ROBUST SPEAKER IDENTIFICATION USING MATRIX COMPLETION UNDER A MISSING DATA IMPUTATION FRAMEWORK

Computer Science Department University of Crete

Christos Tzagkarakis and Athanasios Mouchtaris

Department of Computer Science, University of Crete and Institute of Computer Science, Foundation for Research and Technology - Hellas

email: {tzagarak,mouchtar}@ics.forth.gr

FO.R.T.H. Institute of Computer Science

Overview

- Design and development of efficient speaker recognition algorithms characterized by increased robustness in diverse environments and conditions
- Focus is given on *noise robust text-independent speaker identification* (SID) using *short training/testing* sessions
  - speech signals are often corrupted by noise due to, *e.g.*, the environment where the speaker is present, the voice transmission medium etc.
  - it is often not feasible to obtain large amounts of training data from all speakers
  - speaker utterance to be identified should be as short as possible to reduce the SID response time
- Reduce noise corruption adopting *matrix completion* under a missing data imputation framework
  - exploit low-rank features' property

Low-rank matrix completion

- Let $\mathrm{M} \in \mathbb{R}^{p\times q}$ be the *data matrix*
  - in general, one cannot recover the pq entries of M from a smaller number of k entries, where $k \ll pq$
  - such recovery is possible when rank $r = \mathrm{rank}(\mathrm{M})$ of the matrix is small compared to its dimensions
    - possible recovery of M from $k \geq O(nr\log(n))$, where $n = \max\{p,q\}$
  - consider the following assumptions:
    - the known entries of matrix M are denoted as $\mathrm{M}_{i,j}$, where $(i,j) \in \Omega \subset \{1,\dots,p\} \times \{1,\dots,q\}$ and Ω is the set of sampled values
    - linear map $\mathcal{A} : \mathbb{R}^{pq} \to \mathbb{R}^k$ is defined as a sampling operator recording a small number of entries from the matrix

$$(\mathcal{A}_\Omega(\mathrm{M}))_{i,j} = \begin{cases} M_{i,j}, & (i,j) \in \Omega \\ 0, & \text{otherwise} \end{cases}$$

