline system to language mismatch in a non-contemporaneous setup. This has been done with the help of ISIS/NISIS, a bilingual dual-channel speech corpus with multi-session speech recordings.**

***Index Terms*—Mel frequency cepstral coefficients, perceptual linear predictive coefficients, Gaussian mixture models, ensemble classifiers, classification accuracy, trimmed mean.**

I. Introduction

Automatic speaker identification/recognition (ASI/ASR) refers to the collection of methodologies for the automatic inference of the identity of a person from an utterance made by him, exploiting speaker-specific information inherent in the corresponding speech signal. It is an important area of research with significant real-life practical applications. Algorithms and applications are well-documented in literature, for example, in [1], [2].

One effective approach was proposed by Reynolds [3], Reynolds and Rose [4], which was based on MFCCs as features and GMMs as speaker models. By implementing it on the benchmark speech corpora TIMIT and NTIMIT, they demonstrated that it works almost flawlessly on clean speech (TIMIT) and quite well on noisy telephone speech NTIMIT). This approach is still one of the best available in the literature.

In an earlier work [5], the authors have proposed an extension of the baseline MFCC-GMM speaker recognition system of Reynolds, which shows significantly enhanced recognition accuracy on NTIMIT by

1) *augmenting the original feature set consisting of MFCCs with a set of* Perceptual Linear Predictive Coefficients (PLPCs) [6];

2) *implementing robust statistical estimation techniques like the trimmed mean, to eliminate the effect of outliers*, that is, observations that are too different from the majority of observations and may be due to the inherent variability in the data set or to measurement error.

The following two ideas from a previous work [7] were also used:

1) *Incorporation of the individual correlation structures of the feature sets of each speaker into the corresponding speaker models*: This correlation structure is a significant aspect of the speaker models which was completely ignored by Reynolds by assuming the MFCCs to be independent. This is achieved through the well-known *Principal Component Transformation* (PCT) [8].

2) *Using ensemble classification:* It is a well-known fact that the use of an ensemble of classifiers instead of a single classifier can improve the accuracy to a great extent. A clever device has been employed to design an ensemble classifier which further improved the classification accuracy.

This paper is organized as follows. The features used, namely, MFCCs and PLPCs, are briefly described in the following section. GMMs are covered in Section III, which also outlines the baseline speaker recognition system with MFCC features and GMM speaker models. The extension of the baseline MFCC-GMM proposed by the authors in [5] is described in Section IV. Section V presents the salient features of the ISIS/NISIS bilingual dual-channel multi-session speech corpus used to demonstrate empirically that the algorithm in Section IV, apart from being statistically robust, is also significantly more robust than the baseline system to language mismatch in a non-contemporaneous setup. Results that support this are provided in Section VI. Concluding remarks are made in Section VII.

II. Features Used

A. Mel Frequency Cepstral Coefficients (MFCCs)

The Mel Frequency Cepstrum (MFC) is a representation of

Manuscript received June 11, 2019; revised February 7, 2020.

Smarajit Bose and Amita Pal are with the Interdisciplinary Statistical Research Unit (ISRU), Applied Statistics Division, Indian Statistical Institute, Kolkata, India (e-mail: {smarajit,pamita}@isical.ac.in).

Anish Mukherjee is with the Department of Statistics, University of Missouri, Columbia, MO, USA (e-mail: anishmk9@gmail.com).

Debasmita Das is with the Department of Statistics, University of Connecticut, Storrs, CT, USA (e-mail: debasmita88@yahoo.com).

doi: 10.18178/ijmlc.2020.10.5.984

the short-term power spectrum of a speech signal, based on a linear cosine transform of a log-energy spectrum on a nonlinear mel scale of frequency. It exploits auditory principles, as well as the decorrelating property of the cepstrum, and is amenable to compensation for convolution distortion. As such, it has turned out to be one of the most effective feature representations in speech-related recognition tasks [9].

