Reconstruction of Missing Features Based on a Low-Rank Assumption for Robust Speaker Identification

Christos Tzagkarakis and Athanasios Mouchtaris
University of Crete, Department of Computer Science, Heraklion, Greece
Institute of Computer Science, Foundation for Research and Technology - Hellas, Heraklion, Greece
{tzagarak, mouchtar}@ics.forth.gr

***Abstract*—Reconstruction of missing features promotes robustness in speaker recognition applications under noisy conditions. In this paper, we aim at enhancing the reliability of speech features for noise robust speaker identification under short training and testing sessions restrictions. Towards this direction, we apply a low-rank matrix recovery approach to reconstruct the unreliable spectrographic data due to noise corruption. This is performed by leveraging prior knowledge that the speech log-magnitude spectrotemporal representation is low-rank. Experiments on real speech data show that the proposed method improves the speaker identification accuracy especially for low signal-to-noise ratio (SNR) scenarios when compared with a sparse imputation approach.**

I. Introduction

Speaker recognition is a very challenging task especially in environments dominated by noise. This is even more difficult in the case where a limited amount of training and testing data is available in order to take correct decisions. The quality of speech features plays a key role for acquiring good recognition results. As a consequence, it is of high importance to provide a classification system with features which are as reliable as possible. However, the reliability of speech features is inversely proportional to the level of environmental noise, enhancing low recognition accuracy.

Missing data techniques (MDT) overcome this limitation by enabling the computation of reliable speech features under adverse noisy conditions. They assume that a noisy speech signal can be decomposed into speech-and noise-dominated time-frequency components. The speech-dominated components are considered reliable and can be directly exploited for further use, while the noise-dominated elements are categorized as unreliable, and labeled as missing spectrotemporal data. MDT have been extensively applied in the context of robust automatic speech recognition (ASR) as a solution to performance degradation due to noisy speech features, and they are distinguished in two main categories, namely, marginalization and imputation. In marginalization [1]–[3], speech decoding is based on the reliable components of a noisy time-frequency representation, while the unreliable components are eliminated or marginalized up to the observed values. The imputation approach [4]–[10] is associated with the estimation of the missing data, so that decoding can be performed in a conventional manner. These methods exploit various speech signals properties to estimate the missing features, from the data correlation expressed through statistical models to sparsity-based estimation where the features are sparsely represented in a given dictionary. It is of high importance to notice that the estimation of a reliability mask plays a key role during the discrimination between reliable and unreliable spectrotemporal components. The interested reader can find an overview of MDT for ASR in [11].

Recently, a lot of research has been carried out in the field of speaker recognition wherein the MDT strategy has been followed to minimize the side effects caused due to noise presence in speech signals. In specific, speaker identification is examined in [12]–[14], while in [15], [16] speaker verification is studied in the light of missing feature theory for improvement of recognition performance, while in [17] both tasks are evaluated. In all these works, the main steps include the use of a time-frequency binary mask to distinguish the reliable from the unreliable spetrographic data which in most cases is followed by a marginalization procedure to compensate for the missing spectrotemporal information.

In the current paper, we are interested in extending our previous work [18], where a novel imputation scheme based on matrix completion [19] is proposed for recovering the missing log-scale speech magnitude spectrographic data. This method exploits the low-rank behaviour of the speech spectrotemporal representation and proposed in the context of noise robust text-independent speaker identification under the assumption of short training and testing sessions restrictions examined in our previous work [20]. Here, we compare our low-rank based approach with a deterministic imputation method which is heavily based on sparsity assumptions as a consequence of verifying the missing-feature reconstruction efficiency of low-rank matrix recovery techniques. Thus, during performance evaluation we conduct a large number of simulations compared to our previous work [18] on a small-sized corpus revealing the efficiency of the proposed method compared to the sparse imputation technique which has been shown to achieve or even to exceed the state-of-the-art accuracy regarding ASR [8].

This research has been co-funded by the European Union (European Social Fund ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Thalis-DISFER

The remainder of this paper is organized as follows. Section II describes the low-rank matrix completion problem, while in Section III is analysed the proposed scheme for missing-feature recovery applied in robust speaker identification. A brief overview of the sparse imputation approach is given in Section IV. An experimental evaluation of the proposed technique is presented in Section V. Finally, Section VI summarizes the main conclusions and gives directions for future work. Regarding the notation, we use $\|\cdot\|_0$, $\|\cdot\|_1$ and $\|\cdot\|_2$ to denote the ℓ_0, ℓ_1 and Euclidean vector norms, respectively. The Frobenius matrix norm is denoted by $\|\cdot\|_F$.

