Robust speaker identification using matrix completion under a missing data imputation framework

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

***Abstract*—In this work, we examine the problem of noise robust speaker identification under short training sessions restrictions. In order to improve the identification performance, the effects of noise prior to identification must be alleviated. Towards this direction, we exploit matrix completion to recover the unreliable spectrographic data due to noise corruption. This is done by taking advantage of the low rank property of the speech magnitude STFT spectrogram.**

I. Introduction

Text-independent speaker identification under the assumption of short training sessions is very important in applications where it is often not feasible to have large amounts of training data from all the speakers and speech data are obtained in adverse noisy environments. This problem was studied in our previous work [1], where interesting results were presented under very short training and testing conditions. Here, we adopt a matrix completion approach [2], which is applied on the log-scale speech magnitude STFT spectrogram for recovering the missing (or unreliable) spectrographic data. The interesting point of the current study is that the proposed method does not rely on training data, in contrast to previous sparsity-based recovery methods with missing spectrographic data [3].

II. Low-rank matrix completion

Matrix completion (MC) [2] enables the recovery of a low rank or approximately low rank matrix M of size $p \times q$ from $O(nr \log n)$ entries selected uniformly at random, where $n = \max\{p, q\}$ and $\text{rank}(M) = r$. Suppose that the known entries of matrix M are denoted as $M_{i,j}$, where $(i,j) \in \Omega \subset \{1,\ldots,p\} \times \{1,\ldots,q\}$. We use the sampling operator $\mathcal{A}_\Omega : \mathbb{R}^{p\times q} \to \mathbb{R}^{p\times q}$ defined as

$$(\mathcal{A}_\Omega(X))_{i,j} = \begin{cases} X_{i,j}, & (i,j) \in \Omega \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

to determine the subset selection of the complete set of entries of the matrix X. The original matrix can be recovered from the partially observed matrix by solving the following convex optimization problem

$$\min_X \|X\|_*, \text{ s.t. } \mathcal{A}_\Omega(X) = \mathcal{A}_\Omega(M), \quad (2)$$

where the nuclear norm is defined as $||X||_* = \sum_{k=1}^{\min\{p,q\}} \sigma_k$ and $\sigma_1, \ldots, \sigma_{\min\{p,q\}} \geq 0$ are the singular values of matrix X. Numerous algorithms have been recently proposed to deal with the solution of the nuclear minimization problem in (2). Here, we use the Singular Value Thresholding (SVT) algorithm [4], which at each step performs a soft-thresholding operation on the singular values of the estimated matrix.

III. Proposed Method

In practical applications, speech signals are often degraded due to environmental noise which weakens the speaker identification performance. Thus, it is crucial to reduce the noise effects during the feature extraction process. In the current work, we apply an oracle mask prior to spectral reconstruction in order to distinguish the reliable from the unreliable (or missing) spectrographic speech data. In particular, the observed speech data can be considered in the spectral domain as $Y(k,t) = S(k,t) + N(k,t)$, where

This work was funded by the IAPP CS-ORION (PIAP-GA-2009-251605) grant within 7th Framework Program of the European Community.

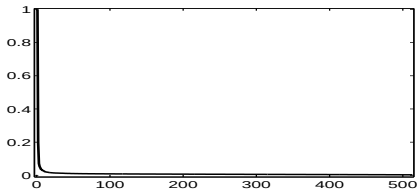

Fig. 1. Ordered normalized singular values of the log-magnitude STFT spectrogram.

$Y(k,t) \in \mathbb{C}^{K\times T}$ is the spectrographic representation of the observed (noisy) speech signal, $S(k,t) \in \mathbb{C}^{K\times T}$ corresponds to the clean speech signal and $N(k,t) \in \mathbb{C}^{K\times T}$ is the contaminating noise, respectively. Frequency band is denoted by index k and t is the frame number. The oracle mask is computed as follows

$$\mathbb{R}^{K\times T} \ni W(k,t) = \begin{cases} 1 := \text{reliable}, & \frac{|S(k,t)|^2}{|N(k,t)|^2} > \lambda \\ 0 := \text{unreliable}, & \text{otherwise} \end{cases} \quad (3)$$

where the SNR-threshold λ is set to -3 dB. Next, we recover the missing spectrographic data by solving optimization problem (2) as follows

$$\hat{X} = \min_X \|X\|_*, \text{ s.t. } \mathcal{A}_\Omega(X) = \mathcal{A}_\Omega(10\log_{10}(|W \odot Y|)), \quad (4)$$

where $\odot$ denotes the element-wise product of the two matrices. In Fig. 1, we observe that the majority of the (normalized) singular values of the log-magnitude STFT spectrogram are approximately close to zero. As a result, MC can be successfully applied to recover the missing data of the incomplete matrix $10\log_{10}(|W \odot Y|)$. The estimated matrix $\hat{X}$ is further used to compute the MFCC features for speaker identification on a voices corpus of 12 speakers, where UBM-GMM model of order 16 is applied on 10 sec clean speech training data (per speaker) and the MFCC order is 30. The proposed MC-based method is compared with a baseline approach based on the same UBM-GMM model, but using a standard MFCC front-end.

TABLE I
Mean correct identification rates (%) using UBM-GMM comparing MC and baseline system after 10 Monte Carlo runs.

noise	SNR (dB)	speech babble	car engine	factory floor	train passing
baseline	5	76.15	94.01	96.71	93.80
	0	54.72	92.51	85.28	58.89
	-5	27.29	90.21	46.50	36.13
MC	5	98.45	99.58	95.98	97.33
	0	90.68	99.57	89.20	88.02
	-5	78.80	98.21	73.71	74.16

References

[1] C. Tzagkarakis and A. Mouchtaris, "Robust text-independent speaker identification using short test and training sessions," *in Proc. of European Signal Processing Conference (EUSIPCO '10)*, Aalborg, Denmark, August 2010.

[2] E. Candés and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, pp. 717-772, 2009.

[3] J. F. Gemmeke, H. Van Hamme, B. Cranen, and L. Boves, "Compressive sensing for missing data imputation in noise robust speech recognition," *IEEE Selected Topics in Sig. Proc.*, vol. 4, no. 2, pp. 272-287, April 2010.

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