Mel-frequency Cepstral Coefficients (MFCCs) [10] are coefficients that collectively make up an MFC. A given speech signal is partitioned into overlapping segments or frames, and MFCCs are computed for each such frame. Based on a bank of *K* filters, a set of *M* MFCCs is computed from each frame as follows:

Let $x[m], w[m]$ denote respectively the speech signal and a window function at a time point m within the frame. The speech waveform $x[m]$ is windowed with $w[m]$, and its short-time Fourier transform (STFT), $Y(n,\omega_k), n=1,2,\ldots,N,$ is computed as

$$Y(n,\omega_k)=\sum_{m=-\infty}^{\infty} x[m]w[n-m]e^{-j\omega_k m},$$

where

$$\omega_k=\frac{2\pi}{n}k,$$

N being the length of the discrete Fourier transform. The magnitude of $Y(n,\omega_k)$ is then weighted by a series of filter frequency responses whose center frequencies and bandwidths roughly match those of the auditory critical band filters, that is, the so-called mel-scale filters, collectively referred to as a mel-scale filter bank (see below). If the frequency response of the ℓ^{th} mel-scale filter is denoted by $V_\ell(\omega),$ then its energy at n is

$$E_{mel}(n,\ell)=\frac{1}{A_\ell}\sum_{k=L_\ell}^{U_\ell}\left|V_\ell(\omega_k)\,Y(n,\omega_k)\right|^2,$$

where L_ℓ and U_ℓ denote respectively the lower and upper frequency indices over which the ℓ^{th} filter is non-zero, and

$$A_\ell=\sum_{k=L_\ell}^{U_\ell}\left|V_\ell(\omega_k)\right|^2.$$

Finally, for $i=1,2,\ldots,M,$ the i^{th} MFCC computed from the frame is

$$MFCC_i=\sum_{k=1}^{K}\log(E_{mel}(i,k)\cos\left[i\left(k-\frac{1}{2}\right)\frac{\pi}{K}\right].$$

Mel-scale Filter Banks

A mel-scale filter bank (Fig. 1) is a set of filters spaced uniformly on the mel scale (described below), which has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval.

The Mel Scale

Psychophysical studies show that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. For each tone with an actual frequency, f, measured in hz, a subjective pitch is measured on the so-called 'mel' scale. The mel scale is a scale of pitches judged by listeners to be equal in distance from one another. The word mel comes from the word melody to reflect this. This scale is a linear frequency spacing below 1000 hz and a logarithmic spacing above 1000 hz (Fig. 2).

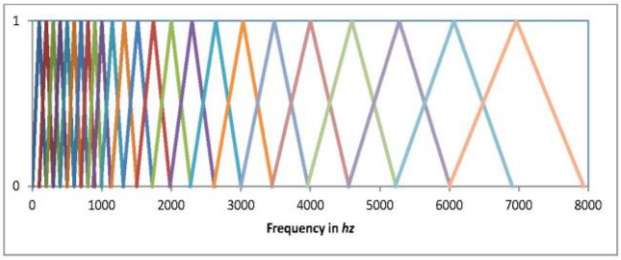

Fig. 1. A mel scale filter bank.

A popular formula to convert *f* hertz into *m* mel is:

$$m=2595\log_{10}\left(1+\frac{f}{100}\right).$$

Computation of MFCCs

This involves the following steps:

1. Partitioning the speech signal into overlapping segments or frames
2. Taking the Fourier transform of signal from each frame.
3. Mapping the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
4. Taking the logs of the powers at each of the mel frequencies.
5. Taking the discrete cosine transform of the list of mel log powers, as if it were a signal.

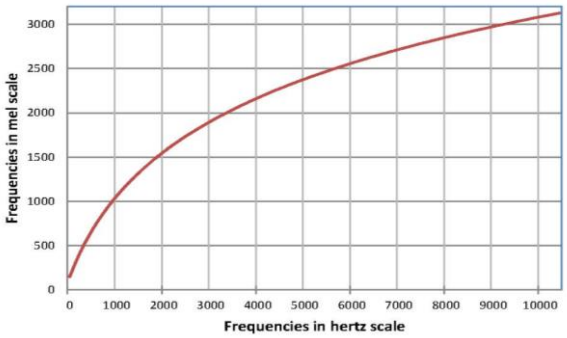

Fig. 2. The mel scale.