II. Low-rank matrix recovery

Matrix completion (MC) enables the recovery of a low-rank or approximately low-rank matrix $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ from at least $O(nr\nu \ln^2 n)$ entries selected uniformly at random (with ν corresponding to the so-called degree of incoherence) [21], where $n = \max\{n_1, n_2\}$ and $r = \operatorname{rank}(\boldsymbol{M})$. Throughout the rest of the paper we will assume that all the scalars, vectors and matrices are real-valued. The original matrix can be recovered from the partially observed matrix by solving the following convex optimization problem

$$\begin{aligned} \min_{\boldsymbol{X}} \quad & \|\boldsymbol{X}\|_* \\ \text{s.t.} \quad & X_{ij} = M_{ij}\,, \ (i,j) \in \mathcal{I} \subset \{1,\ldots,n_1\} \times \{1,\ldots,n_2\}, \end{aligned} \tag{1}$$

where $k = |\mathcal{I}| \geq Cnr \ln^2 n$ denotes the number of observed entries (C is a positive constant), $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$ is the decision variable and the nuclear norm is defined as $\|\boldsymbol{X}\|_* = \sum_{q=1}^{\min\{n_1,n_2\}} \sigma_q$ with $\sigma_1, \ldots, \sigma_{\min\{n_1,n_2\}} \geq 0$ corresponding to the singular values of $\boldsymbol{X}$.

In the following, let the standard matrix completion *linear map* $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^k$. The constraints $X_{ij} = M_{ij}\,, \ \forall (i,j) \in \mathcal{I}$ in (1) can be represented by using the linear map $\mathcal{A}_{\mathcal{I}}$ as follows

$$\min_{\boldsymbol{X}} \ \|\boldsymbol{X}\|_* \ \text{ s.t. } \mathcal{A}_{\mathcal{I}}(\boldsymbol{X}) = \boldsymbol{b}, \tag{2}$$

where $\boldsymbol{b} \coloneqq \mathcal{A}_{\mathcal{I}}(\boldsymbol{M})$ contains the sample values extracted from $\boldsymbol{M}$. Each row of $\mathcal{A}_{\mathcal{I}}(\boldsymbol{M})$ corresponds to the sampling of a single (i,j) element of $\boldsymbol{M}$.

The equality constraint in (2) can also be written in matrix form

$$\mathcal{A}_{\mathcal{I}}(\boldsymbol{X}) \equiv \boldsymbol{A}\boldsymbol{x}, \ \boldsymbol{x} \coloneqq \operatorname{vec}(\boldsymbol{X}) \ \ \forall \boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}, \tag{3}$$

where $\boldsymbol{A} \in \mathbb{R}^{k \times n_1 n_2}$ and $\operatorname{vec}(\cdot) : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 n_2 \times 1}$ denotes the vectorization mapping; any vectorization mapping (e.g., row major order or column major order) is acceptable as long as it is fixed. In matrix completion, each row of $\boldsymbol{A}$ contains exactly 1 non-zero entry.

We also make use of the adjoint of $\mathcal{A}_{\mathcal{I}}$ which takes a vector and maps it to a sparse matrix with the nonzero entries of the sparse matrix corresponding to $\mathcal{I}$. Specifically,

$$\mathcal{A}^*_{\mathcal{I}}(\cdot) : \mathbb{R}^{k \times 1} \to \mathbb{R}^{n_1 \times n_2} \ \text{ with } \ k = |\mathcal{I}| \leq n_1 n_2,$$

and we have the property

$$\boldsymbol{h} = \mathcal{A}_{\mathcal{I}}(\mathcal{A}^*_{\mathcal{I}}(\boldsymbol{h})) \ \ \forall \boldsymbol{h} \in \mathbb{R}^{k \times 1}.$$

Singular value thresholding (SVT) [22] algorithm can be used for solving MC problems since SVT is efficient and can be successfully applied in solving large-scale matrix problems arising in speech features enhancement. Specifically, SVT minimizes the following constraint optimization problem

$$\min_{\boldsymbol{X}} \ \tau \|\boldsymbol{X}\|_* + \frac{1}{2}\|\boldsymbol{X}\|_F^2 \ \text{ s.t. } \mathcal{A}_{\mathcal{I}}(\boldsymbol{X}) = \mathcal{A}_{\mathcal{I}}(\boldsymbol{M}), \tag{4}$$

where the positive constant τ is a trade off between the nuclear and Frobenius norm. The solution to problem (4) converges to that of (1) as $\tau \to \infty$. SVT comprises the two following iterative steps

$$\begin{cases} \boldsymbol{X}_t = \mathcal{D}_\tau(\mathcal{A}^*_{\mathcal{I}}(\boldsymbol{y}_{t-1})) \\ \boldsymbol{y}_t = \boldsymbol{y}_{t-1} - \delta(\mathcal{A}_{\mathcal{I}}(\boldsymbol{X}_t) - \boldsymbol{b}). \end{cases} \tag{5}$$

In the above equation the shrinkage operator $\mathcal{D}_\tau$, also known as *soft-thresholding operator*, is denoted as $\mathcal{D}_\tau = \boldsymbol{U}\boldsymbol{\Sigma}_\tau \boldsymbol{V}^T$ where $\boldsymbol{U}$ and $\boldsymbol{V}$ are matrices with orthonormal columns and $\boldsymbol{\Sigma}_\tau = \operatorname{diag}(\max\{\sigma_i - \tau, 0\})$ with $\{\sigma_i\}_{i=1}^{\min\{n_1,n_2\}}$ corresponding to the singular values of the decomposed matrix. The step size of the iterative algorithmic process is given by δ.

