University of Crete
Department of Computer Science

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

Sparse and Low-Rank Techniques for Robust Speaker Recognition and Missing-Feature Reconstruction

Christos Tzagkarakis

A thesis submitted for the degree of
Doctor of Philosophy

Heraklion, July 2014

University of Crete
Department of Computer Science

Sparse and Low-Rank Techniques for Robust Speaker Recognition and Missing-Feature Reconstruction

Submitted to the Department of Computer Science
in partial fulfillment of the requirements for the degree of Doctor of Philosophy

July 1, 2014

Graduate Committee Approval

Author:

Christos Tzagkarakis, Dept. of Computer Science

Date

Athanasios Mouchtaris (Assistant Professor–Advisor)
Dept. of Computer Sc., Univ. of Crete

Date

Yannis Stylianou (Professor), Dept. of Computer Sc., Univ. of Crete

Date

Panagiotis Tsakalides (Professor), Dept. of Computer Sc., Univ. of Crete

Date

Apostolos Traganitis (Professor), Dept. of Computer Sc., Univ. of Crete

Date

Georgios Tsihrintzis (Professor), Dept. of Informatics, Univ. of Piraeus

Date

Aggelos Pikrakis (Lecturer), Dept. of Informatics, Univ. of Piraeus

Date

Panayiotis Georgiou (Assistant Professor), Dept. of E. E., Univ. of Southern California

Approved by:

Panos Trahanias
Chairman of the Department

…to all those who wisely and patiently inspired me so far

Ευχαριστίες (acknowledgments)

Μια διδακτορική διατριβή κατά πολλούς αποτελεί έναν προορισμό όπου σημασία έχει το ταξίδι προς αυτόν. Έχοντας, λοιπόν, αυτό το ταξίδι φτάσει στο τέλος του έπειτα από πολλές μικρές και μεγάλες περιπετειώδεις στιγμές, θα ήθελα παρακάτω να ευχαριστήσω τα άτομα που συνέβαλαν ώστε να γίνει όσο πιο δυνατόν συναρπαστικό. Πρωτίστως τον επόπτη της παρούσας διατριβής και επίκουρο καθηγητή του τμήματος Επιστήμης Υπολογιστών κ.Αθανάσιο Μουχτάρη, Θανάση κατ΄ εμέ αφού η οικειότητα εκατέρωθεν αποτέλεσε όλα αυτά τα χρόνια ένα σημαντικό συστατικό στοιχείο της συνεργασίας μας.

Θα ήθελα, επίσης, να ευχαριστήσω τον καθηγητή του τμήματος Επιστήμης Υπολογιστών κ.Παναγιώτη Τσακαλίδη, ο οποίος μου έδωσε τη δυνατότητα στο τελευταίο έτος των προπτυχιακών σπουδών μου, να εργαστώ ερευνητικά στο Ίδρυμα Τεχνολογίας και Έρευνας, δίνοντάς μου την ευκαιρία να γνωρίσω από πρώτο χέρι τη διαδικασία έρευνας σε ένα υψηλού επιπέδου επιστημονικό περιβάλλον, ανοίγοντάς μου με αυτόν τον τρόπο την πόρτα για την μελλοντική μου παρουσία στο Ινστιτούτο Πληροφορικής του Ιδρύματος Τεχνολογίας και Έρευνας. Δε θα μπορούσα να μην αναφερθώ στον επίσης καθηγητή του τμήματος Επίστημης Υπολογιστών κ.Ιωάννη Στυλιανού, του οποίου το παροιμιώδες πάθος για τη διδασκαλία του μαθήματος *Εφαρμοσμένα Μαθηματικά για Μηχανικούς* αποτέλεσε το εφαλτήριο για να προσεγγίσω με περίσσιο ενδιαφέρον την ευρύτερη περιοχή της επεξεργασίας σήματος τα μετέπειτα χρόνια, αποτελώντας πλέον την κύρια ερευνητική ενασχόλησή μου.

Τα υπόλοιπα μέλη της επταμελούς εξεταστικής επιτροπής, διετέλεσαν σημαντικό ρόλο στην ποιοτική αξιολόγηση της παρούσας διατριβής. Συγκεκριμένα, θα ήθελα να ευχαριστήσω τους καθηγητές κ.Απόστολο Τραγανίτη (τμήμα Επιστήμης Υπολογιστών - Πανεπιστήμιο Κρήτης) και κ.Γεώργιο Τσιχριντζή (τμήμα Πληροφορικής - Πανεπιστήμιο Πειραιά), τον επίκουρο καθηγητή κ.Παναγιώτη Γεωργίου (τμήμα Ηλεκτρολόγων Μηχανικών - Πανεπιστήμιο της Νότιας Καλιφόρνιας) και τον λέκτορα κ.Άγγελο Πικράκη (τμήμα Πληροφορικής - Πανεπιστήμιο Πειραιά).

Ένα μεγάλο ευχαριστώ αξίζει σε όλους τους φίλους και συνεργάτες των εργαστήριων Επεξεργασίας Σήματος και Τηλ/νιών & Δικτύων του Ινστιτούτου Πληροφορικής του Ι.Τ.Ε., καθώς και αυτών του εργαστηρίου Πολυμέσων και Πληροφορικής του τμήματος Επιστήμης Υπολογιστών.

Ξεχωριστές ευχαριστίες αξίζουν στους γονείς μου και στον αδελφό μου, των οποίων η υποστηρίξη όλα αυτά τα χρόνια ήταν τεράστια και η παρακίνησή τους για την επίτευξη υψηλών στόχων συνεχής.

