Digital Signal Processing 50 (2016) 1–11

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

# Local spectral variability features for speaker verification

Md Sahidullah *, Tomi Kinnunen

*Speech and Image Processing Unit, School of Computing, University of Eastern Finland, P.O. Box 111, FI-80101 Joensuu, Finland*

ARTICLE INFO

*Article history:*
Available online 18 November 2015

*Keywords:*
Feature extraction
Covariance analysis
Eigenstructure
Mel-frequency cepstral coefficients (MFCC)
Speaker verification

ABSTRACT

Speaker verification techniques neglect the short-time variation in the feature space even though it contains speaker related attributes. We propose a simple method to capture and characterize this spectral variation through the eigenstructure of the sample covariance matrix. This covariance is computed using sliding window over spectral features. The newly formulated feature vectors representing local spectral variations are used with classical and state-of-the-art speaker recognition systems. Results on multiple speaker recognition evaluation corpora reveal that eigenvectors weighted with their normalized singular values are useful in representing local covariance information. We have also shown that local variability features can be extracted using mel frequency cepstral coefficients (MFCCs) as well as using three recently developed features: frequency domain linear prediction (FDLP), mean Hilbert envelope coefficients (MHECs) and power-normalized cepstral coefficients (PNCCs). Since information conveyed in the proposed feature is complementary to the standard short-term features, we apply different fusion techniques. We observe considerable relative improvements in speaker verification accuracy in combined mode on text-independent (NIST SRE) and text-dependent (RSR2015) speech corpora. We have obtained up to 12.28% relative improvement in speaker recognition accuracy on text-independent corpora. Conversely in experiments on text-dependent corpora, we have achieved up to 40% relative reduction in EER. To sum up, combining local covariance information with the traditional cepstral features holds promise as an additional speaker cue in both text-independent and text-dependent recognition.

© 2015 Elsevier Inc. All rights reserved.

## 1. Introduction

Speaker verification systems use speech features extracted from short-term power spectrum [1]. Commonly used short-term spectral features, such as mel-frequency cepstral coefficients (MFCCs) [2] and perceptual linear prediction (PLP) [3] features, are extracted from speech segments of 20–30 ms duration and they represent spectral characteristics associated with the speech segment [4]. But temporal variation of spectrum also contains useful information about the dynamics of the speech production system. A common way to incorporate this information is to augment delta and double-delta coefficients with the static features computed over a temporal window of 50–100 ms [5,6]. MFCCs along with deltas and double-deltas remain as the primary features in state-of-the-art speaker verification, due to reasonably high recognition accuracy and straightforward computation. Subsequently, this has sparked great research interest into further ideas such as feature post-processing. For example, cepstral mean and variance normalization (CMVN) [7] and feature warping [8] help to suppress channel and session variations. Different computational blocks of MFCC algorithms have also been explored. For instance, [9] used alternative multiple windowing technique in place of the conventional Hamming window while [10] used regularized linear prediction (LP) analysis for power spectrum estimation. Classical triangular filter bank in MFCC can be replaced with Gaussian-shaped filters [11], gammatone filters [12] and cochlear filters [13]. Root compression technique is prescribed for reducing the dynamic range of mel filter energies as opposed to the logarithmic compression [14]. An improved transformation technique on filter bank log-energies is proposed in [15] which was reported to yield higher recognition accuracy compared to conventional discrete cosine transform (DCT) in clean and noisy conditions.

Recently, further investigations have been carried out for extracting new features for speaker recognition [16,14] that utilize internally some form of *long-term processing* before extracting the short-term features. For instance, in *frequency domain linear prediction* (FDLP) [17,18], the speech signal is first transformed into frequency domain with DCT operation directly on the speech signal. The subband Hilbert envelopes are computed followed by short-term energy computation from each band. In a more recent work, short-term features called *mean Hilbert envelope coefficients* (MHECs) are proposed from subband Hilbert envelope of

* Corresponding author.
*E-mail addresses:* sahid@cs.uef.fi (M. Sahidullah), tkinnu@cs.uef.fi (T. Kinnunen).

http://dx.doi.org/10.1016/j.dsp.2015.10.011
1051-2004/© 2015 Elsevier Inc. All rights reserved.

auditory filter output [14]. Here gammatone filter are employed simulating the effect of auditory nerve. Both the FDLP and MHEC features were reported to give high accuracy in both clean and noisy conditions. Another feature set, *power-normalized cepstral coefficients* (PNCCs), was recently proposed for robust speech recognition [19] and subsequently applied to speaker recognition with success [20]. A common characteristic of these long-term processing ideas, from a practical point of view, is that they have a large number of user-definable parameters that should be carefully chosen, and the settings for different environmental effects and conditions vary widely [16,21,22,14,19]. This makes the end-users task difficult when finding best feature configuration for a certain environment. In this paper, we introduce a new feature extraction technique which models the *local feature-space variability* and can be computed from any spectral features, similar to delta features. The variability of features is calculated directly from the *covariances of the pre-computed cepstral features*.

The use of covariance information has a long history in speech processing and speaker verification is no exception. Since the speech signal varies a lot depending on spoken content, channel, background noise and various other situational parameters, the acoustic features computed from the signal for the same speaker are never exact replicas across training and test utterances. To compensate for such nuisance variations, the speaker and language community has put considerable effort into (co)variance modeling of features and speaker models [23,24]. In the classic techniques, uncertainty of speaker means is captured by covariance matrices in a Gaussian mixture model (GMM) [25]. In state-of-the-art systems, covariance modeling plays a major role at the later stages of the recognizer pipeline. For instance, nuisance attribute projection (NAP) [26] and within-class covariance normalization (WCCN) [27] utilize, respectively, the estimated channel and within-speaker covariance matrices to suppress the respective effects from GMM supervectors [26] or i-vectors [28]. Similarly, taking into account the uncertainty propagation at the PLDA model [29] helps to improve speaker recognition score with the use of posterior covariance estimation.

In most of the above-cited studies, covariance information has been used as a secondary tool for the purpose of suppressing nuisance variations from the primary acoustic features (such as MFCCs) or higher-level compact representations derived from them (such as i-vectors). In contrast to these prior studies, a new viewpoint of our work is a study of covariance *features* for speaker characterization. To this end, the proposed features are obtained using a low-cost procedure from time-localized covariance information of arbitrary acoustic features, such as MFCCs. To this end, our input acoustic features include not only standard MFCC features but also the recently studied alternative parameterizations, so-called FDLP, MHEC and PNCC. Our method is inspired by the successful use of covariance-based features in applications outside of speech technology, such as movement detection and image classification [30–32], blind source separation [33], anomaly detection in a network [34], similarity analysis of multivariate time-series [35] and brain-computer interfacing applications [36]. To this end, the intention of the present study is to provide a feasibility study of such features for speaker characterization. We first motivate and detail our proposed approach in Sections 2 and 3. We describe the experimental set up in Section 4 followed by Section 5 that provides extensive experimentation on three of the standard NIST speaker recognition evaluation (SRE) corpora (2001, 2008 and 2010) and recently released RSR2015. Section 6 provides a summary of our findings. Finally, for reproducibility and to spark further research interest to this direction, we provide an open-source implementation of the proposed method.[1]

## 2. Local variability features: motivation

In speaker recognition, the total variation in feature space is captured by the covariances computed over all the features. But this neglects the variations of the features for a short time duration during the articulation of various speech segments. A previous study has suggested that these variations might be more related to the spoken text [37]. But as each individual has his or her own unique articulatory behavior even for the same spoken content, we argue that measuring that variation could be useful for speaker characterization. To this end, our features are a low-dimensional parameterization of the short-term covariance matrix. A similar method is used in image processing applications where the segments of an image are described by covariance matrices of features of the region, known as *region covariance* [30,32]. The property of local-covariance matrix is also explored in other applications with reasonable success [33–36]. However, to the best of our knowledge, this has not been explored yet for speaker characterization. Here, we first analyze the property of local covariance with different known speech segments of different the speakers. We further provide an analysis how the local covariance is related to the global covariance matrix. We have also analyzed the property of local-covariance in the presence of different kinds of speech segments. This leads to the derivation of feature vector in the next section.

### 2.1. Analysis of covariance for different speech segments

In Fig. 1, we have shown global covariances of three different speakers from NIST SRE 2001. The utterances are *long* and have more than two minutes of speech data. To compute covariance, we consider first two dimensions of MFCC, excluding the energy coefficient. The global covariance does not vary much even when the utterances contain completely different contents (and channels). Now, when CMVN is applied for reducing the convolutional channel effect in feature space, the covariances become even more similar, which indicates that global covariance may not be too effective for speaker characterization. Indeed, some of the early speaker modeling techniques using global covariance for speaker modeling (e.g. [24,23]) have generally been found less effective than methods that rely on speaker means.