B. Perceptual Linear Predictive Coefficients (PLPCs)

Perceptual Linear Prediction is a method of spectral estimation proposed by Hermansky [6]. In different psycho-acoustic experiments it was observed that human frequency resolution varies over different frequency ranges and low frequencies mask higher ones. Moreover, it has been found that hearing is most sensitive at mid-frequencies. While listening people generally integrate 1 bark of spectrum, whereas for discrimination purpose people seem to integrate

about 3.5 barks of spectrum. These observations inspired the development of Perceptual Linear Predictive Coefficients (PLPC) which turned out to be superior in many ways to the Linear Predictive (LP) coefficients in the task of speaker identification.

In this technique of speech analysis, mainly three psycho-acoustic concepts are used to estimate the auditory spectrum which are critical-band spectral analysis, the equal loudness curve and the intensity power law. PLP algorithm can be described using the following steps -- first in the spectral analysis phase the speech signal is partitioned into overlapping segments and each segment is weighted by the Hamming window.

The short-term power spectrum $P(\omega)$ is computed for each of these segments. In the next stage, the spectrum $P(\omega)$ is warped along the frequency axis into the Bark Frequency which is then convolved with power spectrum of the simulated critical band masking curve that results in samples of the critical-band power spectrum. In this step, spectral resolution is significantly reduced. The sampled power spectrum is then pre-emphasized by an equal-loudness curve and a cubic-root amplitude compression is performed simulating the power law of hearing. Finally, in the autoregressive modeling phase, the resulting spectrum is modeled by a 5th order model using the autocorrelation method of all-pole spectral modeling. The following block diagram shows the steps of PLP algorithm.

PLP has an advantage of approximating the speaker independent effective second formant. It reduces the disparity between voiced and unvoiced speech. It has been shown in different experiments that there exists a strong correlation between the perceptually estimated second formant and that estimated by the PLP method.

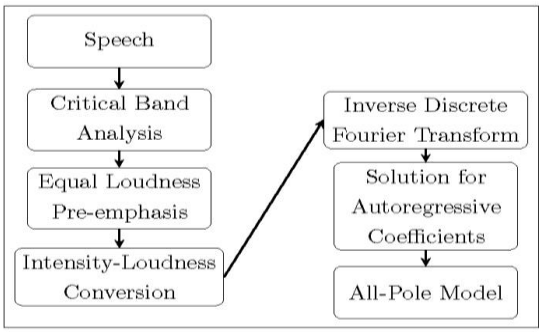

Fig. 3. The PLP algorithm.

III. Speaker Recognition with MFCC-Based GMM Speaker Models

A. Gaussian Mixture Models (GMMs)

If $\boldsymbol{x}$ is a d-dimensional feature vector, then for a S-speaker problem, the probability distribution of the MFCCs obtained from speaker $i, i = 1, 2, \ldots, S$, is modeled as a mixture of C component probability densities as follows:

$$p(\boldsymbol{x} \mid \lambda_i) = \sum_{j=1}^{C} p_{ij} \; f(\boldsymbol{x} \mid \theta_{ij}),$$

with

$$\sum_{j=1}^{C} p_{ij} = 1,$$

where, for the i^{th} speaker, p_{ij} is the prior probability for the j^{th} component of the mixture, $f(\boldsymbol{x} \mid \theta_{ij})$ is the probability density of the feature vector $\boldsymbol{x}$ in the j^{th} component, assumed to be Gaussian in this case and $\lambda_i = \{p_{ij}, \theta_{ij}, \; j = 1, 2, \ldots, C\}$ is the collection of unknown parameters. That is, for a GMM,

$$p(\boldsymbol{x} \mid \lambda_i) = \sum_{j=1}^{C} p_{ij} \frac{1}{(2\pi)^{d/2} \left|\boldsymbol{\Sigma}_{ij}\right|} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{ij})^T \boldsymbol{\Sigma}_{ij}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{ij})},$$

And $\theta_{ij} = \{\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij}\}, i = 1, 2 \ldots, S; \; j = 1, 2, \ldots, C.$

GMM models for all speakers are trained by the Expectation-Maximization algorithm [10]. An unknown speech sample is split into a number of overlapping segments, with MFCCs computed from each segment. The likelihood function for the sample is computed, based on all MFCC vectors obtained from it, and it (the unknown sample) is classified by the principle of maximum likelihood, described below.