Κλείνοντας, θα ήταν αν μη τί άλλο πρόσφορο να παραθέσω μία μαντινάδα του Α.Χαιρέτη, στο νόημα της οποίας πιστεύω ακράδαντα ότι περικλείεται η στάση ζωής που θα πρέπει να ακουλουθούμε ακόμα και αν έχουμε κατ΄ ευφημισμόν φτάσει στην κορυφή, όπως αυτή ορίζεται και χαράσσεται από τον καθένα μας ξεχωριστά.

Και στην ψηλότερη κορφή ο δρόμος να με βγάλει,
΄πό πάνω μου ΄ναι ο ουρανός και χαμηλά ΄μαι πάλι.

The work reported in this thesis has been conducted at the Signal Processing Laboratory (SPL) of the Institute of Computer Science (ICS) of the Foundation for Research and Technology - Hellas (FORTH), and has been financially supported by a FORTH-ICS scholarship.

Abstract

Sparse and Low-Rank Techniques for Robust Speaker Recognition and Missing-Feature Reconstruction

Christos Tzagkarakis

University of Crete

Department of Computer Science

Doctor of Philosophy, 2014

Speaker recognition is the process for recognizing a speaker automatically, based on specific features extracted from the speech signal. It is divided in two distinct categories, namely, speaker *identification* and speaker *verification*. A broad range of applications exploits at its core the process of speaker recognition, where usually the presence of environmental noise in the speech signal impedes the inference of correct decisions. An additional factor, which contributes to the difficulty of recognizing a speaker correctly, is the limited amount of available training and evaluation data. This can be due to either a practical difficulty in obtaining a large volume of training data, or to the need to reduce the overall computational cost by using limited, yet reliable, evaluation data.

Focusing on overcoming the above limitations, while achieving high rates of successful recognition, this dissertation is divided in two main parts. In the first part, the problem of speaker recognition is reduced in an equivalent classification problem. To this end, we develop and study the performance of classification techniques, which are based on the framework of *sparse representations*, where we focus on the task of speaker identification by employing highly limited amounts of training and evaluation data, in environments with high levels of noise. The main assumption that governs these techniques is that the identified speech signal, and specifically the features that have been extracted from this signal, can be expressed as a sparse linear combination in terms of the columns of an overcomplete matrix, which is often referred in the literature with the term "dictionary". This dictionary is constructed appropriately from the available training data, while the computation of the sparse linear combinations is achieved via the solution of an optimization problem based on ℓ_p-norms ($p = 1$ or 2). The optimally estimated sparse weights of the linear combinations, the so-called *sparse codes*, which are obtained as the solutions of the optimization problem, are then employed for the final identification of the speaker based on a minimum reconstruction error criterion.

Extending our previous classification method based on sparse representations, we study the efficiency of a method for *discriminative dictionary learning*. This method estimates jointly the

dictionary comprising of the training data in conjunction with an appropriate linear classifier. The advantage of this approach is that it results in sparse codes, which are characterized by enhanced discriminative capability. For this experimental evaluation of the performance of our proposed method, through extensive simulations, a relatively small-sized database was used. The corresponding data were corrupted by several distinct types of environmental noise, for a wide range of signal-to-noise ratio values. Extensive comparisons with probabilistic models, which are based on the hypothesis that the extracted speech features follow a generalized Gaussian distribution, as well as with some of the state-of-the-art classification methods, such as Gaussian mixture models and joint factor analysis, revealed the superiority of our proposed method in terms of achieving higher correct recognition rates in noisy environments combined with the use of short training and testing speech data.

The second part of this dissertation focuses on the use of *low-rank techniques* as a powerful tool for extracting reliable features from a speech signal. More specifically, a technique for recovering a low-rank matrix is designed, which is employed for the reconstruction of those spectral regions of a speech signal, which are unreliable due to the presence of noise. The discrimination of the spectral regions is achieved by means of a reliability mask, which discriminates the regions characterized by the presence of noise from the regions which are dominated by the speech signal information. The completion of the empty spectral regions is performed based on the assumption that the logarithmic magnitude representation of a speech signal in the time-frequency domain, obtained via the short-time Fourier transform (STFT), is of low rank. Then, the Singular Value Thresholding (SVT) algorithm is exploited for the completion of those regions of the STFT representation, which are considered to be unreliable. The experimental evaluation of the proposed method reveals its power in extracting reliable features, which yield high rates of correct speaker identification in cases of high noise levels. The comparison against the widely used method of sparse imputation, which is based on sparse representations, reveals the superiority of our proposed approach in terms of achieving accurate speaker identification, especially for low levels of signal-to-noise ratios.

The above method does not take into account the existing prior knowledge with respect to the available training data, constituting essentially an unsupervised method. Motivated by this observation, we propose an extension of the matrix completion method, which exploits the prior knowledge that the data matrix is low rank, as well as the knowledge that the data can be represented efficiently in terms of a dictionary. In particular, we proposed an algorithm for joint low-rank representation and matrix completion (J-SVT). J-SVT is superior when compared with the standard SVT with respect to the computation of the low-rank representation of a data matrix in terms of a given dictionary, by employing a small number of observations from the original matrix. Through extensive simulations, we observed an improvement of the reconstruction error achieved by the J-SVT, in contrast to the typical SVT, for several distinct experimental scenarios.