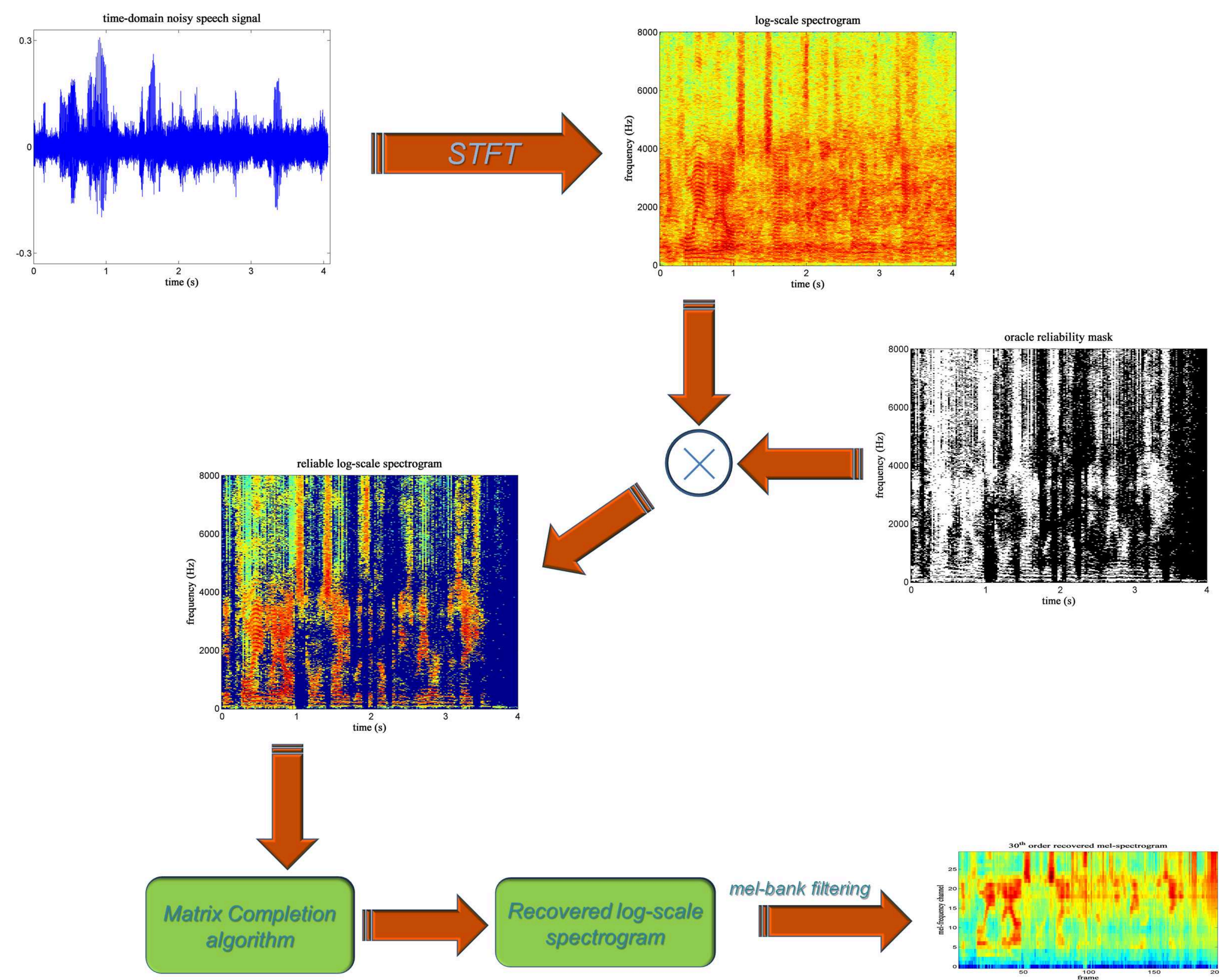
- recover incomplete matrix by computing the tractable *nuclear norm minimization problem*

$$\min_{\mathrm{X}} \|\mathrm{X}\|_*\,, \ \text{s.t.} \quad \| \mathcal{A}_\Omega(\mathrm{X}) - \mathcal{A}_\Omega(\mathrm{M}) \|_F^2 \leq \epsilon$$

where $\sigma_1,\dots,\sigma_{\min\{p,q\}} \geq 0$ are the singular values of X, the nuclear norm is defined as $\|\mathrm{X}\|_* = \sum_{k=1}^{\min\{p,q\}} \sigma_k$, and the Frobenius norm of X is $\|\mathrm{X}\|_F^2 = \sum_{k=1}^{\min\{p,q\}} \sigma_k^2$

Missing data imputation using matrix completion

- Reduce noise effects after feature extraction
- Distinguish reliable from unreliable (missing) spectrographic speech data by applying an oracle *reliability mask*
- Observed spectrographic speech data $\mathrm{M}(l,t) = \mathrm{S}(l,t) + \mathrm{N}(l,t)$, where $l = 1,\dots,L$ and $t = 1,\dots,T$
  - L: FFT size
  - T: number of frames
    - $\mathrm{M}(l,t) \in \mathbb{C}^{L\times T}$: short-time Fourier transform (STFT) of the *noisy* speech signal
    - $\mathrm{S}(l,t) \in \mathbb{C}^{L\times T}$: STFT of the *clean* speech signal
    - $\mathrm{N}(l,t) \in \mathbb{C}^{L\times T}$: STFT of the *noise* signal