III. Missing-features recovery using low-rank matrix completion

As it was mentioned in the introduction, in the current paper our goal is to enhance the reliability of speech features degraded due to environmental (ambient) noise, which are used in speaker identification by adopting the MC framework as described in the previous section. Thus, it is crucial to reduce the noise effects after the feature extraction process by following a missing-feature reconstruction approach.

In particular, the observed speech data can be represented in the time-frequency domain as $Y(f,\rho) = S(f,\rho) + N(f,\rho)$, where $\boldsymbol{Y} \in \mathbb{R}^{F \times P}$, $\boldsymbol{S} \in \mathbb{R}^{F \times P}$ and $\boldsymbol{N} \in \mathbb{R}^{F \times P}$ is the log-magnitude short-time Fourier transform (STFT) of the observed (noisy) speech signal, the clean speech signal and the contaminating noise, respectively. The discrete frequency index is denoted by f and ρ is the frame number.

The first step of spectrotemporal reconstruction is to apply a *binary reliability mask* in order to distinguish the reliable from the unreliable (or missing) spectrographic speech data. We assume that reliable time-frequency (T-F) units are dominated by speech, while unreliable T-F units contain mostly noise. The ideal (oracle) binary mask is computed as follows

$$W(f,\rho) = \begin{cases} 1 := \text{reliable}, & 10\log_{10}\left(\frac{|S(f,\rho)|}{|N(f,\rho)|}\right) > \lambda \\ 0 := \text{unreliable}, & \text{otherwise} \end{cases} \tag{6}$$

where $\boldsymbol{W} \in \mathcal{B}^{F \times P}$ with $\mathcal{B} = \{0,1\}$ and λ is a predefined threshold expressed in dB. We recover the missing spectrotemporal data $\boldsymbol{W} \odot \boldsymbol{Y}$, where $\odot$ denotes the element-wise product of the two matrices by solving the optimization problem (2) as follows

$$\hat{\boldsymbol{Y}} = \arg\min_{\boldsymbol{X}} \ \|\boldsymbol{X}\|_* \quad \text{s.t. } \mathcal{A}_{\mathcal{I}}(\boldsymbol{X}) = \mathcal{A}_{\mathcal{I}}(\boldsymbol{W} \odot \boldsymbol{Y}). \tag{7}$$

The linear map $\mathcal{A}_{\mathcal{I}}$ in (7) is related with matrix $\boldsymbol{A}$ as defined in (3), where the set of indices $\mathcal{I}$ corresponds to the non-zero entries of the binary mask $\boldsymbol{W}$

$$\mathcal{I} = \{(i,j) \mid W(i,j) \neq 0\},\ \forall (i,j) \in \{1,\ldots,F\}\times\{1,\ldots,P\}.$$

Optimization problem (7) can be rewritten as

$$\begin{aligned} \hat{\boldsymbol{Y}} = \arg\min_{\boldsymbol{X}} \quad & \tau \|\boldsymbol{X}\|_* + \frac{1}{2}\|\boldsymbol{X}\|_F^2 \\ \text{s.t.} \quad & \mathcal{A}_{\mathcal{I}}(\boldsymbol{X}) = \mathcal{A}_{\mathcal{I}}(\boldsymbol{W} \odot \boldsymbol{Y}) \end{aligned} \tag{8}$$

adopting the SVT algorithmic framework.

In order to examine the low-rankness of the original data matrix $\boldsymbol{Y}$, we use speech data obtained from the VOICES corpus, which is available from OGIs CSLU [23]. The speech database is comprised of 12 speakers (7 male and 5 female), where 50 utterances per speaker of duration around 4 sec each were recorded under quiet conditions. We take the first 3 utterances per speaker to compute the log-magnitude STFT. The ordered singular values spectra of all the speakers corresponding to an FFT size of 1024, i.e., the number of STFT matrix rows is $F = 513$, are depicted in Fig. 1. We observe that they attain very low values, where the 98% of the energy concentration is manifested around 50. Thus, we can assume that the approximate rank of the original data matrix $\boldsymbol{Y}$ is 50, and thus MC can be potentially applied to recover the missing data of the incomplete matrix $\boldsymbol{W} \odot \boldsymbol{Y}$. The estimated

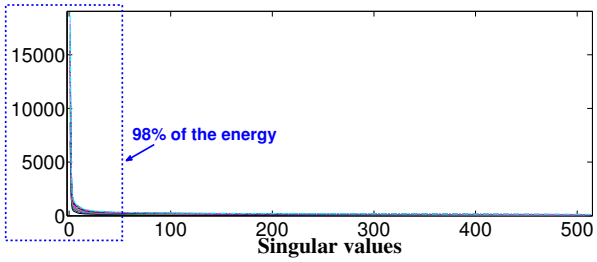

Fig. 1. Ordered singular values spectra of the log-magnitude STFT spectrograms. The concentration of 98% of the energy is around 50.