Περίληψη

Τεχνικές Αραιής και Χαμηλής Τάξης Αναπαράστασης για Εύρωστη Αναγνώριση Ομιλητή και Ανακατασκευή Ελλιπών Χαρακτηριστικών

Χρήστος Τζαγκαράκης

Πανεπιστήμιο Κρήτης

Τμήμα Επιστήμης Υπολογιστών

Διδακτορική Διατριβή, 2014

Η αναγνώριση ομιλητή αποτελεί τη διαδικασία της αυτόματης αναγνώρισης του ατόμου που μιλάει, με βάση κάποια χαρακτηριστικά που εξάγονται από το σήμα φωνής. Χωρίζεται σε δύο επιμέρους κατηγορίες, και συγκεκριμένα στην ταυτοποίηση και στην επαλήθευση του ομιλητή. Ένα ευρύ φάσμα εφαρμογών έχει ως πυρήνα του την αναγνώριση ομιλητή, όπου συνήθως η παρουσία περιβαλλοντικού θορύβου στο σήμα φωνής δυσκολεύει την εξαγωγή σωστών εκτιμήσεων. Ένας επιπρόσθετος παράγοντας που συμβάλει στη δυσκολία σωστής αναγνώρισης αποτελεί η περιορισμένη ποσότητα δεδομένων εκπαίδευσης και δεδομένων αξιολόγησης. Αυτό μπορεί να οφείλεται είτε σε λόγους δυσκολίας απόκτησης μεγάλου όγκου δεδομένων εκπαίδευσης είτε στην ανάγκη να μειώσουμε το υπολογιστικό κόστος μέσω της χρήσης λίγων, αλλά αξιόπιστων, δεδομένων αξιολόγησης.

Στην προσπάθειά μας να αντιμετωπίσουμε τις παραπάνω δυσκολίες, επιτυγχάνοντας υψηλά ποσοστά επιτυχούς αναγνώρισης, η παρούσα εργασία χωρίζεται σε δύο μέρη. Στο πρώτο μέρος, το πρόβλημα της αναγνώρισης ομιλητή ανάγεται σε ένα πρόβλημα ταξινόμησης. Στην κατεύθυνση αυτή αναπτύσσουμε και μελετάμε συμπεριφορά τεχνικών ταξινόμησης που βασίζονται σε υποθέσεις αραιής αναπαράστασης, όπου επικεντρωνόμαστε στην εφαρμογή ταυτοποίησης ομιλητή με χρήση πολύ περιορισμένων δεδομένων εκπαίδευσης και αξιολόγησης, σε περιβάλλοντα με υψηλά επίπεδα θορύβου. Η βασική υπόθεση που διέπει τις συγκεκριμένες τεχνικές είναι πως το υπό ταυτοποίηση σήμα φωνής, και ειδικότερα τα χαρακτηριστικά που έχουν εξαχθεί από αυτό, μπορεί να γραφεί ως αραιός γραμμικός συνδυασμός ως προς ένα υπερπλήρη πίνακα, ο οποίος συχνά αναφέρεται στη βιβλιογραφία με τον όρο λεξικό. Το λεξικό αυτό κατασκευάζεται κατάλληλα από τα διαθέσιμα δεδομένα εκπαίδευσης, ενώ η εύρεση των αραιών γραμμικών αναπαραστάσεων επιτυγχάνεται μέσω της επίλυσης ενός προβλήματος βελτιστοποίησης με βάση την ℓ_p-νόρμα ($p = 1$ ή 2). Τα βέλτιστα εκτιμώμενα αραιά βάρη των γραμμικών συνδυασμών, οι επονομαζόμενοι και αραιοί κώδικες, που προκύπτουν ως λύσεις του προβλήματος βελτιστοποίησης, χρησιμοποιούνται για την τελική ταυτοποίηση του ομιλητή μέσω ενός κανόνα ελάχιστου σφάλματος ανακατασκευής.

Επεκτείνοντας την παραπάνω μέθοδο ταξινόμησης μέσω αραιής αναπαράστασης, εξετάζουμε την εφαρμογή μίας μεθόδου διακριτικής εκμάθησης λεξικού. Με την μέθοδο αυτή εκτιμάται από κοινού το λεξικό που περιέχει τα δεδομένα εκπαίδευσης μαζί με ένα κατάλληλο γραμμικό ταξινομητή. Το πλεονέκτημα αυτής της προσέγγισης είναι ότι οδηγεί στην παραγωγή αραιών κωδίκων οι οποίοι χαρακτηρίζονται από μεγαλύτερη διακριτική ικανότητα. Κατά τη διάρκεια της πειραματικής αξιολόγησης της απόδοσης αυτής της μεθόδου, μέσω προσομοιώσεων, χρησιμοποιήθηκε μία σχετικά ολιγομελής βάση δεδομένων. Στα δεδομένα αυτά προστέθηκαν διάφορα είδη περιβαλλοντικού θορύβου για ένα ευρύ σύνολο τιμών σηματοθορυβικού λόγου. Οι εκτενείς συγκρίσεις που πραγματοποιήθηκαν τόσο με πιθανοτικά μοντέλα, τα οποία βασίζονται στην υπόθεση ότι τα χαρακτηριστικά της φωνής ακολουθούν γενικευμένη Gaussian κατανομή, όσο και με μερικές εκ των κορυφαίων μεθόδων ταξινόμησης, όπως μοντέλα μίξης Gaussian κατανομών και κοινής παραγοντικής ανάλυσης, ανέδειξαν την υπεροχή της προτεινόμενης μεθόδου αναφορικά με την επίτευξη υψηλότερων ποσοστών σωστής ταυτοποίησης σε περιβάλλοντα θορύβου σε συνδυασμό με τη χρήση περιορισμένης ποσότητας δεδομένων εκπαίδευσης και αξιολόγησης.