On the other hand, the geometric structure of *short-term* sample covariance matrices are illustrated for ten different words of TIMIT sentence SA2 of three different speakers in Fig. 2. To extract the local covariance, we use the TIMIT word-level annotations. Clearly, local covariances for different speakers are visually more distinguishable even though they correspond to the same spoken text. This motivates us to explore ways to parameterize the short-term covariance into sequences of feature vectors, to be used with arbitrary classifier back-ends. Before presenting this in Section 3, we shall first elaborate on the relationship of global and local (segment-dependent) covariances.

### 2.2. Relationship between the global covariance and local covariances

Let the $t$-th cepstral feature vector from a speech utterance be $\mathbf{x}_t$. Then the sample estimator of the global covariance matrix of entire feature space with $T$ frames is,

$$\mathbf{C}^{\text{global}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t^{\top} - \boldsymbol{\mu}\boldsymbol{\mu}^{\top}, \tag{1}$$

[1] http://cs.joensuu.fi/~sahid/codes/local_variability.zip.

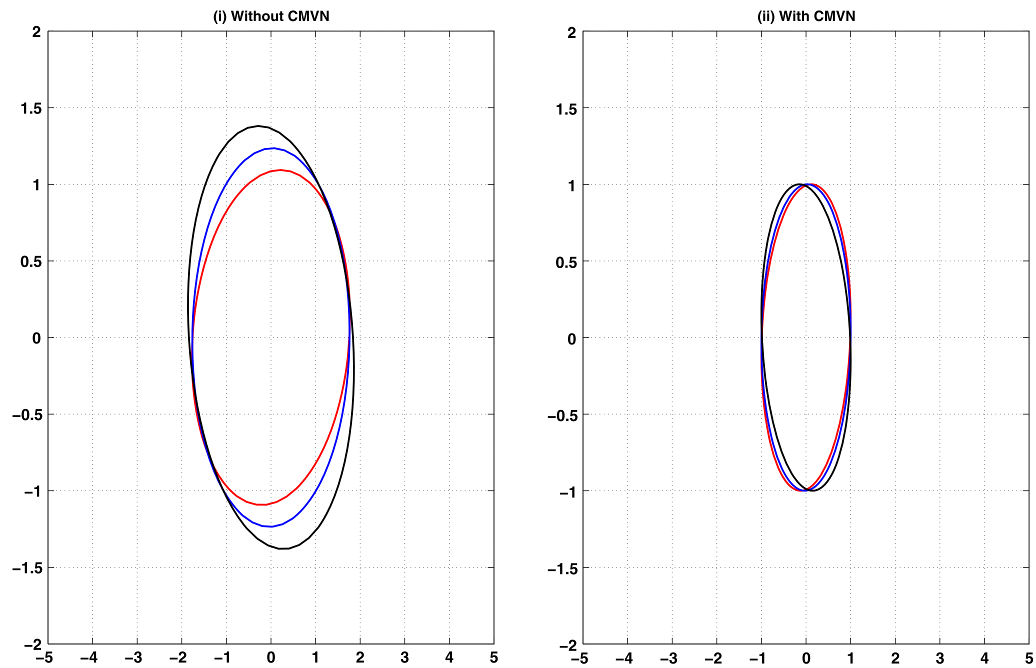


**Fig. 1.** Plot of global covariance matrices of three different speakers of NIST SRE for (i) without CMVN and (ii) with CMVN processing. First two dimensions of feature vector are chosen for this visualization.

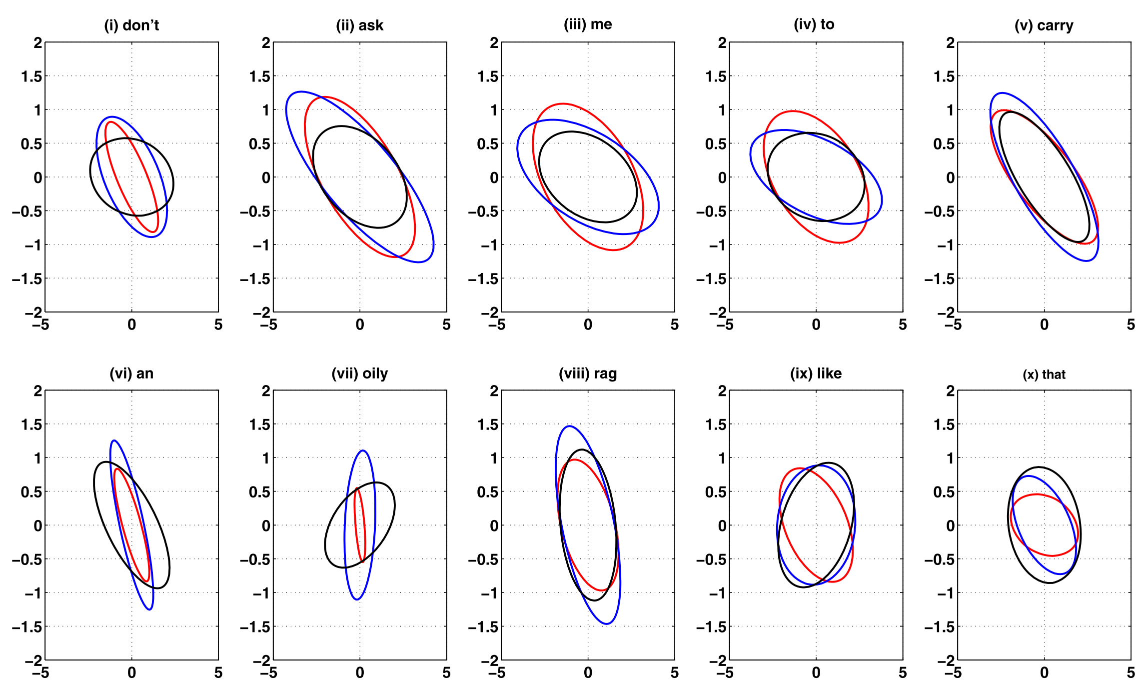


**Fig. 2.** Plot of covariance matrices of three different female speakers corresponding to different words of SA2 sentences ("*Don't ask me to carry an oily rag like that*") in TIMIT corpus. First two dimensions of feature vector are chosen for this visualization.

where $\boldsymbol{\mu}$ is the global mean. If the utterance is divided into $Q$ non-overlapping segments (not necessarily of same length), the sample covariance of $q$-th segment $S_q$ is,

$$\mathbf{C}_q^{\text{local}} = \frac{1}{|S_q|} \sum_{t \in S_q} \mathbf{x}_t \mathbf{x}_t^\top - \boldsymbol{\mu}_q \boldsymbol{\mu}_q^\top, \tag{2}$$

where $|S_q|$ is the number of samples in the $q$-th segment. Now we can write,

$$\sum_{t \in S_q} \mathbf{x}_t \mathbf{x}_t^\top = |S_q| \mathbf{C}_q^{\text{local}} + \boldsymbol{\mu}_q \boldsymbol{\mu}_q^\top. \tag{3}$$

From Eq. (1), we find that,

$$\mathbf{C}^{\text{global}} = \frac{1}{T} \left[ \sum_{t \in S_1} \mathbf{x}_t \mathbf{x}_t^\top + \sum_{t \in S_2} \mathbf{x}_t \mathbf{x}_t^\top + \ldots + \sum_{t \in S_q} \mathbf{x}_t \mathbf{x}_t^\top \right] - \boldsymbol{\mu}\boldsymbol{\mu}^\top. \tag{4}$$

Hence from Eqs. (3) and (4), we get,

$$\mathbf{C}^{\text{global}} = \frac{1}{T} \Big[ |S_1| \mathbf{C}_1^{\text{local}} + \boldsymbol{\mu}_1 \boldsymbol{\mu}_1^\top + |S_2| \mathbf{C}_2^{\text{local}} + \boldsymbol{\mu}_2 \boldsymbol{\mu}_2^\top + \ldots + |S_q| \mathbf{C}_q^{\text{local}} + \boldsymbol{\mu}_q \boldsymbol{\mu}_q^\top \Big] - \boldsymbol{\mu}\boldsymbol{\mu}^\top. \tag{5}$$

Finally, we can express the global covariance as,

$$\mathbf{C}^{\text{global}} = \frac{1}{T}\sum_{q=1}^{Q}|S_q|\,\mathbf{C}_{\text{local}}^{q} + \frac{1}{T}\sum_{q=1}^{Q}\boldsymbol{\mu}_q\boldsymbol{\mu}_q^{\top} - \boldsymbol{\mu}\boldsymbol{\mu}^{\top}. \tag{6}$$

From here we find that when mean-normalization in feature space is performed, $\boldsymbol{\mu}$ becomes 0 and $\sum_{q=1}^{Q}\boldsymbol{\mu}_q\boldsymbol{\mu}_q^{\top}$ is directly related to the between-segment covariance ($C^{\text{bsegs}}$), as given by

$$\mathbf{C}^{\text{bsegs}} = \frac{1}{Q}\sum_{q=1}^{Q}\boldsymbol{\mu}_q\boldsymbol{\mu}_q^{\top}. \tag{7}$$

Note that this $C^{\text{bsegs}}$ is analogous to between-class covariance matrix used in linear discriminant analysis [38]. Now from Eq. (6) and Eq. (7), we can express the global covariance as,

$$\mathbf{C}^{\text{global}} = \frac{1}{T}\sum_{q=1}^{Q}|S_q|\,\mathbf{C}_{q}^{\text{local}} + \frac{Q}{T}\mathbf{C}^{\text{bsegs}}. \tag{8}$$

Therefore, after mean-normalization over the entire feature space, the global covariance is nothing but the linear combination of the local covariances (*i.e.*, within-segment covariance matrix) and between-segment covariance matrix.