log-magnitude STFT matrix $\hat{\boldsymbol{Y}}$ is further used to compute the mel-frequency spectrographic representation, which will be termed as mel-spectrogram. This representation corresponds to a matrix whose columns consist of mel-frequency log spectral vectors, each of which represents the frequency warped log spectrum of a short speech frame

$$\boldsymbol{Q} = 10 \cdot \log_{10}\left(\boldsymbol{B} \cdot 10^{\hat{\boldsymbol{Y}}/10}\right) \in \mathbb{R}^{d\times P}, \tag{9}$$

where the matrix $\boldsymbol{B} \in \mathbb{R}^{d\times F}$ contains the mel-spaced filterbank amplitudes and d is the number of mel-filters[1]. The mel-frequency cepstral coefficients are given by

$$\mathbf{D} = \boldsymbol{\Psi}\boldsymbol{Q}, \tag{10}$$

where $\boldsymbol{\Psi}$ denotes the $d \times d$ discrete cosine transform (DCT) matrix. The features in $\mathbf{D}$ are then used for the text-independent noise robust speaker identification task.

[1]The matrix $\boldsymbol{B}$ is computed using the VOICEBOX toolbox.

IV. Missing-feature recovery based on sparse imputation

In this section, we briefly describe the sparse imputation (SI) method [8] previously applied in the context of missing data imputation for robust speech recognition. The core idea in SI is that a given signal can be represented as a sparse linear combination of basis elements.

If we combine the log-magnitude STFT of the clean speech data $\boldsymbol{S}$ with (9) and (10) the obtained mel-frequency cepstra are given by the matrix $\mathbf{D}_S \in \mathbb{R}^{d\times P}$. By following a "concatenate-then-shift" process the $d \times P$ mel-frequency cepstra matrix $\mathbf{D}_S$ is transformed into a new matrix of size $(dT) \times (\lfloor (P-T)/\xi \rfloor + 1)$, where T is the number of columns used in each iteration during the concatenation procedure and ξ is the sliding amount. Here, we assume that $\xi = 1$, i.e., we shift by one column at a time. The rescaled matrix is denoted by $\tilde{\mathbf{D}}_S$ with the i-th column being equal to $\tilde{\boldsymbol{d}}_{S,i} \in \mathbb{R}^{dT\times 1}$. Each input test sample $\tilde{\boldsymbol{d}}_{S,i}$ can be expressed as a sparse linear combination of an overcomplete matrix, the so-called dictionary, whose columns consist of a set of basis elements, usually referred to as atoms or exemplars. The linear combination is written as

$$\tilde{\boldsymbol{d}}_{S,i} = \sum_{l=1}^{\beta} \alpha_{l,i}\, \boldsymbol{g}_l = \boldsymbol{G}\boldsymbol{\alpha}_i, \tag{11}$$

where $\boldsymbol{\alpha}_i$ is an β-dimensional coefficients vector and $\boldsymbol{G}$ is an overcomplete dictionary of size $dT \times \beta$ with $\beta \gg dT$. Due to the sparsity coefficients vector's assumption, only a few exemplars are active and contribute to the representation of $\tilde{\boldsymbol{d}}_{S,i}$.

The focus is given on estimating reliable speech features further used for speaker identification under noisy conditions. We make the assumption that a set of speech data coming from the same speaker will have a similar sparse representation given the dictionary $\boldsymbol{G}$ which contains the training speech data of all speakers belonging to a database. In specific, $\boldsymbol{G}$ is formed by concatenating all the rescaled training mel-frequency cepstra matrices $\boldsymbol{G}_i$, $i = 1,\ldots,J$,

$$\begin{aligned} \boldsymbol{G} &= [\boldsymbol{g}_{1,1}|\cdots|\boldsymbol{g}_{1,m_1}|\boldsymbol{g}_{2,1}|\cdots|\boldsymbol{g}_{2,m_2}|\cdots|\boldsymbol{g}_{J,1}|\cdots|\boldsymbol{g}_{J,m_J}] \\ &= [\boldsymbol{G}_1|\boldsymbol{G}_2|\cdots|\boldsymbol{G}_J] \in \mathbb{R}^{dT\times\beta}, \end{aligned} \tag{12}$$

where J is the total number of speakers in the corpus and $\beta = m_1 + m_2 + \ldots + m_J$. If $\boldsymbol{\alpha}_i$ is a sufficiently sparse vector then the solution of the following optimization problem

$$\hat{\boldsymbol{\alpha}}_i = \arg\min_{\boldsymbol{a}} \|\boldsymbol{a}\|_1 \quad \text{s.t. } \tilde{\boldsymbol{d}}_{S,i} = \boldsymbol{G}\boldsymbol{a}. \tag{13}$$

gives a unique solution to (11). Efficient ways to solve the convex optimization problem in (13) have been studied extensively. One way is to recast (13) as an ℓ_1 norm constrained least squares problem of the form

$$\hat{\boldsymbol{\alpha}}_i = \arg\min_{\boldsymbol{a}} \left\|\boldsymbol{G}\boldsymbol{\alpha} - \tilde{\boldsymbol{d}}_{S,i}\right\|_2 + \lambda \|\boldsymbol{\alpha}\|_1, \tag{14}$$

where the least absolute shrinkage and selection operator (LASSO) algorithm [24] can be applied to compute its solution.