Το δεύτερο μέρος της διατριβής μελετάει τη χρήση τεχνικών χαμηλής τάξης ως ένα εργαλείο για την εκτίμηση αξιόπιστων χαρακτηριστικών φωνής. Ειδικότερα, εφαρμόζεται μία τεχνική ανάκτησης πίνακα χαμηλής τάξης για την ανακατασκευή εκείνων των φασματικών περιοχών του σήματος φωνής, οι οποίες δεν είναι αξιόπιστες εξαιτίας της έντονης παρουσίας θορύβου. Ο διαχωρισμός αυτών των φασματικών περιοχών επιτυγχάνεται με τη βοήθεια μιας μάσκας αξιοπιστίας, η οποία διακρίνει τις περιοχές που χαρακτηρίζονται από παρουσία θορύβου σε σχέση με τις περιοχές στις οποίες επικρατεί η πληροφορία του σήματος φωνής. Η συμπλήρωση των κενών φασματικών περιοχών πραγματοποιείται βάσει της υπόθεσης ότι η λογαριθμική αναπαράσταση πλάτους ενός σήματος φωνής στο πεδίο χρόνου-συχνότητας μέσω του short-time μετασχηματισμού Fourier (STFT) είναι χαμηλής τάξης. Κατόπιν, ο Singular Value Thresholding (SVT) αλγόριθμος υιοθετείται για την συμπλήρωση των περιοχών της STFT αναπαράστασης που θεωρούνται ως μη αξιόπιστες. Η πειραματική αξιολόγηση της προτεινόμενης μεθόδου αναδεικνύει την ισχύ της στον υπολογισμό αξιόπιστων χαρακτηριστικών τα οποία οδηγούν σε αρκετά υψηλά ποσοστά σωστής ταυτοποίησης ομιλητή σε περιπτώσεις όπου τα επίπεδα θορύβου είναι υψηλά. Η σύγκριση με την ευρέως χρησιμοποιούμενη μέθοδο της sparse imputation, η οποία βασίζεται στην υπόθεση αραιής αναπαράστασης, φανερώνει την ανωτερότητα της προτεινόμενης μεθόδου αναφορικά με την επίτευξη ακριβούς ταυτοποίησης ομιλητή, για χαμηλά επίπεδα σηματοθορυβικού λόγου.

Η παραπάνω μέθοδος δε λαμβάνει υπόψη την εκ των προτέρων γνώση που υπάρχει σχετικά με τα δεδομένα εκπαίδευσης που έχουμε στη διάθεσή μας, αποτελώντας ουσιαστικά μία μέθοδο χωρίς επίβλεψη. Έχοντας αυτή την παρατήρηση ως κίνητρο, προτείνεται μία επέκταση της μεθόδου συμπλήρωσης πίνακα η οποία εκμεταλλεύεται την εκ των προτέρων γνώση ότι ο πίνακας δεδομένων είναι χαμηλής τάξης, καθώς και τη γνώση ότι τα δεδομένα μπορούν να αναπαρασταθούν με αποτελεσματικό τρόπο ως προς ένα λεξικό. Ειδικότερα, προτείνουμε έναν αλγόριθμο από κοινού αναπαράστασης χαμηλότερης τάξης και συμπλήρωσης πίνακα (J-SVT). Ο J-SVT υπερέχει του κλασικού SVT στον υπολογισμό της αναπαράστασης χαμηλότερης τάξης ενός πίνακα δεδομένων ως προς ένα δοσμένο λεξικό χρησιμοποιώντας λίγες παρατηρήσεις από τον αρχικό πίνακα. Μέσω προσομοιώσεων παρατηρείται η βελτίωση του σφάλματος ανακατασκευής που επιτυγχάνει ο J-SVT σε αντίθεση με τον τυπικό SVT, για διάφορα πειραματικά σενάρια.

Contents

List of Figures

List of Tables

List of Abbreviations

APD	**A**mplitude **P**robability **D**ensity
CIR	**C**orrect **I**dentification **R**ate
EM	**E**xpectation-**M**aximization
FFT	**F**ast **F**ourier **T**ransform
GGD	**G**eneralized **G**aussian **D**istribution
GMM	**G**aussian **M**ixture **M**odel
UBM-GMM	**U**niversal **B**ackground **M**odel-**G**aussian **M**ixture **M**odel
JFA	**J**oint **F**actor **A**analysis
J-SVT	**J**oint low-rank representation and matrix completion version of **S**ingular **V**alue **T**hresholding algorithm
KLD	**K**ullback **L**eibler **D**ivergence
LDA	**L**inear **D**iscriminant **A**nalysis
LRR	**L**ow-**R**ank **R**epresentation
MAP	**M**aximum **a** **P**osteriori
MC	**M**atrix **C**ompletion
MFCCs	**M**el-**F**requency **C**epstral **C**oefficients
ML	**M**aximum **L**ikelihood
OMP	**O**rthogonal **M**atching **P**ursuit
PDF	**P**robability **D**ensity **F**unction
SRC	**S**parse **R**epresentation **C**lassification
SI	**S**parse **I**mputation
SNR	**S**ignal-to-**N**oise **R**atio
STFT	**S**hort-**T**ime **F**ourier **T**ransform
SVT	**S**ingular **V**alue **T**hresholding
SVD	**S**ingular **V**alue **D**ecomposition

List of Symbols

$\mathbf{A}$	matrix
$\mathbf{A}^{-1}$	inverse matrix
$\mathbf{A}^{T}$	transpose matrix
$\lvert\mathbf{A}\rvert$	matrix determinant
$\mathbf{x}$	column vector (unless stated otherwise to be a row vector)
$\mathbf{x}^{T}$	transpose (row) vector
$\mathcal{O}(\cdot)$	order of $(\cdot)$
$\lVert\cdot\rVert$	vector norm
$\lVert\cdot\rVert_F$	Frobenius norm
$\lVert\cdot\rVert_*$	nuclear matrix norm
$\mathcal{A}$	linear map
$\mathcal{A}^*$	adjoint of a linear map
$\log$	natural logarithm

List of publications

The following publications were produced during the PhD study:

- C. Tzagkarakis, S. Becker and A. Mouchtaris, “Missing Data Imputation Based on Joint Low-Rank Representation and Matrix Completion”, journal paper under preparation.