## 3. Eigenstructure features from local covariance matrix

### *3.1. Local spectral variation from short-term covariance*

The short-term spectral features of a speech utterance for different frames can be viewed as a multivariate time-series where one spectral frame represents a "snapshot" of the speech production system. The variations in this multivariate data can be measured by computing its covariance matrix [39]. In conventional Gaussian mixture modeling (GMM) that underlies in both classic [25,40] and modern recognizers [28], covariance of each mixture component represents spectral variability within the respective acoustic class. In contrast, we consider the short-term sample covariance matrix computed over a short-segment of speech (about 5 to 11 frames), and parameterize it as a feature vector for use with any recognizer back-end.

The diagonal elements of this short-term covariance matrix (*i.e.*, sample variance of each feature over the temporal window) correspond to feature variation across the frames. They were found useful for speaker characterization in [41]. The off-diagonal elements, representing co-variation in MFCCs, have generally smaller values due to use of (global) decorrelation technique, such as the discrete cosine transform (DCT) in MFCC extraction. However, as DCT achieves perfect decorrelation only when the mel-filter bank log-energies follow a *first-order Markov process* [42], the off-diagonals are also useful for characterizing spectral variations. For steady sounds, such as sustained vowels, the first-order Markov property is reasonable but segments containing relatively more variable spectral contents, such as unvoiced fricatives, stop consonants and diphthongs, will yield non-negligible diagonal elements. We will now describe the proposed method which aims at preserving the important characteristics of the local covariance matrix.

### *3.2. Features from short-term covariance*

Let a sliding window of spectral features be denoted by a $d \times N$ matrix $\mathbf{X}$ centered around the $t$-th speech frame containing $d$-dimensional features in each column corresponding to $N = (2L+1)$ frames. That is,

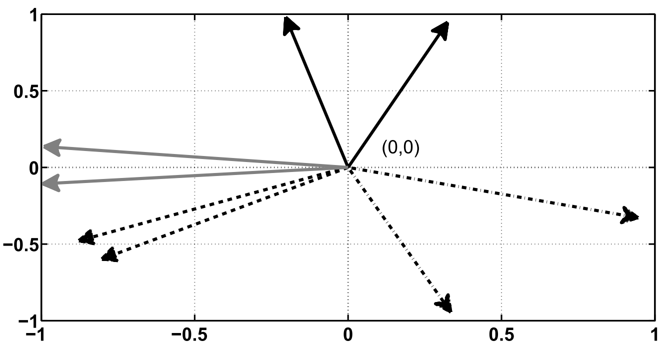


**Fig. 3.** Graph of eigenvectors corresponding to the highest eigenvalues of local covariance matrices for four different speakers (in separate line styles). The data corresponds to the same phoneme /ae/ from SA2 sentences ("*Don't ask me to carry an oily rag like that*") in TIMIT corpus. For each speaker, two separate instances of the same sound are shown with identical line styles.

$$\mathbf{X} = \begin{bmatrix}\mathbf{x}_{t-L} & \ldots & \mathbf{x}_t & \ldots & \mathbf{x}_{t+L}\end{bmatrix}, \tag{9}$$

where each $\mathbf{x}_t$ denotes $d$-dimensional column vector representing spectral feature of $t$-th frame. The sample covariance matrix is $\mathbf{C} = \frac{1}{N-1}\sum_{t=1}^{N}(\mathbf{x}_t - \bar{\mathbf{x}})(\mathbf{x}_t - \bar{\mathbf{x}})^{\top}$, where $\bar{\mathbf{x}}$ is the sample mean, $\bar{\mathbf{x}} = \frac{1}{N}\sum_{t=1}^{N}\mathbf{x}_t$. Now, $\mathbf{C}$ is a $d \times d$ matrix which contains information related to variation of spectral features. As the effect of mean subtraction, only the variable component of spectral features are retained in $\mathbf{C}$. Due to practical limitations, elements of the covariance matrix would form a rather poor parameterization. We employ *eigen-decomposition* property [43] where a positive semi-definite covariance matrix $\mathbf{C}$ is uniquely represented by its eigenvectors and -values as,

$$\mathbf{C}\mathbf{e}_i = \lambda_i\mathbf{e}_i, \qquad i = 1, 2, 3, \ldots, d \tag{10}$$

where $\lambda_i$ is the eigenvalue corresponding to the $d$-dimensional eigenvector $\mathbf{e}_i$. Geometrically, the covariance matrix $\mathbf{C}$ corresponds to a prediction ellipse of a multivariate Gaussian that can be represented by its semi-axes. Specifically, the direction of $i$-th semi-axis is determined by the eigenvector $\mathbf{e}_i$, and its magnitude is the respective *singular value*, $s_i = \sqrt{\lambda_i}$. In our present setup, short-term features of 19-dimensions are computed from speech frame of 20 ms with an overlap of 10 ms. Hence, in order to keep the sliding window at 110 ms, $N$ needs to be fixed at 11. Note that $N < d$ in all the cases considered in this paper since we use temporal windows having length at most 130 ms. Consequently, the rank of the sample covariance matrix $\mathbf{C}$ is at most 13 assuming all the observations are linearly independent [44, p. 103]. That is, $\mathbf{C}$ is always rank-deficient for which it is non-invertible [44, p. 51]. However, this is not a concern as we do *not* need the inverse of $\mathbf{C}$ at any point. Instead, we parameterize the covariance matrix via its eigenvalues and -vectors and treat it as a feature vector, analogous with MFCCs. To this end, note first that at most $N$ of the eigenvalues are nonzero; secondly, only the few top eigenvalues are significant as features of close-by frames are highly correlated. Therefore, eigenstructure of $\mathbf{C}$ can be represented with a fewer eigenvalues.

In Fig. 3, the first two eigenvectors corresponding to the multiple instances of phoneme /ae/ are illustrated for four different speakers. Those speakers are separable by the direction (specified by the eigenvectors). On the other hand, the singular values can be seen as measures of spectral activity. For example, speech regions with less variation will have lower singular values as the covariance matrix is nearly a null matrix (with all singular values equal to zero in the limiting case). For rapidly varying regions,

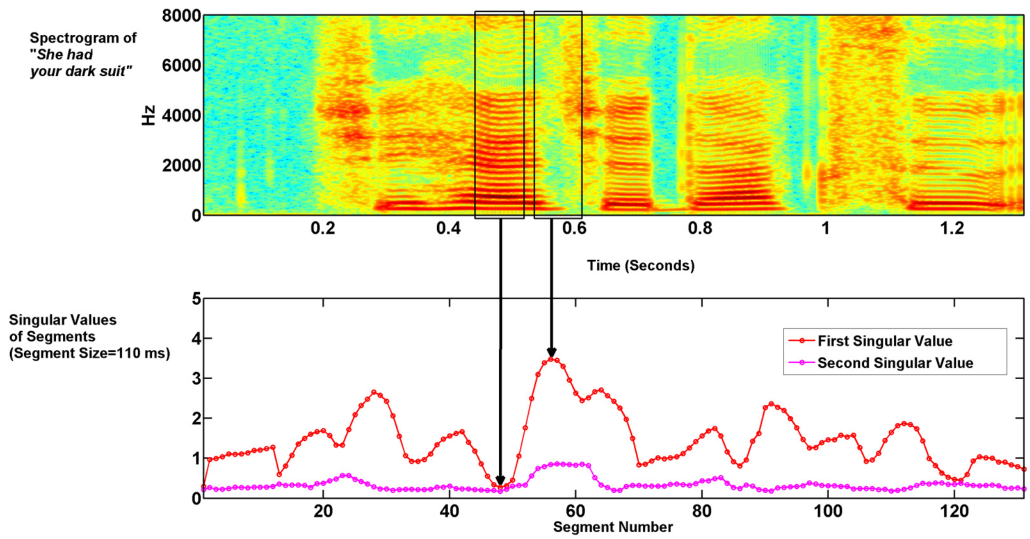


**Fig. 4.** Figure showing speech spectrogram of "*She had your dark suit*" (top) and the corresponding first two singular values of covariances (bottom). Here, the covariance is computed for 110 ms sliding window (*i.e.*, number of frame, $N = 11$ for window size 20 ms with an overlap of 50%) in temporal domain. The figure shows that portions of the speech segment with "stable" spectral information have lower eigenvalue where as regions with "highly varying" speech information correspond to higher singular values.