The mel-frequency cepstra matrix $\mathbf{D}_Y \in \mathbb{R}^{d\times P}$ corresponds to the noisy speech data $\boldsymbol{Y}$. By following the same "concatenate-then-shift" procedure as before, we obtain the rescaled versions $\tilde{\boldsymbol{W}} \in \mathbb{R}^{(dT)\times(\lfloor(P-T)/\xi\rfloor+1)}$ and $\tilde{\mathbf{D}}_Y \in \mathbb{R}^{(dT)\times(\lfloor(P-T)/\xi\rfloor+1)}$ of the mask $\boldsymbol{W}$ and noisy mel-frequency cepstra $\mathbf{D}_Y$, respectively. Then, the element-wise multiplication $\tilde{\mathbf{D}}_Y^r = \tilde{\boldsymbol{W}} \odot \tilde{\mathbf{D}}_Y$ gives a rough estimation of the reliable features. The reliable elements $\tilde{\boldsymbol{d}}_{Y,i}^r$ of the i-th column can be used to approximate the corresponding elements of $\tilde{\boldsymbol{d}}_{S,i}$ by solving the problem

$$\hat{\boldsymbol{\alpha}}_i = \arg\min_{\boldsymbol{a}} \left\| \boldsymbol{G}_r\boldsymbol{\alpha} - \tilde{\boldsymbol{d}}_{Y,i}^r \right\|_2 + \lambda \left\|\boldsymbol{\alpha}\right\|_1, \tag{15}$$

where $\boldsymbol{G}_r$ correspond to the rows of $\boldsymbol{G}$ associated with the reliable features. The obtained sparse representation $\hat{\boldsymbol{\alpha}}_i$ can be used to estimate the clean observation vector as

$$\hat{\tilde{\boldsymbol{d}}}_{S,i} = \boldsymbol{G}\hat{\boldsymbol{\alpha}}_i. \tag{16}$$

It is important to note that by solving (15) the reconstruction error will not be zero in general, thus we only impute the unreliable elements

$$\hat{\tilde{\boldsymbol{d}}}_{S,i} = \begin{cases} \hat{\tilde{\boldsymbol{d}}}_{S,i}^r = \tilde{\boldsymbol{d}}_{Y,i}^r \\ \hat{\tilde{\boldsymbol{d}}}_{S,i}^u = \boldsymbol{G}_u\hat{\boldsymbol{\alpha}}_i, \end{cases} \tag{17}$$

where $\boldsymbol{G}_u$ and $\hat{\tilde{\boldsymbol{d}}}_{S,i}^u$ corresponding to the rows of $\boldsymbol{G}$ and $\hat{\tilde{\boldsymbol{d}}}_{S,i}$ for which the i-th column $\tilde{\boldsymbol{w}}_i$ of $\tilde{\boldsymbol{W}}$ equals zero.

If we apply (15)-(17) for all columns of the features matrix $\tilde{\mathbf{D}}_Y^r$ we end up with a set of $(dT)\times(\lfloor(P-T)/\xi\rfloor+1)$ solutions of the form $\{\hat{\tilde{\boldsymbol{d}}}_{S,i}\}_i$. In matrix form notation the set $\{\hat{\tilde{\boldsymbol{d}}}_{S,i}\}_i$ can be denoted by $\hat{\tilde{\mathbf{D}}}_S$ which reflects a reliable estimation of the noisy speech features. A reshaped $d\times P$ version of $\hat{\tilde{\mathbf{D}}}_S$ can be considered denoised version of the mel-frequency cepstra matrix $\hat{\mathbf{D}}_S$ of the underlying speech signal, which can be used directly for speaker identification.

V. Experimental results

In this section, we show that the proposed low-rank matrix completion approach is an efficient method to reconstruct the missing T-F components of speech signals used during speaker identification. First, the reconstruction performance of the SVT algorithm is evaluated and compared with other matrix completion methods. Then, we demonstrate the superior reconstruction performance of the SVT algorithm against the SI method, in terms of achieving an increased correct identification accuracy over the VOICES corpus.

A. Evaluation of SVT matrix completion on missing data imputation for speaker identification

In this section, we compare the reconstruction performance of the SVT [22] algorithm with the performance obtained by reconstructing the missing data matrix using LMaFit [25] and ScGrassMC [26]. The experimental set-up, also used in our previous work [18], is adopted for the SVT performance assessment. More specifically, we are interested in achieving noise robust speaker identification, where noisy speech features are processed under a missing data imputation framework [8] towards reducing the effects of noise in order to enhance the speaker identification accuracy. In the subsequent experimental evaluations we use UBM-GMM[2] [27] as the main classification process after feature enhancement through missing data imputation.