- C. Tzagkarakis, S. Becker and A. Mouchtaris, “Joint Low-Rank Representation and Matrix Completion Under a Singular Value Thresholding Framework”, *in Proc. of European Signal Processing Conference (EUSIPCO ’14)*, Lisbon, Portugal, September 2014.

- C. Tzagkarakis and A. Mouchtaris, “Reconstruction of Missing Features Based on a Low-Rank Assumption for Robust Speaker Identification”, *invited paper in Proc. of International Conference on Information, Intelligence, Systems and Applications (IISA ’14)*, Chania, Greece, July 2014.

- C. Tzagkarakis and A. Mouchtaris, “Sparsity Based Noise Robust Speaker Identification Using a Discriminative Dictionary Learning Approach”, *in Proc. of European Signal Processing Conference (EUSIPCO ’13)*, Marrakech, Morocco, September 9-13, 2013.

- C. Tzagkarakis and A. Mouchtaris, “Robust Speaker Identification Using Matrix Completion Under a Missing Data Imputation Framework”, *in Proc. 2013 Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS ’13)*, Lausanne, Switzerland, July 8-11, 2013.

- A. Griffin, T. Hirvonen, C. Tzagkarakis, A. Mouchtaris and P. Tsakalides, “Single-channel and Multichannel Sinusoidal Audio Coding Using Compressed Sensing”, *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 19 (5), pp. 1382-1395, July 2011.

- 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.

- A. Griffin, C. Tzagkarakis, T. Hirvonen, A. Mouchtaris and P. Tsakalides, “Exploiting the Sparsity of the Sinusoidal Model Using Compressed Sensing for Audio Coding”, *in Proc. Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS ’09)*, Saint-Malo, France, April 6-9, 2009.

Structure of the thesis

The present thesis deals with the problem of noise robust speaker identification under limited training and testing speech data. Our goal is twofold. First, we would like to study the efficiency of sparsity-based and discriminative dictionary learning classification methods within the context of highly limited amount of training and testing instances under noisy environments. Second, we leverage prior knowledge that the speech log-magnitude spectrotemporal representation is low-rank in order to apply a missing data imputation method based on matrix completion. We aim at enhancing the reliability of speech features using a matrix recovery technique based on singular value thresholding algorithm. An extended version of standard matrix completion that leverages prior knowledge that the matrix is low-rank and that the data samples can be efficiently represented by a fixed known dictionary is also proposed. The thesis is organized as follows:

Chapter 1

This chapter provides the necessary background information of the speaker recognition research area. We describe the main categories of a speaker recognition system associated with the specific task undertaken. Besides, we analyze the basic compensation techniques adopted to overcome the various robustness issues arisen in noisy environments. A brief overview of missing data techniques is also given based on the use of reliability masks for producing reliable speech features fed into a speaker recognition system as well as the core research work corresponding to the speaker recognition problem under short training and evaluation speech data. The chapter concludes by discussing our motivation and listing the main contributions of the this dissertation.

Chapter 2

In this chapter, we describe a sparsity-based classification approach proposed within the context of noise robust speaker identification using a limited amount of training and testing speech data.

We assume that each test instance can be sparsely represented as a linear combination of all the training data (used to construct a dictionary) which have been obtained during the enrollment phase. Specifically, we exploit the fact that the test instances coming from a certain speaker can be expressed as a linear combination of the training instances associated with the same speaker. The optimally estimated sparse weights of this linear combinations, dubbed as sparse codes, are computed as the solutions of a sparse optimization problem. The estimated sparse codes are then employed for the final identification of the speaker based on a minimum reconstruction error criterion. This method is compared with a proposed probabilistic model, which is based on the assumption that the extracted speech features follow a generalized Gaussian distribution, as well as with some of the state-of-the-art speaker identification techniques revealing the superiority of the sparsity-based approach under the constraint of using short test and training sessions in noisy conditions.

Chapter 3

In this chapter, the second proposed method is presented for solving the noise robust speaker identification problem using a limited amount of training and testing utterances. In particular, we aim at learning an overcomplete dictionary, resulting in highly discriminative sparse codes, along with a linear classifier. This estimation is performed in a joint fashion by imposing additional constraints on the associated objective function in order to produce similar sparse codes for those training samples belonging to the same speaker. This is in contrast to the sparse representation classification (SRC) approach introduced in the previous chapter, which do not treat jointly the estimation of the dictionary, the sparse codes, and the classifier parameters. Several experiments comparing the discriminative dictionary learning technique with a UBM-GMM system, as well as with the SRC approach show that the proposed method performs better than the other two methods in the case of small amount of training data, and is very robust to noisy conditions.

Chapter 4

This chapter describes a method for missing-feature reconstruction applied in the context of noise robust speaker identification using short training and testing data. Reconstruction of missing features promotes robustness in speaker recognition applications under noisy conditions. The low-rank behaviour of the log-magnitude spectrotemporal speech data is exploited in the framework of missing data imputation, where a low-rank matrix recovery approach based on singular value thresholding (SVT) algorithm is applied to reconstruct the unreliable spectrographic data due to noise corruption. Experiments on real speech data performed to compare its performance with the recently introduced sparse imputation technique showing that the proposed technique achieves an improved performance in terms of higher correct identification rates especially for low signal-to-noise ratio (SNR) scenarios.