- The oracle binary mask is computed as

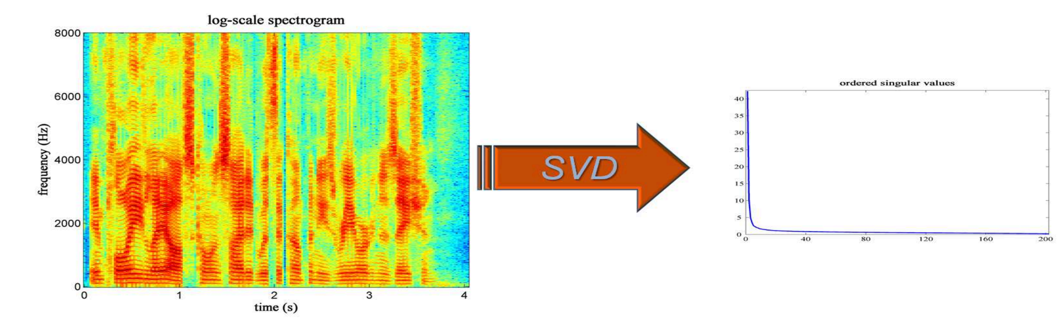
$$\mathbb{R}^{L\times T} \ni \mathrm{W}(l,t) = \begin{cases} 1 := \text{reliable}, & \frac{|\mathrm{S}(l,t)|^2}{|\mathrm{N}(l,t)|^2} > \lambda \\ 0 := \text{unreliable, otherwise} \end{cases}$$

  - λ: SNR-threshold (during the simulations $\lambda = -3$ dB)
- Recover the missing spectrographic data by solving the optimization problem

$$\hat{\mathrm{X}} = \min_{\mathrm{X}} \|\mathrm{X}\|_*\,, \ \text{s.t.} \quad \| \mathcal{A}_\Omega(\mathrm{X}) - \mathcal{A}_\Omega(10\log_{10}(|\mathrm{W} \odot \mathrm{M}|)) \|_F^2 \leq \epsilon$$

  - $\odot$: element-wise product of two matrices

- The *ordered singular values* of the clean spectrographic speech data are depicted in the figure below



  - *low-rankness assumption* is satisfied $\Longrightarrow$ the majority of the singular values tend to approach zero
- Singular Value Thresholding (SVT) algorithm used to recover the low-rank log-scale spectrographic speech data matrix

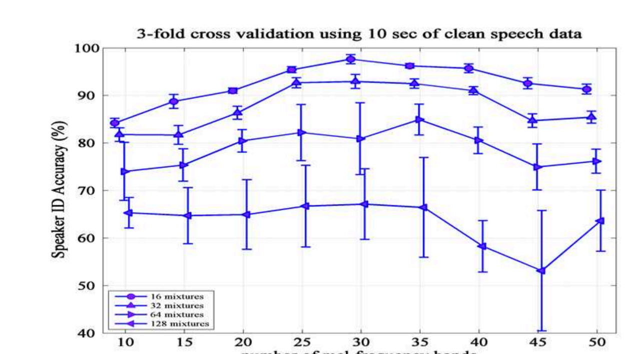
Experimental results

- VOICES corpus: 12 speakers
  - sampling rate @ 22 kHz $\longrightarrow$ downsampling @ 16 kHz
  - analysis window: 640 samples with 50% overlapping
  - number of mel-frequency bands: 30
  - training data per speaker: 10 sec
    - Universal Background Model (UBM)-GMM training model: 16 Gaussian mixtures were used using diagonal covariance matrix
  - test segment length: 400 frames (~ 8 sec)

$$\underbrace{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_Q}_{1^{\text{st}}\ \text{segment}}, \mathbf{x}_{Q+1}, \mathbf{x}_{Q+2}, \dots$$

$$\mathbf{x}_1, \underbrace{\mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_Q, \mathbf{x}_{Q+1}}_{2^{\text{nd}}\ \text{segment}}, \mathbf{x}_{Q+2}, \dots$$

$$\boxed{\text{correct ident. rate} = \frac{\#\ \text{correctly identified segments}}{\text{total}\ \#\ \text{of segments}} \cdot 100\%}$$



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

Future work

- Training-data based matrix completion
- Comparisons with other missing data techniques
- Use practical binary masks during the simulations
- Extend the experiments to moderate-size speech databases