The original speech signals are sampled at 22 kHz, and downsampled to 16 kHz. During feature extraction, an analysis window of 40 msec (equivalent to 640 samples), with a step size of 20 msec (corresponding to 320 samples), is employed to compute a mel-frequency spectrogram of 30 bands. For the UBM-GMM classifier a diagonal covariance matrix of 16 Gaussian mixtures was chosen during the simulations, where 10 sec of clean speech training data (per speaker) were used. We selected the last five utterances as testing data per speaker. Speech babble noise and factory floor noise were used to additively corrupt the test utterances. The SNR of the distorted speech is set to -15, -10, -5, 0, 5, and 10 dB, while the noise signals belong to the NOISEX-92 database [28]. For each combination of noise type and SNR level, the sampling ratio of the observed matrix $\boldsymbol{W} \odot \boldsymbol{Y}$ is defined as

$$\text{Sampling ratio} = \frac{\text{number of observed values } (k)}{\text{matrix size } (F \times P)}. \tag{18}$$

We note that the sampling ratio (18) is inversely proportional to the number of zeros in the binary mask $\boldsymbol{W}$ as defined in (6), i.e., for smaller SNR values the amount of unreliable features increases, and thus the number of observed values k corresponding to the reliable features decreases.

The performance evaluation follows the strategy described in [20]. In particular, having solved (7) each completed matrix $\hat{\boldsymbol{Y}}$ corresponds to a sequence of feature vectors (columns) $\{\hat{\boldsymbol{y}}_t \in \mathbb{R}^{F\times 1}\}_{t=1}^P$ of the form

$$\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{y}}_2, \hat{\boldsymbol{y}}_3, \ldots, \hat{\boldsymbol{y}}_{P-1}, \hat{\boldsymbol{y}}_P.$$

Each sequence of that form is divided into overlapping segments of Q feature vectors, where the segments have the following form

$$\begin{gathered} \underbrace{\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{y}}_2, \hat{\boldsymbol{y}}_3, \ldots, \hat{\boldsymbol{y}}_Q}_{1^{\text{st}}\text{ segment}} \hat{\boldsymbol{y}}_{Q+1}, \ldots, \hat{\boldsymbol{y}}_{P-1}, \hat{\boldsymbol{y}}_P \\ \hat{\boldsymbol{y}}_1, \underbrace{\hat{\boldsymbol{y}}_2, \hat{\boldsymbol{y}}_3, \ldots, \hat{\boldsymbol{y}}_Q, \hat{\boldsymbol{y}}_{Q+1}}_{2^{\text{nd}}\text{ segment}}, \ldots, \hat{\boldsymbol{y}}_{P-1}, \hat{\boldsymbol{y}}_P \\ \vdots \\ \hat{\boldsymbol{y}}_1, \hat{\boldsymbol{y}}_2, \hat{\boldsymbol{y}}_3, \ldots, \hat{\boldsymbol{y}}_Q, \hat{\boldsymbol{y}}_{Q+1}, \ldots, \hat{\boldsymbol{y}}_{P-Q}, \underbrace{\hat{\boldsymbol{y}}_{P-Q+1}, \ldots, \hat{\boldsymbol{y}}_{P-1}, \hat{\boldsymbol{y}}_P}_{\text{P}-\text{Q}+1^{\text{th}}\text{ segment}} \end{gathered} \tag{19}$$

The segment length Q is set to 400 during the testing simulations, which corresponds to approximately 8 sec. The correct identification rate (CIR) of the j-th speaker is computed as

[2]Universal Background Model for Gaussian Mixture Model

the percentage of the correctly identified segments of length Q over the total number of segments

$$\text{CIR}_j = \frac{\#\text{ cor. identified segments}}{\text{total }\#\text{ of segments}} \cdot 100\%, \qquad (20)$$

where the total number of segments equals $P - Q + 1$. The

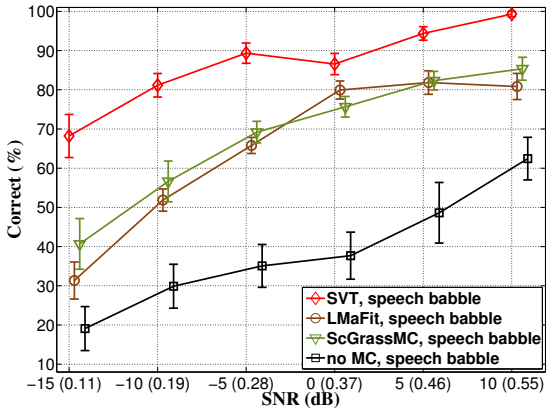

Fig. 2. Mean correct identification rates (%) for the SVT, LMaFit, ScGrassMC and no MC for six different number of SNR values, where speech babble noise is added. The numbers inside the parentheses represent the sampling ratios (18).

total mean correct identification rate is used as an evaluation metric during the test simulations, which is given by

$$\text{mean CIR} = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{1}{J}\sum_{j=1}^{J}\text{CIR}_j^r\right), \qquad (21)$$

where R and J denote the total number of Monte Carlo runs and speakers, respectively. The correct identification rate CIR_j^r of speaker j during the r-th Monte Carlo run is given by (20).