Chapter 5

In this chapter, an extension of the SVT-based low-rank matrix completion method for missing-feature recovery is described. In particular, the approach analyzed in the previous chapter does not take into account the existing prior knowledge with respect to the available training data, constituting essentially an unsupervised method. This observation motivate us propose an extension of the matrix completion method, which exploits the prior knowledge that the data matrix is low rank, as well as the knowledge that the data can be represented efficiently in terms of a dictionary. In particular, we propose an algorithm for joint low-rank representation and matrix completion (J-SVT). J-SVT is superior when compared with the standard SVT with respect to the computation of the low-rank representation of a data matrix in terms of a fixed dictionary, by employing a small number of observations from the original data matrix. Through several simulations, we show that the reconstruction error achieved by the J-SVT is lower with respect to the typical SVT, for several distinct experimental scenarios.

Chapter 6

This chapter serves as a conclusion and summarization of the main results of this thesis and provides directions for future work.

Chapter

1

Introduction

Somewhere, something incredible is waiting to be known.

Carl Sagan (1934-1996)

Speaker recognition concerns the task of recognizing the identity of a claimed speaker. The voice signal constitutes the core ingredient on which recognition is based. Vocal quality characteristics associated with the rhythm and verbal idioms, pronunciation and intonation style, etc. strongly affects the recognition accuracy. All these quality parameters should be jointly taken into consideration for building a reliable practical speaker recognition system.

Speaker recognition [1, 2, 3, 4] can be categorized into *speaker identification* and *speaker verification*. Generally speaking, speaker verification is a one-to-one matching process where one speaker's voice is matched to one template whereas speaker identification is a one-to-many match where the voice is compared against a specific number of voice patterns. Speaker identification [5, 6, 7, 8, 9, 10, 11] is defined as the task of determining an unknown speaker's identity. It can in turn be distinguished into two categories with respect to the speakers' set structure. Specifically, in the case of closed-set speaker identification we make the assumption that the voice signal coming from the unknown speaker must belong to a fixed and known set of speakers. Otherwise, we have an open-set speaker identification [12, 13, 14], where the speakers that are not members of the set of known speakers are categorized as impostors.

In speaker verification, a speaker claims to be of a certain identity and his/her voice is used to verify this claim. It can be considered as a binary hypothesis problem, where the goal is to discern whether the voice of the speaker under verification comes from a person whose voice has been enrolled into the speech corpus and as a result is known to the system from a potentially large group of voices unknown to the system. The combination of speaker verification with open-set speaker identification leads to a system where the speaker is initially detected as an impostor or non-impostor, and if the speaker is positively accepted then the specific speaker's identity is estimated according to the speakers enrolled in the database.

Speaker recognition can be adopted in a broad range of applications. First of all, *security* constitutes the main core of diverse applications spanning from the control of entrance to restricted areas (e.g. military facilities, governmental buildings etc.) to telephone banking and commerce where the individual's voice is used to ensure a secure financial transaction. Typically, a specific password or special phrase should be uttered in order to proceed with the whole process. A second possible application is *media indexing* based on voiceprints. A multimedia collection such as online movie database and broadcast news, audio books archives etc. can be automatically indexed by using a speaker recognition system allowing the user to navigate and access the audiovisual material based on content. Speaker recognition can also be applied in the context of *forensics*, where a sample of a suspect's voice can be used as evidential material within a court room or as an investigation tool during a criminal investigation.

Apart from the above, speaker recognition could play a crucial role in *ambient intelligence environments*. During the last years a large interdisciplinary effort has been carried out towards researching and studying problems in which the computer ceases to constitute a physical object and turns into a pervasive presence in the surrounding field interacting with the user in various ways. Imagine, for example, a typical meeting room where a simple equipment based on a speaker recognition system could track the current speaker and take on-the-fly decisions about changing camera orientation, automatic change of presentation slides, a personalized interaction with the teleconference system etc. Besides, all the speaker-centric information produced during the meeting could easily be used in an off-line mode for meeting transcription as well as for speaker diarization, namely estimating who spoke when.

Speaker recognition methods can also be divided into *text-dependent* and *text-independent* methods. The former [15] requires the speaker to provide fixed utterances of keywords or sentences, the same text being used for both training and recognition. However, in text-independent recognition, the decision does not rely on a specific text being spoken. Thus, the speech training data and the testing utterances of the same speaker may have completely different linguistic and phonetic content which should be taken into account during the recognition process. Text-independent method constitutes a more challenging task and occurs in most practical situations compared to the text-dependent one. In the next sections we will describe the main reasons affecting the accuracy of a speaker recognition system and we will also mention our motivation as well as the basic contribution points of the current thesis.

1.1 Robustness in speaker recognition

The efficiency of practical speaker recognition systems in most cases is strongly affected by the presence of noise, reverberation or other distortion factors usually associated with the transmission medium of the speech signal or phonetic variability issues. Let us imagine, for example, the case of an individual located in an external environment (e.g. a sidewalk near a busy street) talking at the mobile phone and trying to accomplish a financial transaction. The ambient noise permeates the entire voice signal and as a result reducing the recognition performance which in turn causes a transaction failure. As an extension of the previous example, consider the case of channel/handset mismatch between training and testing phases, where the training data have been recorded via a mobile phone but during the recognition process the speaker uses a landline phone to communicate with the system.