however, the singular values will be higher as they correspond to the standard deviations in the directions of the eigenvectors. In Fig. 4, variation of two highest singular values are illustrated for one speech segment. Clearly, regions having slowly varying or "stable" spectral characteristics under a temporal window of 110 ms, have lower singular values in comparison with the "rapidly varying" section of spectrum.

To incorporate information from both the eigenvectors and singular values, we proceed as follows. Assuming $\{\mathbf{e}_i\}_{i=1}^{K}$ are the eigenvectors corresponding to the $K$ largest eigenvalues, our proposed $dK$-dimensional feature is,

$$\mathbf{f} = \left[\mathbf{f}_1^\top \mathbf{f}_2^\top \ldots \mathbf{f}_K^\top\right]^\top \tag{11}$$

where $\mathbf{f}_i = \alpha_i \mathbf{e}_i$ are *weighted* eigenvectors. We consider three kinds weighting schemes:

1. $\alpha_i = 1$,
2. $\alpha_i = s_i$,
3. $\alpha_i = s_i / \sum_{n=1}^{d} s_n$.

The first variant discards any information of the singular values and we call it *uniformly weighted eigenvector coefficient* (UWEC). In the second case, eigenvectors are weighted with the corresponding singular value, leading to *singular value weighted eigenvector coefficient* (SWEC). In the last case, the weights are further normalized by the sum of singular values or *trace-norm* to incorporate the influence of discarded eigenvectors. We call it *normalized singular value weighted eigenvector coefficient* (NSWEC).

In practice, we use *singular value decomposition* (SVD) to simultaneously compute singular values and eigenvectors [45]. SVD represents a $d \times N$ rectangular matrix $\tilde{\mathbf{X}}$ as a multiplication of three matrices, *i.e.*, $\tilde{\mathbf{X}} = \mathbf{USV}^\top$, where $\mathbf{U}$ is a $d \times d$ orthogonal matrix containing eigenvectors of $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top$, $\mathbf{V}$ is an $N \times N$ orthogonal matrix containing eigenvectors of $\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}$, and $\mathbf{S}$ is a diagonal matrix containing the singular values of both $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top$ and $\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}$ [46]. Now if $\tilde{\mathbf{X}}$ has zero mean in row space, $\mathbf{U}$ will represent eigenvectors of $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top$ (*i.e.*, covariance matrix) and $\mathbf{S}$ will contain the corresponding singular values. Therefore, the steps to calculate the new features from any feature matrix $\mathbf{X}$ of size $d \times N$ are:

**Step 1:** Compute normalized feature matrix $\tilde{\mathbf{X}}$ by (i) subtracting sample mean $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$ from each column of $\mathbf{X}$, (ii) dividing them by $\sqrt{N-1}$.

**Step 2:** Perform SVD: $\tilde{\mathbf{X}} = \mathbf{USV}^\top$, where $\mathbf{U} \in \mathbb{R}^{d\times d}$, $\mathbf{S} \in \mathbb{R}^{d\times N}$ and $\mathbf{V} \in \mathbb{R}^{N\times N}$.

**Step 3:** Get the singular values $s_i$ from the diagonal of $\mathbf{S}$. Likewise, the $i$-th row of $\mathbf{U}$ is $\mathbf{e}_i$.

**Step 4:** Form the feature vector $\mathbf{f}$ by concatenating $\mathbf{f}_i = \alpha_i \mathbf{e}_i$ for $i = 1, 2, \ldots, K$ with appropriate value of $\alpha_i$.

Note that the SVD step requires additional computations with time complexity $O(\min\{Nd^2, N^2d\})$ per frame [47] for the commonly used implementation. But there are faster algorithms to compute SVD, specifically for our case, where singular vectors corresponding to the top $K$ singular values are only required [45,48,49]. Computational cost can be further reduced here as SVD is calculated on the data using sliding window [50]. Another issue with SVD is that the inherent sign ambiguity associated with its decomposition can be a problem [51]. However in practice, it will not affect the recognition performance, if exactly same implementation of robust SVD algorithm[2] that gives deterministic output is employed for training and testing.

## 4. Experimental setup

### 4.1. Database description

We evaluate speaker verification accuracy on three NIST corpora. First, we perform extensive experiments on NIST SRE 2001[3] to find out optimal parameter configurations. Then we apply it on the telephone sub-conditions of NIST SRE 2008[4] and 2010.[5] We have selected C6 sub-condition from NIST SRE 2008 containing all the telephone speech trials. From NIST SRE 2010, we have chosen C5 and C6. Here, C5 corresponds to normal vocal effort in both enrolment and verification samples while in C6 the target speakers

[2] For example, SVD implementation in MATLAB which uses LAPACK (Linear Algebra Package). We have also found that this algorithm is stable and signs of the singular vectors are not affected due to small amount of random perturbation [52].

[3] http://www.itl.nist.gov/iad/mig/tests/spk/2001/.

[4] http://www.itl.nist.gov/iad/mig/tests/sre/2008/.

[5] http://www.itl.nist.gov/iad/mig//tests/sre/2010/.

**Table 1**
Description of NIST speech corpora used for the performance evaluation. (♂: Male, ♀:Female.)

| | NIST SRE 2001 | NIST SRE 2008 (C6) | NIST SRE 2010 (C5) | NIST SRE 2010 (C6) |
|---|---|---|---|---|
| Target models | 74♂, 100♀ | 648♂, 1140♀ | 290♂, 290♀ | 181♂, 184♀ |
| Test segments | 850♂, 1188♀ | 895♂, 1674♀ | 355♂, 357♀ | 147♂, 185♀ |
| Target trials | 850♂, 1188♀ | 874♂, 1840♀ | 353♂, 355♀ | 178♂, 183♀ |
| Non-target trials | 8500♂, 11 880♀ | 11 637♂, 21581♀ | 13 707♂, 15 958♀ | 12 825♂, 15 486♀ |

**Table 2**
Description of the RSR2015 speech corpus (Part I) used for the performance evaluation. (♂: Male, ♀:Female.)

| | | Development | Evaluation |
|---|---|---|---|
| Target models | | 1492♂, 1405♀ | 1708♂, 1470♀ |
| Test segments | | 8979♂, 8448♀ | 10 256♂, 8810♀ |
| Target trials | Target Correct (TC) | 8931♂, 8419♀ | 10 244♂, 8810♀ |
| Non-target trials | Target Wrong (TW) | 259 001♂, 244 123♀ | 297 076♂, 255 490♀ |
| | Impostor Correct (IC) | 437 631♂, 387 230♀ | 573 664♂, 422 880♀ |
| | Impostor Wrong (IW) | 6 342 019♂, 5 612 176♀ | 8 318 132♂, 6 131 760♀ |

are enrolled with normal vocal effort but tested with high vocal effort speech [53]. The details of the NIST corpora are summarized in Table 1.

We have also performed experiments with the recently released text-dependent RSR2015 corpus [54]. It contains trials with lexical constraints in training and test. Part 1 of the database is used for the experiments where the speakers use a fixed pass-phrase for authentication. The details of the Part 1 of the corpora are summarized in Table 2. It consists of four different kinds of trials. The first type of trial, target correct (TC), consists of target speakers tested with correct pass-phrase from the same speaker, considered as the target trial. There are three different non-target trials. The first one is target wrong (TW) where the same speakers with different pass-phrase try to authenticate. The two remaining ones are impostor correct (IC) and impostor wrong (IW), where impostor speakers try to authenticate, respectively, using correct and wrong pass-phrases.