B. Speaker Recognition by Maximum Likelihood

Consider a speaker database consisting of S speakers where the i^{th} speaker is being represented by a GMM $p(\boldsymbol{x} \mid \lambda_i)$ as defined above. If a speech utterance of unknown origin is presented, and it is known that is speaker is represented in the speaker database, the objective of speaker recognition is to identify which of the K speakers could have uttered it.

Suppose the unknown utterance is split into P overlapping frames using the same procedure as for the training samples, and MFCCs are computed from each segment. If $\boldsymbol{x}_p$ denotes the MFCC vector computed from the p^{th} segment, then the overall likelihood of the unknown utterance under the i^{th} speaker model is

$$L(\lambda_i) = \prod_{p=1}^{P} p(\boldsymbol{x}_p \mid \lambda_i),$$

assuming the MFCC vectors from different segments to be mutually independent statistically.

Speaker number k is identified as the speaker of the unknown speech utterance if

$$L(\lambda_k) = \max_{1 \le i \le S} L(\lambda_i).$$

Since the logarithm function is monotonically increasing in its argument, maximizing the likelihood function $L(\lambda_i)$ is equivalent to maximizing the log-likelihood

$$\ell(\lambda_i) = \log L(\lambda_i) = \sum_{p=1}^{P} \log p(\boldsymbol{x}_p \mid \lambda_i).$$

Thus speaker number k is identified as the speaker of the unknown speech utterance if

$$\ell(\lambda_k) = \max_{1 \le i \le S} \ell(\lambda_i).$$

IV. Robust Speaker Recognition by Fusion of Features and Classifiers

A novel approach for robust speaker recognition by fusion of features and classifiers was presented by the authors in an earlier work [5]. In this approach, significant enhancement in classification accuracy of the baseline MFCC-GMM speaker recognition system was made possible by a combination of the following:

1. *Augmentation of the MFCC-based feature set with a PLPCs*: Experiments were conducted separately with both these feature sets. Further investigations revealed that the classifiers built by fusing both feature sets could identify different speakers even more accurately
2. *Fusion of classifiers:* Since there were quite a few parameters in the MFCC-GMM model, one can build new classifiers by choosing different combination of values for the parameters. An ensemble classifier based on 3-4 such classifiers was employed for the final classification.

There are several ways of combining the decisions of different classifiers in an ensemble classifier. Majority voting is quite popular. The authors opted for fusion through aggregation of likelihood values of different classifiers, followed by maximization of the aggregated likelihood values.

Thus, if there are *R* number of classifiers in an ensemble, then the aggregated likelihood $L'(\lambda_i)$ is

$$L'(\lambda_i) = \sum_{j=1}^{R} L(\lambda_j).$$

1. *Incorporation of the individual correlation structures of the feature sets for each speaker into the corresponding speaker model*: This is achieved through the use of the *Principal Component Transformation* (PCT) [8], which is described below. The baseline MFCC-GMM system ignores this totally by assuming the MFCCs to be independent.

Since correlation structures differ from speaker to speaker, these transformations are also different for different speakers. The GMM for a particular speaker is fitted on the MFCCs transformed by the principal component transformations for that speaker, rather than the original MFCCs. As far as testing is concerned, to determine the likelihood values with respect to a given target speaker model, the MFCCs from the test utterance are transformed by the principal component transformation corresponding to that speaker.

2. *Implementation of robust estimation procedures like the trimmed mean to eliminate the effect of outliers*: Outliers are observations that are too different from the majority of observations and may be due to the inherent variability in the data set or to measurement error.