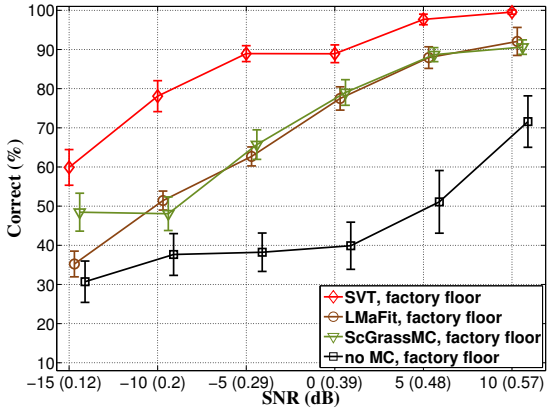

Fig. 3. Mean correct identification rates (%) for the SVT, LMaFit, ScGrassMC and no MC for six different number of SNR values, where factory floor noise is added. The numbers inside the parentheses represent the sampling ratios (18).

The average correct identification rates, computed as the percentage of the correctly identified segments over the total number of test segments, for 10 Monte Carlo runs are depicted in Figures 2 and 3. The SVT algorithm is compared with LMaFit and ScGrassMC, as well as with the no matrix completion (no MC) technique where the missing data matrix $\boldsymbol{W} \odot \boldsymbol{Y}$ is used explicitly for the speaker identification task. Fig. 2 shows the results corresponding to the speech babble noise, while Fig. 3 corresponds to the correct identification rates in the case of factory floor noise. The vertical bars indicate the 95% confidence intervals. It is clear that the SVT matrix completion algorithm outperforms substantially the other three evaluated methods across all the SNR noise levels. In particular, we can see that in both noise cases at -10 dB SNR, i.e., when approximately 80% of the data is missing, the speaker identification accuracy is around 80%. For all other cases, where the SNR is at least -5 dB the achieved correct identification rates are above 87%.

B. Evaluation of SVT against sparse imputation

In this section, we examine the reconstruction performance of the proposed low-rank matrix completion method as described in Sections II and III, with respect to the resulting correct identification rates compared with the SI approach overviewed in Section IV. Fig. 4 and Fig. 5 show the identification accuracy corresponding to speech babble and factory floor noise, respectively. In this simulation, we consider six different SNR values (-16, -12, -8, -4, 0, 4, 8 and 12 dB). Specifically, we focus on examining the reconstruction performance of SVT matrix completion compared with SI mainly in noisy conditions, i.e. for values of SNR below -4 dB.

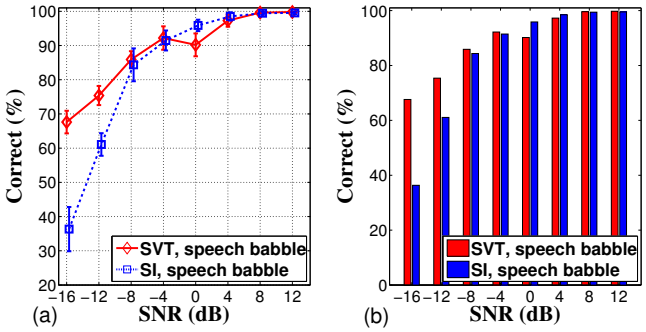

Fig. 4. Mean correct identification rates (%) for the SVT vs. SI for eight different number of SNR values, where speech babble noise is added.

In Fig. 4.(a) and Fig. 5.(a) the solid line corresponds to the identification rates achieved by the proposed SVT matrix completion approach, while the dotted line represents the performance of the sparse imputation method. In all cases, the vertical bars indicate the 95% confidence intervals. The difference in performance between the two methods especially

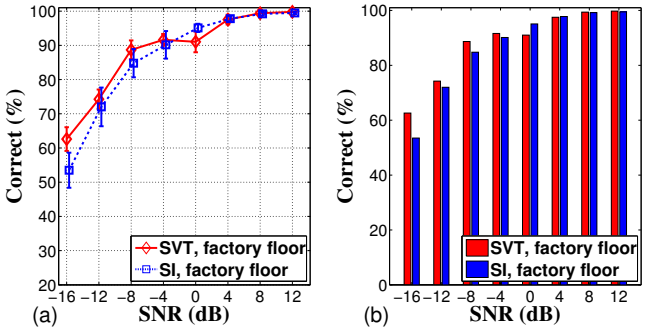

Fig. 5. Mean correct identification rates (%) for the SVT vs. SI for eight different number of SNR values, where factory floor noise is added.

in low SNR values appear more clearly in the bar plots as depicted in Fig. 4.(b) and Fig. 5.(b). It is important to address that low-rank matrix recovery performs better than SI for SNR values below -4 dB for both noise types, especially in the case of speech babble noise where SVT achieves 30% and 15% higher identification rates than SI for -16 dB and -12 dB, respectively. Similarly, SVT achieves an increase of 10% in the identification accuracy when compared with SI, for the factory floor noise at -16 dB. Clearly, for all the SNR values greater than -4 dB, SVT is slightly better than SI except for the case of 0 dB and 4 dB wherein SI slightly outperforms SVT.

As an overall conclusion, our experimental evaluation revealed that low-rank matrix recovery can compete other state-of-the-art missing data imputation methods like SI even without exploiting the a priori knowledge of training data as extra information which could enhance the identification performance.