All these and many other practical examples indicate the fact that the robustness issue and thus the accuracy of a speaker recognition system is related with the mismatch conditions between the training data and the available data during the recognition process often termed as *session variability* [16, 17]. The emotional status of the speaker such anxiety, sadness, happiness, etc. can also produce session variations. Even in the case of ambient intelligence environments such as smart office or smart rooms where noise levels are almost negligible, some session variations can occur because of the air conditioner operation sound or possible changes in the speaker and acquisition terminal distance resulting in important differences in recorded voice signals and thus leading to poor recognition performance [18].

1.1.1 Compensation methods

A plethora of techniques have been developed through the last decades to compensate for the training/testing mismatch conditions. In particular, many approaches have been proposed in the feature, model and match-score context in order to deal with the robustness issue. In *feature-based* compensation methods the sequence of feature vectors (generally corresponding to a short-term spectral representation) associated with a speaker's utterance is subject to invariance enhancement to non-speaker vocal quality information within the input speech signal. In specific, the Cepstral Mean Normalization (CMN) compensation methods [19, 20, 21, 22] exploit the fact that the noise is additive in the log-spectral domain, reducing in this way the linear filtering effects as expressed in various channel distortions. Under the noise additivity assumption in the log-spectral domain and the fact that the channel signal does not significantly vary over the duration of an utterance, CMN aims to alleviate the distortion effects by subtracting from each feature vector the average mean of those. CMN can be extended to Cepstral Mean and Variance Normalization (CMVN) process [23] for equalizing the variances of the features by dividing each feature vector by its standard deviation where a sliding window strategy is followed. The window should be long enough to allow good estimates for the mean and variance, yet short enough to capture time-varying properties of the channel.

The relative spectral (RASTA) filtering approach [24] is a channel compensation method exploiting the inherent differences between the temporal properties of distortion effects and the temporal properties of the speech. Specifically, RASTA performs a band-pass filtering where highly and slowly varying frequency components lying out of the filter bandwidth are eliminated as considered to contain non-speech information. The RASTA method can be regarded in general as an evolved version of CMN where except from the noise components those components which evolve in such a way as to be considered that do not contain information about the speech are also eliminated.

Another group of methods is related with the modification of the speech signal's power spectral representation. The source-filter speech production model is assumed towards adopting the linear prediction (LP) methodology during the feature extraction process. The pitch information is captured within the LP residual, while the LP filter response models the vocal tract characteristics [25]. This information discrimination reveals the vocal tract properties in noisy conditions. Additional weighting approaches can improve robustness such as liftering [22] which enforces the computation of low order coefficients against the noise sensitive higher order coefficients as well as postfiltering [26] which gives emphasis on formant regions based on the assumption that the noise effect is eliminated in these regions.

Channel variability compensation and enhanced speaker recognition accuracy can be achieved in light of feature transformation approaches. A feature transform is computed to convert speaker-dependent features to speaker-independent features. The application of inverse transform into the noisy speech features can reduce the distortion effects [22]. In [27] the transformation parameters are utilized in the feature domain to perform maximum a-posteriori adaptation from a channel independent model to a set of channel dependent models. Feature warping [28] and short-time Gaussianization [29] have also been proposed which involve modification of the short-term feature distribution to match a target distribution. It is assumed that the clean (cepstral) features follow a specific distribution (for example a Gaussian distribution), which is altered by the additive noise and channel distortions. This modification is achieved by warping the cumulative distribution function of the features in order to match the reference distribution function.

The key idea behind *model-based* compensation is the modification of speaker model parameters instead of handling speech features per se as a solution in learning the noise characteristics without requiring the explicit identification and labeling of different conditions. Examples of such methods include the speaker-independent variance transformation [30] and the transformation for synthesizing supplementary speaker models for other channel conditions that have not been presented during the enrollment phase [31]. This can be practically achieved by building a channel-independent multicondition background model using all training data from a collection of various channels, while channel-dependent models are constructed via maximum a-posteriori adaptation and used to learn transformations between different channels. The training/testing mismatch may be lessened by synthesizing a training channel type to a testing channel type during the recognition process. Multicondition training data are also utilized to jointly model the inter-speaker (i.e., the set of special characteristics distinguishing different speakers) and channel variability under the factor analysis scheme [32, 33]. The basic idea lies in the decomposition of the session variability component in a low-dimensional acoustic subspace.

There have also been proposed other techniques in which the focus is on noise compensation, for example, parallel model combination [34, 35, 36], or Jacobian environmental adaptation [37, 38], assuming the availability of a statistical model of the environment or noise.

Score-based compensation methods are mainly used in speaker verification task and the main goal is to enforce scores from different speakers to fall into a similar range so that a common speaker-independent threshold can be used. Before proceeding further it would be helpful to mention that during the training process in a speaker recognition system we usually built a (probabilistic) model for each speaker belonging to the database. When we want to recognize the speaker we have to evaluate the likelihood of the test utterance with respect to the trained model and thus the so-called likelihood ratio scores are produced. The most dominant score-domain methods include handset dependent score normalization (H-norm) providing robustness to channel variability through the construction of Gaussian mixture models (GMMs) to model non-linear uncompensated channel effects within each of the relevant conditions [39]. During recognition the test segment is assigned a handset type classification based on the handset GMMs, and the speaker GMM likelihood is modified by normalization with the handset model parameters. The offline estimation of the normalization parameters is allowed in Z-norm [40] method, where the explicit labeling of each test utterance according to its channel type is not required. Z-norm approach can be extended by scaling the score distribution with the variance of the imposter scores giving rise to the T-norm [41] method.