### 4.2. Feature extraction

Short-term spectral features are extracted from speech frames of 20 ms with 50% overlap. The Hamming window is used for discrete Fourier transform (DFT) based power spectrum estimation. Baseline MFCCs are extracted first using 20 triangular filters in mel scale [15]. Discarding the energy coefficient, the remaining 19 coefficients are processed further with relative spectral (RASTA) filtering [55]. Then delta and double-delta coefficients are computed with temporal window of three frames and augmented with the static coefficients to create 57-dimensional feature vector. Then, features corresponding to non-speech frames are discarded using a speech activity detection (SAD) technique that utilizes bi-Gaussian modeling of log-energies [56]. Finally, utterance level *cepstral mean and variance normalization* (CMVN) is performed. The proposed local covariance based features are extracted using the static part of the MFCCs (after processing with RASTA and CMVN) using a sliding window of fixed length in the temporal domain.

The proposed features are also extracted using other recently studied features to extract the base coefficients. MHEC [14], FDLP [21], and PNCC [19] features are implemented with the optimized configurations reported in literature. In MHEC, first the speech signal is passed through a gammatone filter bank consisting of 32 filters. Then the mean energy of Hilbert envelope of each subband is computed. Finally, DCT is performed on 15th root compressed energy coefficients to compute 20-dimensional cepstral features. We have also appended delta and double-delta coefficients to get final 60 dimensional feature vector as in [14]. In the case of FDLP, first the speech signal is divided into very long segments of length 10 s. Then, DCT is performed to transform the signal into frequency domain. After this, linear prediction analysis of order 30 is performed for each of the 17 subbands spaced linearly in Bark scale. Then, short-term energy of each envelope is computed and they are used for creating 13-dimensional cepstral features using DCT (discarding the energy coefficient). Finally, delta and double-delta coefficients are added to create 39-dimensional features. In PNCC, we use a temporal window of five frames. 32 filters are used to compute the energy coefficients which further undergo 15th root power compression. Finally, we get 57-dimensional features similar to our MFCCs. We set the frame size and the frame shift same for all the compared features. CMVN is performed in all cases. For all the compared base feature sets, we compute the NSWEC features in the same manner as from the MFCCs using a window of 60 ms.

### 4.3. Classifier description

We evaluate speaker recognition performance using two different classifiers. First, to enable a large number of preliminary experiments, we use the classic lightweight Gaussian mixture model with universal background model (GMM–UBM) on NIST SRE 2001. The main purpose is to find an optimized configuration for our proposed feature. We then use an up-to-date i-vector [28] recognizer with *probabilistic linear discriminant analysis* (PLDA) [57,58] back-end to assess the recognition accuracies of the baseline MFCCs and proposed features in the two newer NIST corpora. In the i-vector system, two gender-dependent UBMs with 512 mixture components are trained using 20 EM iterations with speech data from NIST SRE 2004–06, FISHER, and Switchboard. Then, a total variability matrix with 400 factors is trained using five EM iterations from the same data. The i-vectors are processed using linear discriminant analysis (LDA) to reduce their dimensions to 200, followed by radial Gaussianization [58]. For PLDA training, the same data as in T-matrix estimation is utilized. The dimensionality of the speaker subspace is set to 150 and 20 EM iterations are used for estimating the PLDA hyper-parameters. For performing experiments with RSR2015, we have used GMM–UBM system as the speech files are short in duration and i-vector shows poor performance [54]. In this case, gender-dependent UBM is trained using TIMIT corpora and target speakers are created using maximum-a-posteriori (MAP) algorithm with relevance factor 14 [40]. The reason for selecting TIMIT is that it also contains microphone quality speech signals with sampling frequency of 16 kHz similar to that of RSR2015. In order to conduct experiments with this corpus, we have made necessary changes to the cepstral feature extractor used for NIST evaluation, *i.e.*, the sampling rate is set at 16 kHz.

**Table 3**
Speaker verification performance on NIST SRE 2001 using a GMM–UBM system for the baseline MFCCs and the proposed eigenstructure features. Temporal window length is set to 100 ms. Top three eigenvectors are chosen, *i.e.*, $K = 3$. Dimensionality of all the four feature sets is the same, 57.

| Feature (dimensionality) | EER (in %) | minDCF × 100 |
|---|---|---|
| MFCC (57) | 8.01 | 3.66 |
| UWEC (57) | 9.86 | 4.25 |
| SWEC (57) | 10.05 | 4.51 |
| NSWEC (57) | **9.42** | **4.12** |

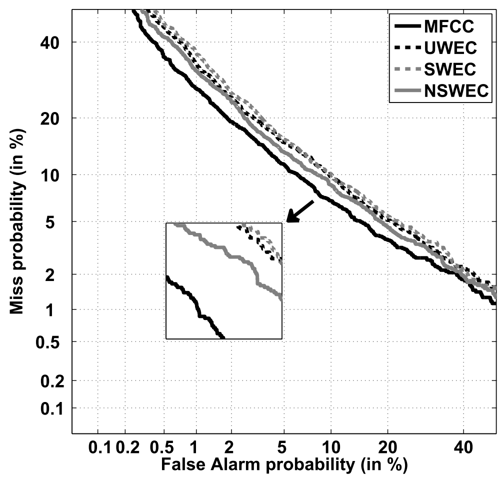


**Fig. 5.** DET plot showing speaker verification performance on NIST SRE 2001 for 57-dimensional MFCC (static + Δ + ΔΔ) and three 57-dimensional proposed features computed from first three eigenvectors using temporal window of 100 ms.

However, other parameters including the number of static features are kept identical as before.

### 4.4. Performance evaluation

We use equal error rate (EER) and minimum detection cost function (minDCF) to assess speaker recognition accuracy. EER is calculated when the false alarm ($P_{\text{fa}}$) and the false rejection rates ($P_{\text{miss}}$) are equal, whereas minDCF is the minimum of $w_{\text{miss}} \times P_{\text{miss}} + w_{\text{fa}} \times P_{\text{fa}}$ over all detection thresholds. Here, $w_{\text{miss}}$ and $w_{\text{fa}}$ are weights for the miss and false alarm rates. These values are set according to the evaluation plans of the respective corpora. For NIST SRE 2001 and RSR2015, $w_{\text{miss}} = 0.10$ and $w_{\text{fa}} = 0.99$ while for NIST SRE 2008 and 2010, $w_{\text{miss}} = 0.001$ and $w_{\text{fa}} = 0.999$.

## 5. Results

We first study the newly proposed eigenstructure features for an arbitrarily chosen temporal window length (here 100 ms). As in our present setup, the speech frame size is 20 ms with 10 ms overlap, we consider nine frames (*i.e.*, context of four frames in each direction) for covariance computation. In order to keep the dimensionality of the proposed feature same to that of the baseline MFCCs (*i.e.*, 57), only the first three eigenvectors corresponding to the highest eigenvalues are considered (*i.e.*, $K = 3$). Different weighting schemes along with the baseline MFCCs are compared on NIST 2001 in Table 3, and the corresponding DET plots are shown in Fig. 5. Features based on normalized singular value weighting (NSWEC) outperforms the other two variants, UWEC and SWEC, in terms of both EER and minDCF. The proposed features yield generally higher error rates compared with our MFCC baseline but, as we will see shortly, they capture complementary cues to MFCCs that help in a fusion mode.

**Table 4**
Speaker verification performance on NIST SRE 2001 using the proposed NSWEC features with GMM–UBM system for different length of temporal window and number of eigenvectors (*i.e.*, $K$). Note that for 40 ms temporal window (*i.e.*, using three frames), performance cannot be evaluated for $K > 2$ as higher eigenvalues become zero or close to zero.

| Window length (in ms) | $K = 2$ | | $K = 3$ | | $K = 4$ | |
|---|---|---|---|---|---|---|
| | EER (in %) | minDCF × 100 | EER (in %) | minDCF × 100 | EER (in %) | minDCF × 100 |
| 40 | 9.42 | 4.10 | – | – | – | – |
| 60 | 8.83 | 3.96 | **8.29** | **3.82** | 9.47 | 4.14 |
| 80 | 8.98 | **3.82** | 8.93 | 3.90 | 8.98 | 3.87 |
| 100 | 9.51 | 4.08 | 9.42 | 4.12 | 9.03 | 4.22 |
| 120 | 10.26 | 4.48 | 10.50 | 4.47 | 10.84 | 4.73 |
| 140 | 10.70 | 4.50 | 10.55 | 4.72 | 10.90 | 4.83 |
| 160 | 10.45 | 4.77 | 10.56 | 4.81 | 10.94 | 4.90 |

The proposed NSWEC feature combines information from both the eigenvalues and -vectors of the local covariance matrix. In a separate experiment, we also studied whether eigenvalues only (without eigenvectors) could also serve as useful features. To this end, we set the temporal window length again to 100 ms (9 adjacent frames), used square root of the top eigenvalues (*i.e.*, the singular values) as features, and varied dimensionality from 3 to the maximum value of 9. This leads to EERs larger than 32% in all the cases, which indicates that eigenvalues alone perform poorly.