This is motivated by the observation that the log-likelihood function $\ell(\lambda_i)$ can be interpreted as being equal to $P\rho(\lambda_i)$, where

$$\rho(\lambda_i) = \frac{1}{P} \sum_{p=1}^{P} \log p(\boldsymbol{x}_p \mid \lambda_i),$$

which is nothing but the average or arithmetic mean of the *P* $\log p(\boldsymbol{x}_p \mid \lambda_i)$ values. Also, maximizing $\ell(\lambda_i)$ over *i* is equivalent to maximizing $\rho(\lambda_i)$ over *i*. To obtain a more robust estimate of $\rho(\lambda_i)$, we make use of the well-known trimmed mean procedure [12], which is described below.

A. Principal Component Transformation (PCT)

This is a widely-used linear orthogonal transformation for converting a set of observations on possibly correlated variables into a set of observations on linearly uncorrelated variables called *principal components* [8]. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (i.e., uncorrelated with) the preceding components. Principal components are guaranteed to be independent only if the data set is jointly normally distributed. PCT is sensitive to the relative scaling of the original variables. Depending on the field of application, it is also called the Karhunen–Loève transform (KLT), and so on.

Let **X** be a $m \times n$ data matrix each of whose *n* columns represents an observation on an *m*-variate random variable ***U***. It is assumed that the columns have zero empirical mean (that is, the arithmetic mean of the *n* observations has been subtracted from each of them). If the $m \times m$ matrix $\boldsymbol{\Sigma}$ is the dispersion matrix of the observations, with eigenvalues $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_m$, the corresponding eigenvectors being $\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots, \boldsymbol{P}_m$, then the principal component transformation of **X** that preserves dimensionality (that is, gives the same number of principal components as the original variables) is given by

$$\mathbf{Y} = \mathbf{PX},$$

where $\mathbf{P}$ is a $m \times m$ orthogonal matrix having $\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots, \boldsymbol{P}_m$ as its columns, and the columns of **Y** are the transformed versions of the original *m*-variate observations on ***U*** forming the columns of **X**.

B. The Trimmed Mean

Let $x_1, x_2, \ldots, x_n$ be *n* univariate observations, and let the corresponding ordered observations be $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. Then, for some $\alpha \in (0, 0.5)$, where $\alpha = \alpha_1 + \alpha_2$ for some $\alpha_1, \alpha_2 \in [0, 0.5)$, the α-trimmed mean of the *n* observations is defined as

$$\bar{x}_{\alpha} = \frac{1}{n - A_1 - A_2} \sum_{i=A_1+1}^{n-A_2} x_{(i)},$$

where $A_j = \lfloor n\alpha_j / 2 \rfloor$, $j = 1, 2$, $\lfloor \bullet \rfloor$ being the floor function. It is nothing but the average of observations excluding the A_1 smallest and A_2 largest, so that a proportion α of the observations (that are supposedly extreme) are excluded.

V. The ISIS/NISIS Speech Corpus

ISIS (an acronym for Indian Statistical Institute Speech) and **NISIS** (Noisy ISIS) are speech corpora that respectively contain simultaneously recorded microphone and telephone speech (spontaneous as well as read), over multiple sessions, in two languages (Bangla and English), in a typical office environment with moderate background noise. They were created in the Indian Statistical Institute, Kolkata, as a part of a project funded by the Department of Information Technology, Ministry of Communications and Information Technology, Government of India, during 2004-07. The speakers generally had Bangla or another Indian language as their mother tongue, and so were non-native English speakers. Details of the methodology of collection are given in [13].

Particulars of both corpora are given below:

- Number of speakers: 105 (53 male + 52 female)
- Recording environment: moderately quiet computer room
- Sessions per speaker: 4 (numbered I, II, III and IV)
- Interval between sessions: 1 week to about 2 months
- Types of utterances in Bangla and English per session:

10 isolated words (randomly drawn from a specific text corpus, and generally different for all speakers and sessions)

answers to 8 questions (these answers included dates, phone numbers, alphabetic sequences, and a few words spoken spontaneously)

12 sentences (first two sentences common to all speakers, the remaining randomly drawn from the text corpus, duration ranging from 3-10 seconds)

Thus, for each session, there are two sets of recordings per speaker, one each in Bangla and English, containing 21 files each.