VI. Conclusion

In this paper, we proposed a method for missing-feature reconstruction applied in the context of robust speaker identification using short training and testing data. The low-rank behaviour of the log-magnitude spectrotemporal speech data is exploited in the framework of data imputation. We compared its performance with the recently introduced sparse imputation technique showing that the proposed method achieves an improved performance in terms of higher correct identification rates. As a future work, we are interested in extending the SVT matrix completion approach to a dictionary-based version which will take into consideration the training data of all speakers. A further experimental investigation could be also conducted by applying simulations in data with a wider range of noise types and by comparing with other statistical-based imputation methods using estimated reliability masks.

References

[1] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, 2001.

[2] J. P. Barker, M. P. Cooke, and D. P. W. Ellis, "Decoding speech in the presence of other sources," *Speech Communication*, vol. 45, no. 1, pp. 5–25, 2005.

[3] N. Ma, J. Barker, H. Christensen, and P. Green, "Combining speech fragment decoding and adaptive noise floor modeling," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 20(3), pp. 818–827, March 2012.

[4] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 275–296, 2004.

[5] B. J. Borgström and A. Alwan, "Utilizing compressibility in reconstructing spectrographic data, with applications to noise robust ASR," *IEEE Signal Proc. Letters*, vol. 16, no. 5, pp. 398–401, May 2009.

[6] F. Faubel, J. McDonough, and D. Klakow, "Bounded conditional mean imputation with Gaussian mixture models: A reconstruction approach to partly occluded features," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, April 2009, pp. 3869–3872.

[7] W. .Kim and J. H. L. Hansen, "Missing-feature reconstruction by leveraging temporal spectral correlation for robust speech recognition in background noise conditions," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18(8), pp. 2111–2120, November 2010.

[8] J. F. Gemmeke, H. V. Hamme, B. Cranen, and L. Boves, "Compressive sensing for missing data imputation in noise robust speech recognition," *IEEE Journal of Selected Topics in Sig. Proc.*, vol. 4(2), pp. 272–287, April 2010.

[9] U. Remes, K. J. Palomäki, T. Raiko, A. Honkela, and M. Kurimo, "Missing-feature reconstruction with a bounded nonlinear state-space model," *IEEE Signal Proc. Letters*, vol. 18, no. 10, pp. 563–566, October 2011.

[10] J. A. González, A. M. Peinado, N. Ma, A. M. Gómez, and J. Barker, "MMSE-based missing-feature reconstruction with temporal modeling for robust speech recognition," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 21(3), pp. 624–635, March 2013.

[11] B. Raj and R. M.Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, September 2005.

[12] Y. Shao, S. Srinivasan, and D. Wang, "Incorporating auditory feature uncertainties in robust speaker identification," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, Hawaii, USA, April 2007, pp. 277–280.

[13] D. Pullella, M. Kühne, and R. Togneri, "Robust speaker identification using combined feature selection and missing data recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, March 2008, pp. 4833–4836.

[14] T. May, S. van de Par, and A. Kohlrausch, "Noise-robust speaker recognition combining missing data techniques and universal background modeling," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 20(1), pp. 108–121, January 2012.

[15] M. T. Padilla, T. F. Quatieri, and D. A. Reynolds, "Missing feature theory with soft spectral subtraction for speaker verification," in *Proc. Int. Conf. on Spoken Language Processing (Interspeech)*, Pittsburgh, PA, USA, September 2006, pp. 913–916.

[16] D. Ribas, J. A. Villalba, E. Lleida, and J. R. Calvo, "Speaker verification in noisy environment using missing feature approach," in *CIARP*, 2010, pp. 220–227.

[17] Y. Shao and D. Wang, "Robust speaker recognition using binary time-frequency masks," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, May 2006.

[18] C. Tzagkarakis and A. Mouchtaris, "Robust speaker identification using matrix completion under a missing data imputation framework," in *Proc. Workshop on Sig. Proc. with Adaptive Sparse Structured Representations (SPARS '13)*, Lausanne, Switzerland, July 2013.

[19] E. J. Candés and B. Recht, "Exact matrix completion via convex optimization," *Journal on Foundations of Computational Mathematics*, vol. 9(6), pp. 717–772, December 2009.

[20] C. Tzagkarakis and A. Mouchtaris, "Robust text-independent speaker identification using short test and training sessions," in *Proc. European Signal Processing Conf. (EUSIPCO)*, August 2010, pp. 586–590.

[21] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Trans. on Information Theory*, vol. 57, no. 3, pp. 1548–1566, March 2011.

[22] J. F. Cai, E. J. Candés, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20(4), pp. 1956–1982, March 2010.

[23] A. Kain, "High resolution voice transformation," Ph.D. dissertation, OGI School of Science and Engineering at Oregon Health and Science University, October 2001.

[24] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58(1), pp. 267–288, 1996.

[25] Z. Wen, W. Yin, and Y. Zhang, "Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm," *Mathematical Programming Computation*, vol. 4, no. 4, pp. 333–361, December 2012.

[26] T. Ngo and Y. Saad, "Scaled gradients on Grassmann manifolds for matrix completion," in *Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, Nevada, December 2012, pp. 1421–1429.

[27] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[28] A. Varga and H. J. M. Steeneken, "Assesment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, pp. 247–251, 1993.