### 5.1. Effect of temporal window length and number of eigenvector

The difference in performance of our proposed feature w.r.t. MFCC is high. The length of window (*i.e.*, 100 ms) for computing the proposed feature may not be optimal. The temporal window length should be carefully chosen: too long a window may be influenced by the context while too short a window will not effectively represent information related to temporal variation. Table 4 shows the results for NSWEC with temporal window size varried from 40 to 140 ms. We also vary the number of eigenvectors, *i.e.*, $K$. The highest recognition accuracy on NIST SRE 2001 is obtained with temporal window of 60 ms (*i.e.*, 5 frames) with $K = 3$. Interestingly, this is the same length of speech from which our baseline MFCCs are computed (by taking delta and double-deltas into account).

### 5.2. Comparison with conventional dynamic features

Keeping in mind that the proposed features are similar to deltas in the sense that they capture temporal characteristics of the base coefficients, it is interesting to compare them to deltas only (without the base MFCCs). To this end, we compare the performances of the proposed eigenvector-based feature (NSWEC) with the conventional dynamic coefficients (deltas, double-deltas, and triple-deltas) in Table 5. When used separately, the traditional deltas and double deltas achieve lower error rates in most cases. But comparing the combination of traditional dynamic coefficients with the equivalent variant of the proposed features (*i.e.*, first two eigenvectors or first three eigenvectors), we observe reductions in both EER and minDCF. We also observe that when these dynamic coefficients are further augmented with the static MFCCs, the performance improves substantially.

### 5.3. Complementarity and compatibility with other robust features

The proposed feature set conveys information associated with the variation of spectral features neglected in MFCCs. Therefore, we expect a gain in speaker verification accuracy when the two feature sets are combined. We furthermore hypothesize that the proposed local-covariance based feature can be computed from other

**Table 5**
Comparison of speaker verification performance on NIST SRE 2001 using conventional dynamic features, proposed eigenstructure-based features, and their combinations with static MFCC. The proposed features are computed for 60 ms temporal window.

| Feature (dimension) | Stand-alone | | Input fusion with static MFCC | |
|---|---|---|---|---|
| | EER (in %) | minDCF × 100 | EER (in %) | minDCF × 100 |
| $\Delta$ (19) | 10.37 | 4.41 | 8.29 | 3.71 |
| $\Delta\Delta$ (19) | 11.68 | 5.11 | 8.29 | 3.97 |
| $\Delta\Delta\Delta$ (19) | 14.37 | 6.24 | 8.63 | 4.06 |
| $\Delta + \Delta\Delta$ (38) | 9.37 | 4.07 | 8.01 | 3.66 |
| $\Delta + \Delta\Delta + \Delta\Delta\Delta$ (57) | 9.76 | 4.25 | 8.39 | 3.55 |
| First eigenvector (19) | 10.01 | 4.56 | 8.15 | 3.61 |
| Second eigenvector (19) | 14.52 | 6.37 | 9.14 | 4.18 |
| Third eigenvector (19) | 26.64 | 9.37 | 9.14 | 4.32 |
| First two eigenvectors (38) | 8.83 | 3.96 | 7.62 | 3.55 |
| First three eigenvectors (57) | 8.29 | 3.82 | 7.90 | 3.52 |
| First four eigenvectors (76) | 9.03 | 4.22 | 7.89 | 3.69 |

**Table 6**
Speaker verification performance on NIST SRE 2001 using different fusion technique. The results are also shown when the proposed technique is applied to the MHEC, FDLP and PNCC features. In all cases, proposed NSWEC features are computed with temporal window of 60 ms and three eigenvectors are retained.

| Base feature | Mode | EER (in %) | minDCF × 100 |
|---|---|---|---|
| MFCC | Baseline | 8.01 | 3.66 |
| | NSWEC | 8.29 | 3.82 |
| | Input fusion | **7.94** | **3.41** |
| | Score fusion | 7.90 | 3.50 |
| MHEC | Baseline | 11.34 | 4.84 |
| | NSWEC | 9.96 | 4.60 |
| | Input fusion | **9.67** | **4.19** |
| | Score fusion | 10.21 | 4.39 |
| FDLP | Baseline | 9.42 | **3.88** |
| | NSWEC | 13.00 | 5.59 |
| | Input fusion | **9.13** | 3.92 |
| | Score fusion | 9.57 | 3.98 |
| PNCC | Baseline | 9.23 | 3.88 |
| | NSWEC | 10.83 | 4.78 |
| | Input fusion | **8.06** | **3.58** |
| | Score fusion | 8.87 | 3.83 |

features not limited to MFCCs. In Table 6, results of the combination schemes and results with other features are shown for NIST SRE 2001 with the optimized temporal window size (60 ms) obtained above. We find that the proposed eigenstructure-based feature extraction technique works well with MHEC, FDLP and PNCC as well. Performance is also improved when fused with the conventional features for both input and output fusion schemes. Here, input fusion is done by concatenating the base cepstral and local variability based features. Alternatively, output fusion is performed by linearly combining the recognition scores (*i.e.*, likelihood ratio) of the two systems with equal weights. We also note that input fusion (*i.e.*, frame-level concatenation of two 57-dimensional feature vectors) yields lower error rates compared with score fusion.

### 5.4. Performance evaluation on i-vector framework

We evaluate the performance in i-vector framework with the optimized feature configuration obtained in the initial experiments with the GMM–UBM system. We conduct experiments with MFCC as base feature as it outperforms other features in the preliminary experiments on NIST SRE 2001 (Table 6). In addition to the input fusion and equal weights (EW) score fusion, we have also done experiments with linear regression (LR) based score fusion. Here, fusion weights are optimized in one speech corpus (development data) by minimizing logistic loss function. Then the weights are applied in experiments with evaluation data. We have used Focal toolkit[6] for this purpose. NIST SRE 2008 is used as development data for optimizing the fusion parameters. We have further conducted experiments with i-vector fusion. Here, i-vectors from two systems are first concatenated. Then they are processed with LDA, whitening followed by length normalization before they are used with PLDA system. The results for different fusion schemes along with baseline system are shown in Table 7 and Table 8 for NIST SRE 2008 and SRE 2010, respectively. The combined systems give higher recognition accuracy than the baseline MFCC-based system. Highest relative improvement has been achieved for male section of NIST SRE 2010 (C6). In this case, the relative reduction in EER is 12.28% for i-vector fusion based combined system.

**Table 7**
Speaker verification performance on telephone speech sub-condition (C6) of NIST SRE 2008 using i-Vector system for baseline MFCC and eigenstructure-based proposed features. Results are also shown for input fusion, equal weighted score fusion, and i-vector fusion.

| System | EER (in %) | | minDCF × 100 | |
|---|---|---|---|---|
| | Male | Female | Male | Female |
| Baseline | 4.81 | 6.26 | 7.38 | 9.79 |
| NSWEC | 5.63 | 7.57 | 7.36 | 9.80 |
| Input fusion | 4.92 | **6.20** | 7.60 | 9.79 |
| Score fusion (EW) | 4.67 | 6.60 | **6.92** | **9.78** |
| i-Vector fusion | **4.58** | 6.43 | 7.49 | 9.85 |

### 5.5. Text-dependent speaker recognition results on RSR2015

Experimental results on the text-dependent RSR2015 corpus are shown in Tables 9 and 10, respectively, for the development and evaluation sections. Speaker verification performance using our baseline GMM–UBM system is outperforms the previously reported results in most of the sub-conditions [54]. We see increased recognition accuracy in all cases using our proposed feature in fused mode. In many cases, the improvement is remarkably higher compared to improvements obtained on the text-independent NIST corpora. For example, we have obtained 40% relative reduction in EER over our MFCC baseline system in the female part of the evaluation section for the third sub-condition (*i.e.*, where wrong pass-phrases from impostors are used as non-target trials). Considerable improvement is also obtained in the other cases. This is most likely due to the lexical constraints in text-dependent mode. Here, the proposed features capture speaker-related variability and the contribution from this complementary information is helpful to improve speaker verification accuracy.

[6] https://sites.google.com/site/nikobrummer/focal.