To conduct the experiments whose results are reported in the following section, ten utterances from the "sentences" folder were used for 100 speakers enrolled in the NISIS corpus. Two separate series of experiments were conducted by splitting the 10 recordings for each speaker at each session as follows:

- The first 6 were used for training and rest for testing
- The first 8 were used for training and rest for testing

For brevity, these experiments will subsequently be referred to as 6:4 and 8:2 respectively.

For brevity, we shall subsequently refer to the set of all Bangla sentence utterances recorded in session no. *i* as BS-*i*, *i*=I, II, III, IV, for each speaker. The corresponding notation for the set of all English sentence utterances recorded in session no. *i* is ES-*i*.

VI. Results

In [13], the performance with NISIS in the 6:4 setup of the baseline MFCC-GMM system based on 32-component GMMs built with 38 MFCCs, was reported. The corresponding results, with training and test data mismatched in respect of session and/or language, are given in the light grey cells of Table I.

For conducting further experiments, a 39-dimensional feature set (FS-I), formed by combining 13 MFCCs, 13 delta MFCCs and 13 PLPCs, has been used. For building ensemble classifiers, we have also made use of 39-dimensional feature set (FS-II) consisting only of MFCCs. Both feature sets have been used in conjunction with different values of underlying parameters to construct a number of classifiers for use in the ensembles. A GMM with 32 components based on one or both of these 39-dimensional feature vectors has been used for each speaker in all cases.

TABLE I: Comparative Recognition Accuracy with NISIS of the Baseline MFCC-GMM System[A] and the MFCC-PLPC-GMM System[B] (in the 6:4 Setup)

Train Set ↓ \ Test Set →	BS-I	BS-II	BS-III	BS-IV	ES-I	ES-II	ES-III	ES-IV
BS-I	**71**	41	42	39	51	41	36	29
	94	67	64	64	73	53	52	47
BS-II	47	**70**	45	38	43	45	37	36
	66	**90**	67	62	52	70	55	53
BS-III	36	39	**69**	42	43	42	41	25
	67	68	**89**	65	50	59	71	52
BS-IV	45	44	48	**70**	37	39	42	41
	66	71	71	**86**	47	58	60	72
ES-I	33	29	29	25	**68**	40	35	33
	74	52	48	48	**83**	60	53	52
ES-II	34	37	28	27	41	**71**	37	31
	52	75	51	53	62	**86**	63	60
ES-III	34	32	36	28	45	40	**69**	34
	54	55	74	53	58	64	**84**	65
ES-IV	27	28	32	34	29	36	38	**64**
	46	52	51	66	58	65	62	**80**

[a] in light grey background
[b] in a deeper grey background

For comparison, in the darker grey cells of Table I, we provide the results of the MFCC-GMM system, based on FS-I, which will henceforth be referred to as the MFCC-PLPC-GMM system. It is very evident from this Table that the MFCC-PLPC-GMM system has substantially improved recognition accuracy relative to the baseline MFCC-GMM system. This pattern has also been observed in the 8:2 case. In other words, by combining the PLPC features with the MFCC features, significant improvement is achieved in the accuracy across all the experiments. This has been used as the baseline in further experiments.

To investigate the effect of language and session mismatch on speaker identification accuracy of the approach proposed in [5] and described in Section IV, extensive experiments were conducted in which the training/test data were taken from different sessions/languages. For brevity, this approach will henceforth be referred to subsequently as PREF (an

acronym for ***P**rincipal component-transformed **R**obust **E**nsemble of **F**eatures*).

As will be amply evident from the following discussion, PREF based on FS-I (or PREF-I, in short) leads to even greater improvement in recognition accuracy. Overall, the degradation due to language and temporal mismatch is seen to have been restored to a great extent by PREF-I in most of the experiments.

Performance of the MFCC-PLPC-GMM system (as baseline) as well as PREF-I, with training and test data varying across languages and sessions, for both the 6:4 and 8:2 setups, are reported in Tables II and III respectively.