**Table 8**
Same as Table 7 but for the sub-conditions C5 and C6 of NIST SRE 2010.

| System | C5 | | | | C6 | | | |
|---|---|---|---|---|---|---|---|---|
| | EER (in %) | | minDCF × 100 | | EER (in %) | | minDCF × 100 | |
| | Male | Female | Male | Female | Male | Female | Male | Female |
| Baseline | 2.55 | 3.68 | 4.73 | 5.44 | 4.48 | 5.75 | 7.19 | 8.13 |
| NSWEC | 3.72 | 5.64 | 7.72 | 6.60 | 7.78 | 8.64 | 8.36 | 9.40 |
| Input fusion | 2.64 | 3.94 | 4.51 | 5.35 | 4.49 | 7.10 | 7.81 | 7.49 |
| i-Vector fusion | 2.67 | **3.66** | 4.47 | 4.96 | **3.93** | 6.56 | 7.35 | **6.99** |
| Score fusion (EW) | 2.74 | 3.94 | 4.59 | **4.70** | 4.40 | 5.46 | **7.08** | 8.36 |
| Score fusion (LR) | **2.27** | **3.66** | **4.33** | 5.10 | 3.95 | **5.35** | 7.13 | 8.24 |

**Table 9**
Speaker verification performance on *development section* of RSR2015 using GMM–UBM system for baseline MFCC (static $+\,\Delta + \Delta\Delta$), eigenstructure-based proposed features and combined systems. The proposed features are computed by considering top three eigenvectors from temporal window of 60 ms.

| System | Target: TC, Non-target: TW | | | | Target: TC, Non-target: IC | | | | Target: TC, Non-target: IW | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | EER (in %) | | minDCF × 100 | | EER (in %) | | minDCF × 100 | | EER (in %) | | minDCF × 100 | |
| | Male | Female | Male | Female | Male | Female | Male | Female | Male | Female | Male | Female |
| Baseline | 2.86 | 0.84 | 1.40 | 0.43 | 2.83 | 1.98 | 1.35 | 1.03 | 0.36 | 0.10 | 0.16 | 0.05 |
| NSWEC | 4.38 | 2.13 | 2.05 | 1.34 | 4.28 | 3.23 | 2.08 | 1.78 | 0.76 | 0.38 | 0.35 | 0.19 |
| Input fusion | **2.33** | **0.64** | **1.11** | **0.32** | **2.64** | **1.61** | **1.22** | **0.86** | **0.31** | **0.06** | **0.13** | **0.03** |
| Score fusion (EW) | 2.73 | 0.77 | 1.30 | 0.43 | 2.70 | 1.76 | 1.28 | 0.94 | 0.32 | 0.11 | **0.13** | 0.04 |

**Table 10**
Same as Table 9 but for evaluation section of RSR2015.

| System | Target: TC, Non-target: TW | | | | Target: TC, Non-target: IC | | | | Target: TC, Non-target: IW | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | EER (in %) | | minDCF × 100 | | EER (in %) | | minDCF × 100 | | EER (in %) | | minDCF × 100 | |
| | Male | Female | Male | Female | Male | Female | Male | Female | Male | Female | Male | Female |
| Baseline | 1.31 | 0.60 | 0.63 | 0.27 | 1.67 | 1.76 | 0.86 | 0.91 | 0.17 | 0.08 | 0.07 | 0.04 |
| NSWEC | 2.64 | 1.76 | 1.30 | 0.88 | 3.09 | 3.56 | 1.59 | 1.84 | 0.39 | 0.34 | 0.19 | 0.17 |
| Input fusion | **0.97** | **0.49** | **0.47** | **0.23** | **1.58** | 1.72 | 0.80 | **0.86** | **0.12** | **0.05** | **0.05** | **0.03** |
| Score fusion (EW) | 1.25 | 0.53 | 0.59 | 0.27 | 1.62 | **1.67** | **0.79** | 0.89 | 0.14 | 0.08 | 0.06 | **0.03** |
| Score fusion (LR) | 1.23 | 0.53 | 0.59 | 0.27 | 1.62 | 1.68 | **0.79** | 0.89 | 0.14 | 0.08 | **0.05** | **0.03** |

## 6. Conclusion

Most speaker verification methods rely on speaker means, for instance, in the form of GMM supervectors or i-vectors, while the use of (co)variance features has been much less explored. To this end, the main intention of this study was to investigate feasibility of local covariance features for speaker characterization. We have proposed a new straightforward speech parameterization from short-term covariance matrix based on eigenstructure analysis. Similar to delta features, the proposed features can be computed from arbitrary base cepstral coefficients not limited to MFCCs as demonstrated in this study with the MHEC, FDLP, and PNCC features.

When used as stand-alone features, the speaker verification error rates were higher than our MFCC baseline (including deltas and double deltas), but comparable and complementary with the most standard "dynamic" features – deltas and double-deltas. Different from the delta coefficients, our features are – by construction – invariant to frame re-ordering within the observation window and they capture the uncertainty in the window. Fusion experiments were conducted out to find out the compatibility of our features with three recently investigated features. We got consistently better results for GMM–UBM based speaker recognition system for both input (feature) and output (score) fusion when evaluated on the NIST SRE 2001 data. We then performed experiments with an up-to-date i-vector system on NIST SRE 2008 and 2010 corpora. The proposed features were again found helpful when fused with MFCCs at frame or score level. In experiments with text-dependent RSR2015 corpus, we have also observed considerable reductions in EER and minDCF.

Summing up our study, the use of local covariance information for speaker characterization holds promise in speaker characterization. The local covariance features capture spreading (uncertainty) information of the local short-term features absent from the cepstral features and deltas. It is worthwhile emphasizing that we observed performance improvements (in fusion mode with baseline features) for four different state-of-the-art base feature sets, two different classifiers and four different corpora including both text-independent and text-dependent set-ups. Some of the interesting future directions are robustness analysis, variable window size, optimized back-end parameters and studying the applicability of our features in other tasks, such as language and accent recognition.

## Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions which have greatly helped in improving the content of this paper. This work was funded from Academy of Finland (proj. nos. 253120 and 283256).

## References

[1] J.R. Deller, J.H. Hansen, J.G. Proakis, Discrete-Time Processing of Speech Signals, Wiley India Pvt. Ltd., 2011.
[2] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoust. Speech Signal Process. 28 (4) (1980) 357–366.
[3] H. Hermansky, Perceptual linear predictive PLP analysis of speech, J. Acoust. Soc. Am. 87 (4) (1990) 1738–1752.
[4] T. Kinnunen, H. Li, An overview of text-independent speaker recognition: from features to supervectors, Speech Commun. 52 (1) (2010) 12–40.