TABLE II: COMPARATIVE RECOGNITION ACCURACY WITH NISIS OF THE BASELINE MFCC-PLPC-GMM SYSTEM[A] AND PREF-I[B] (6:4 SETUP)

Train Set ↓ \ Test Set→	BS-I	BS-II	BS-III	BS-IV	ES-I	ES-II	ES-III	ES-IV
BS-I	**94**	67	64	64	73	53	52	47
	98	75	70	73	84	66	69	62
BS-II	66	**90**	67	62	52	70	55	53
	79	**97**	79	76	68	90	70	67
BS-III	67	68	**89**	65	50	59	71	52
	80	82	**96**	82	68	74	89	70
BS-IV	66	71	71	**86**	47	58	60	72
	77	82	78	**96**	64	74	77	88
ES-I	74	52	48	48	**83**	60	53	52
	92	66	63	63	**96**	73	70	68
ES-II	52	75	51	53	62	**86**	63	60
	68	87	67	72	74	**93**	76	78
ES-III	54	55	74	53	58	64	**84**	65
	66	70	90	73	69	83	**95**	82
ES-IV	46	52	51	66	58	65	62	**80**
	67	69	69	85	67	81	79	**95**

[a]in light grey background
[b]in a deeper grey background

As expected, the best results for both approaches are obtained when both training and test data are in the same language and are contemporary, that is, recorded in the same session, and there is a general deterioration when

- the training and test data correspond to different languages;
- the training and test data are recorded at different points of time, the degradation in performance becoming more pronounced in general as the degree of non-contemporaneity increases;
- training and test data differ both in language and in session, the decrease in accuracy being generally more pronounced.

Moreover, there is very significant improvement in classification accuracy relative to the baseline system even in cases of language and/or session mismatch in training and test data. This is consistently observed in all experiments and provides empirical justification for the claim that the statistically robust approach of Section IV is also language-independent even if the setup is non-contemporaneous.

VII. CONCLUSION

In real-life speaker recognition problems, non-contemporaneity of the training data with respect to the test data is almost always a very important concern, as is language mismatch. In view of this, the robust approach proposed in [5] and described in Section IV shows great promise of providing greater recognition accuracy in such cases.

TABLE III: COMPARATIVE RECOGNITION ACCURACY WITH NISIS OF THE BASELINE MFCC-PLPC-GMM SYSTEM[A] AND PREF-I[B] (8:2 SETUP)

Train Set ↓ \ Test Set→	BS-I	BS-II	BS-III	BS-IV	ES-I	ES-II	ES-III	ES-IV
BS-I	**93**	70	64	67	77	58	59	50
	98	87	85	82	90	79	83	75
BS-II	72	**96**	76	70	57	78	60	53
	91	**97**	92	86	82	91	82	79
BS-III	75	75	**95**	73	55	65	77	60
	93	91	**96**	90	85	82	92	85
BS-IV	75	79	75	**91**	56	62	65	79
	89	93	92	**96**	81	82	89	91
ES-I	85	57	58	58	**89**	67	60	63
	92	82	83	76	**96**	84	82	78
ES-II	56	80	59	57	67	**84**	68	65
	87	94	87	86	85	**93**	88	87
ES-III	59	60	80	61	62	74	**90**	74
	90	87	97	87	86	91	**95**	92
ES-IV	55	56	56	73	61	73	73	**88**
	87	87	89	91	88	89	89	**95**

[a]in light grey background
[b]in a deeper grey background

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Pal and Bose were responsible for the design and acquisition of the ISI and NISIS speech corpora as well as initial research ideas and implementation, under a project funded by Department of Information Technology, Ministry of Communication and Information Technology, Government of India. Bose proposed the application of the PCT, ensemble methods and trimmed mean. Mukherjee proposed the inclusion of PLPC features and their combination with MFCCs, and also did the initial coding in MATLAB. Mukherjee and Das conducted the experiments and did the relevant data analysis. Pal prepared the manuscript. All authors approved the final version.

ACKNOWLEDGMENT

The authors gratefully acknowledge the funding provided for the initial exploratory work in this area by Department of Information Technology, Ministry of Communication and Information Technology, Government of India. They are also indebted to Prof. Hema Murthy of the Indian Institute of Technology Madras for her invaluable assistance in the form of useful discussions and for loan of software related to the MFCC-GMM speaker recognition system.

REFERENCES

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, pp. 12-40, 2010.
[2] J. P. Campbell, "Speaker recognition: A tutorial," in *Proc. the IEEE*, 1997, vol. 85, pp.1437-1462.

[3] D. A. Reynolds, "Large population speaker identification using clean and telephone speech," *IEEE Signal Processing Letters*, vol. 2, pp. 46-48, 1995.
[4] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72-83, 1995.
[5] S. Bose, A. Pal, A. Mukherjee, and D. Das, "Robust speaker identification busing fusion of features and classifiers," *International Journal of Machine Learning and Computing*, vol. 7, pp. 133-138, Oct. 2017.
[6] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J.Acoustical Society of America*, vol. 87, no. 4, pp. 1738-1752, 1990.
[7] A. Pal, S. Bose, G. K. Basak, and A. Mukhopadhyay, "Improved speaker identification by Aggregating Gaussian Mixture Models (GMMs)," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 28, no. 4, 2014.
[8] C. R Rao, *Linear Statistical Inference and Its Applications*, 2nd (reprint) edition, John Wiley & Sons, New York, 2001.
[9] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Pearson Education, Inc, 2008.
[10] S. B. Davies and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, pp. 357-366, 1980.
[11] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, pp. 1-38, 1977.
[12] P. J. Huber, *Robust Statistics*, Springer Berlin Heidelberg, 2011.
[13] A. Pal, S. Bose, M. Mitra, and S Roy, "ISIS and NISIS: New Bilingual dual-channel speech corpora for robust speaker recognition," in *Proc. the 2012 International Conference on Image Processing, Computer Vision, & Pattern Recognition (IPCV-2012)* , 2012, Vol. I, pp. 936-939.

Smarajit Bose received B. Stat. and M. Stat. degrees from the Indian Statistical Institute, Calcutta in 1984 and 1986 respectively, and a Ph. D degree from the University of California, Berkeley in 1992, all in Statistics. Between 1991 and 1992, he was a visiting scientist in IBM Almaden Research Center, San Jose where he worked with the Advance Process Monitoring Group. He visited the Statistics Departments of the University of Washington, Seattle during 1992–1993 and the Ohio State University, Columbus during 1993–1996 where he worked on classification and clustering problems and their applications in medical imaging and ergonomics. In 1996, he joined the Indian Statistical Institute, Calcutta and now is a professor in the applied statistics division of the Institute. He has also visited the University of California, Santa Barbara in 2001, and also in 2002–2003. His current research interests are pattern recognition and its applications in image processing and speaker identification.

Amita Pal obtained a B.Sc. degree with honours and a M.Sc. degree in statistics from the University of Calcutta in 1979 and 1981 respectively, and a Ph.D degree in statistics from the Indian Statistical Institute, Kolkata in 1991. She joined the Indian Statistical Institute as a lecturer in 1994 and is currently working as a professor in the Interdisciplinary Statistical Research Unit of the same Institute. She visited the Imperial College of Science, Technology and Medicine, London in 1994 on a six-month UNDP fellowship. Her research interests included pattern recognition, image processing and statistical machine learning.

Anish Mukherjee received a bachelor of science (B.Sc.) degree in statistics from St. Xavier's College, Kolkata in 2011, followed by a B.Tech. degree in computer science and engineering from the University of Calcutta in 2014. He had been involved in research on robust speaker identification during 2014-16 in the Interdisciplinary Statistical Research Unit of the Indian Statistical Institute, Kolkata. His areas of interest are pattern recognition, computer vision, high-dimensional data analysis and Bayesian nonparametrics. He is currently pursuing a Ph. D. degree in statistics in University of Missouri, Columbia, USA.

Debasmita Das received a bachelor of science (B.Sc.) degree in statistics from Ashutosh College, Kolkata in 2012 and a master of science (M.Sc.) degree in statistics from University of Calcutta in 2014. She has been involved in research on robust speaker identification during 2014-16 in the Interdisciplinary Statistical Research Unit of the Indian Statistical Institute, Kolkata. Her areas of interest are pattern recognition, computer vision, high-dimensional data analysis and biostatistics. She is currently pursuing a Ph. D degree in statistics in University of Connecticut, Storrs, USA.