[5] F. Soong, A. Rosenberg, On the use of instantaneous and transitional spectral information in speaker recognition, IEEE Trans. Acoust. Speech Signal Process. 36 (6) (1988) 871–879.
[6] H. Hermansky, Mel cepstrum, deltas, double-deltas,.. -what else is new?, in: Proceedings Robust Methods for Speech Recognition in Adverse Condition, 1999.
[7] O. Viikki, K. Laurila, Cepstral domain segmental feature vector normalization for noise robust speech recognition, Speech Commun. 25 (1–3) (1998) 133–147.
[8] J. Pelecanos, S. Sridharan, Feature warping for robust speaker verification, in: 2001: A Speaker Odyssey – The Speaker Recognition Workshop, 2001.
[9] T. Kinnunen, R. Saeidi, F. Sedlak, K.A. Lee, J. Sandberg, M. Hansson-Sandsten, H. Li, Low-variance multitaper MFCC features: a case study in robust speaker verification, IEEE Trans. Audio Speech Lang. Process. 20 (7) (2012) 1990–2001.
[10] C. Hanilci, T. Kinnunen, F. Ertas, R. Saeidi, J. Pohjalainen, P. Alku, Regularized all-pole models for speaker verification under noisy environments, IEEE Signal Process. Lett. 19 (3) (2012) 163–166.
[11] S. Chakroborty, Some studies on acoustic feature extraction, feature selection and multi-level fusion strategies for robust text-independent speaker identification, Ph.D. thesis, Indian Institute of Technology Kharagpur, 2008.
[12] L. Qi, H. Yan, An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions, IEEE Trans. Audio Speech Lang. Process. 19 (6) (2011) 1791–1801.
[13] X. Zhao, Y. Shao, D. Wang, CASA-based robust speaker identification, IEEE Trans. Audio Speech Lang. Process. 20 (5) (2012) 1608–1616.
[14] S.O. Sadjadi, J.H. Hansen, Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification, Speech Commun. 72 (2015) 138–148.
[15] M. Sahidullah, G. Saha, Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition, Speech Commun. 54 (4) (2012) 543–565.
[16] S. Ganapathy, J. Pelecanos, M. Omar, Feature normalization for speaker verification in room reverberation, in: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2011, pp. 4836–4839.
[17] M. Athineos, D. Ellis, Autoregressive modeling of temporal envelopes, IEEE Trans. Signal Process. 55 (11) (2007) 5237–5245.
[18] S. Ganapathy, Signal analysis using autoregressive models of amplitude modulation, Ph.D. thesis, Johns Hopkins University, January 2012.
[19] C. Kim, R. Stern, Power-normalized cepstral coefficients (PNCC) for robust speech recognition, in: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2012, pp. 4101–4104.
[20] M. McLaren, N. Scheffer, M. Graciarena, L. Ferrer, Y. Lei, Improving speaker identification robustness to highly channel-degraded speech through multiple system fusion, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2013, pp. 6773–6777.
[21] S.H.R. Mallidi, S. Ganapathy, H. Hermansky, Robust speaker recognition using spectro-temporal autoregressive models, in: INTERSPEECH, 2013, pp. 3689–3693.
[22] S.O. Sadjadi, T. Hasan, J. Hansen, Mean Hilbert envelope coefficients (MHEC) for robust speaker recognition, in: INTERSPEECH, 2012.
[23] J.P. Campbell, Speaker recognition: a tutorial, Proc. IEEE 85 (9) (1997) 1437–1462.
[24] R.D. Zilca, Text-independent speaker verification using covariance modeling, IEEE Signal Process. Lett. 8 (4) (2001) 97–99.
[25] D. Reynolds, R. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models, IEEE Trans. Speech Audio Process. 3 (1) (1995) 72–83.
[26] W. Campbell, D. Sturim, D. Reynolds, A. Solomonoff, SVM based speaker verification using a GMM supervector kernel and NAP variability compensation, in: 2006 IEEE International Conference on Acoustics, Speech and Signal Processing. Proceedings, Vol. 1, ICASSP 2006, 2006, pp. I–I.
[27] A.O. Hatch, S.S. Kajarekar, A. Stolcke, Within-class covariance normalization for SVM-based speaker recognition, in: Interspeech, 2006.
[28] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front–end factor analysis for speaker verification, IEEE Trans. Audio Speech Lang. Process. 19 (4) (2011) 788–798.
[29] P. Kenny, T. Stafylakis, P. Ouellet, M. Alam, P. Dumouchel, PLDA for speaker verification with utterances of arbitrary duration, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2013, pp. 7649–7653.
[30] O. Tuzel, F. Porikli, P. Meer, Region covariance: a fast descriptor for detection and classification, in: A. Leonardis, H. Bischof, A. Pinz (Eds.), Computer Vision, ECCV 2006, in: Lecture Notes in Computer Science, vol. 3952, Springer, Berlin, Heidelberg, 2006, pp. 589–600.
[31] A. Cherian, S. Sra, A. Banerjee, N. Papanikolopoulos, Jensen–Bregman logdet divergence with application to efficient similarity search for covariance matrices, IEEE Trans. Pattern Anal. Mach. Intell. 35 (9) (2013) 2161–2174.
[32] Y. Pang, Y. Yuan, X. Li, Gabor-based region covariance matrices for face recognition, IEEE Trans. Circuits Syst. Video Technol. 18 (7) (2008) 989–993.
[33] J. Stone, Blind source separation using temporal predictability, Neural Comput. 13 (7) (2001) 1559–1574.
[34] M. Xie, J. Hu, S. Guo, Segment-based anomaly detection with approximated sample covariance matrix in wireless sensor networks, IEEE Trans. Parallel Distrib. Syst. 26 (2) (2015) 574–583.
[35] K. Yang, C. Shahabi, A PCA-based similarity measure for multivariate time series, in: Proceedings of the Second ACM International Workshop on Multimedia Databases, MMDB 2004, 2004, pp. 65–74.
[36] A. Barachant, S. Bonnet, M. Congedo, C. Jutten, Classification of covariance matrices using a Riemannian-based kernel for BCI applications, Neurocomputing 112 (0) (2013) 172–178.
[37] M. Banbrook, S. McLaughlin, I. Mann, Speech characterization and synthesis by nonlinear methods, IEEE Trans. Speech Audio Process. 7 (1) (1999) 1–17.
[38] R. Duda, P. Hart, D.G. Stork, Pattern Classification, 2nd edition, John Wiley & Sons Inc., 2006, pp. 111–112, Ch. 3.
[39] R. Johnson, D. Wichern, Applied Multivariate Statistical Analysis, 6th edition, PHI Learning Pvt. Ltd., 2007.
[40] D. Reynolds, T. Quatieri, R. Dunn, Speaker verification using adapted Gaussian mixture models, Digit. Signal Process. 10 (1–3) (2000) 19–41.
[41] M. Sahidullah, Enhancement of speaker recognition performance using block level, relative and temporal information of subband energies, Ph.D. thesis, Indian Institute of Technology Kharagpur, India, 2015.
[42] N. Jayant, P. Noll, Digital Coding of Waveforms: Principles and Applications to Speech and Video, PHI, 1984.
[43] R. Horn, C. Johnson, Matrix Analysis, 2nd edition, Cambridge University Press, 2012.
[44] G. Strang, Linear Algebra and Its Applications, 4th edition, Cengage Learning, 2005.
[45] G. Golub, C. Van Loan, Matrix Computations, 3rd edition, JHU Press, 2007.
[46] G. Strang, Introduction to Applied Mathematics, Wellesley-Cambridge Press, 1986.
[47] M. Holmes, A. Gray, C. Isbell, Fast SVD for large-scale matrices, in: Workshop on Efficient Machine Learning at NIPS, 2007.
[48] L. Hogben, Handbook of Linear Algebra, CRC Press, 2006.
[49] B. Datta, Numerical Linear Algebra and Applications, SIAM, 2010.
[50] R. Badeau, G. Badeau, B. David, Sliding window adaptive SVD algorithms, IEEE Trans. Signal Process. 52 (1) (2004) 1–10.
[51] D. Jeong, C. Ziemkiewicz, W. Ribarsky, R. Chang, Understanding principal component analysis using a visual analytics tool, Charlotte Visualization Center, UNC Charlotte, 2009.
[52] G. Stewart, Perturbation theory for the singular value decomposition, in: SVD and Signal Processing II: Algorithms, Analysis and Applications, Elsevier Science, New York, 1991, pp. 99–109.
[53] J. Pohjalainen, C. Hanilci, T. Kinnunen, P. Alku, Mixture linear prediction in speaker verification under vocal effort mismatch, IEEE Signal Process. Lett. 21 (12) (2014) 1516–1520.
[54] A. Larcher, K. Lee, B. Ma, H. Li, Text-dependent speaker verification: classifiers, databases and RSR2015, Speech Commun. 60 (2014) 56–77.
[55] H. Hermansky, N. Morgan, RASTA processing of speech, IEEE Trans. Speech Audio Process. 2 (4) (1994) 578–589.
[56] M. Sahidullah, G. Saha, Comparison of speech activity detection techniques for speaker recognition, arXiv:1210.0297.
[57] S. Prince, J. Elder, Probabilistic linear discriminant analysis for inferences about identity, in: IEEE 11th International Conference on Computer Vision, 2007, ICCV 2007, 2007, pp. 1–8.
[58] D. Garcia-Romero, C. Espy-Wilson, Analysis of i-vector length normalization in speaker recognition systems, in: INTERSPEECH, Florence, Italy, 2011, pp. 249–252.

**Md Sahidullah** received the Ph.D. degree in the area of speech processing from the Department of Electronics and Electrical Communication Engineering of Indian Institute Technology Kharagpur in 2015. Prior to that he obtained the Bachelors of Engineering degree in Electronics and Communication Engineering from Vidyasagar University in 2004 and the Masters of Engineering degree in Computer Science and Engineering (with specialization in Embedded System) from West Bengal University of Technology in 2006. In 2007–2008, he was with Cognizant Technology Solutions India PVT Limited. Since 2014, he is working as a post-doctoral researcher in the School of Computing, University of Eastern Finland. His research interest includes speaker recognition, voice activity detection.

**Tomi Kinnunen** received the Ph.D. degree in computer science from the University of Eastern Finland (UEF, formerly Univ. of Joensuu) in 2005. From 2005 to 2007, he was an associate scientist at the Institute for Infocomm Research (I2R) in Singapore. Since 2007, he has been with UEF.

In 2010–2012, his research was funded by the Academy of Finland in a post-doctoral project focusing on speaker recognition. In 2014, he chaired Odyssey 2014: The Speaker and Language Recognition workshop. He also acts as an associate editor in two journals, IEEE/ACM Transactions on Audio, Speech and Language Processing and Digital Signal Processing. His primary research interests are in the broad area of speaker and language recognition where he has authored about 100 peer-reviewed scientific publications.