1.1.2 Missing data methods

It is of high importance to notice that all the methods described above were developed during a long-term research effort to deal with the problem of robustness in the context of speaker recognition. However, it still remains quite difficult in many practical cases to successfully apply these compensation approaches in order to achieve high accuracy recognition results. This is mainly due to the fact that treatment of the environmental noise is hard to be induced by the majority of the aforementioned compensation methods as opposed to the protection against channel distortions. More specifically, feature-based compensation techniques cannot handle speech signals corrupted by environmental noise without the availability of matched models although their modeling behaviour is quite robust in the case of linear channel effects. Additionally, a limiting factor of model-based compensation methods is the requirement of noise characteristics' knowledge which is needed for adaptation performance. As a general conclusion, it could be stated that in dynamic and mutable environments dominated by highly non-stationary and transient noise it is very difficult to provide sufficient levels of robustness by using these methods.

As a step towards building a more robust speaker recognition system in order to remove the effects of non-stationary and transient environmental noise behaviour we could consider the idea of using only those special feature components which are supposed to contain *reliable information* about the voice signal at hand. *Missing data techniques* are based on this features reliability assumptions and the effort is given on achieving enhanced robustness by enabling the computation of reliable speech features under adverse noisy conditions. This technique was firstly proposed in the context of computer vision and especially for recovering partially occluded images for recognition tasks [42, 43] Missing data approaches were later extended in order to mimic the ability of human auditory system which can efficiently process distorted speech signals [44, 45]. In particular, consider a two-dimensional spectrotemporal representation of a noisy speech signal which 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 in a speaker recognition system [46], while other regions of the time-frequency representation are mostly corrupted by background noise and thus labeled as unreliable or missing spectrotemporal data. Missing data techniques are heavily based on the *missing data mask* which constitutes a matrix indicating the reliable as well as the unreliable spectrotemporal elements of a noisy speech signal. The accurate estimation of the reliability mask is very crucial for the labeling of missing spectrographic regions.

Missing data techniques were firstly introduced in automatic speech recognition (an overview can be found in [47]). They can be distinguished into two main categories, namely imputation and marginalization. *Imputation* [48, 49, 50, 51, 52, 53, 54, 55] is defined as the technique of substituting missing time-frequency components with an estimate of the time-frequency component value based on speech signal's high degree of redundancy. In *marginalization* [56, 57, 58, 59], missing spectrotemporal regions are ignored and thus, recognition is based on the reliable components of the noisy speech signal's time-frequency representation, where observation likelihoods are computed by integrating over the range of possible values of the missing components. All 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.

Recently, a lot of research has been carried out in the field of speaker recognition wherein the missing data strategy has been followed to minimize the side effects caused due to noise presence in speech signals. In specific, speaker identification is examined in [60, 61, 62], while in [63, 64, 65] speaker verification is studied in the light of missing feature theory for improvement of recognition performance, while in [66] 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.

Imputation appears to be somewhat advantageous compared with marginalization especially due to the fact that after the reconstruction of the missing time-frequency components with clean estimates, the new (reconstructed) time-frequency features can be directly applied to any recognition system which has been trained on undistorted speech data. Hence, we can deduce that imputation can be characterized as a system-independent method operating as a "black box" which can be inserted in any recognition system as a noise robustness tool. Another benefit of imputation is that recognition accuracy is not affected at high and moderate signal-to-noise ratio's (SNRs) regimes. However, typical imputation methods fail to preserve the recognition performance at lower SNR values approximately below 5 dB. This decline in performance is primarily attributed to the fact that at low SNRs the percentage of spetrotemporal regions assigned as missing (or unreliable) is too high in relation to the total number of time-frequency components, and therefore it is difficult to achieve good clean estimates as a consequence of limited reliable data. Besides, the stochastic nature of both speech and noise signal produces heterogeneous speech-dominated and noise-dominated time-frequency regions. This complicates the modeling of the problem which is usually based on local information and various correlation properties of the reliable time-frequency areas.

Another reason for imputation inefficiency at low SNR values constitutes the practical reliability mask estimation. Practically, reliability mask should be estimated algorithmically based upon the noisy voice signal as well as the available speech training data. In other words, there is an analogy between mask's quality estimation and performance of the imputation method. In this thesis, we are mainly interested in using an ideal (or oracle) reliability mask and thus, we do not intend to deepen into a rigorous description of practical masks estimation algorithms. For a more detailed overview on mask estimation methods, the interested reader is referred to [67] and the citations therein.

The concept of *sparse representation* has also been exploited recently in the realm of missing data imputation, attempting to recover missing data spectrotemporal areas. The basic assumption is that the signal's spectral representation can be expressed as a sparse linear combination of elements from an appropriately chosen dictionary. Sparse representation techniques falls into the compressive sensing framework [68, 69] which states that signals that are sparse or compressible in a suitable transform basis can be recovered from a highly reduced number of incoherent linear random projections, as opposed to the traditional signal processing paradigms, which are dominated by the typical Shannon-Nyquist sampling theorem. In [50, 53] the solution of an ℓ_1-norm optimization problem is proposed towards solving missing data imputation under the concept of sparse representation. In specific, *sparse imputation* –a term introduced in [53]–states that missing speech spectra can be reconstructed by expressing them as a sparse linear combination of dictionary elements called examples. After several experiments it was found that sparse imputation could produce quite good performance especially for low SNR values in the context of automatic speech recognition.

1.2 Speaker recognition using limited data

In the previous section, general information about speaker recognition systems was presented along with how various robustness issues arise in noisy conditions can be dealt with using compensation methods and missing data techniques. An additional key factor that also puts at risk the performance of speaker recognition systems is the available amount of training and testing data used during the recognition process. An obvious rule of thumb is that the more the amount of data we have, the more accurate recognition rates will occur. However, in several practical scenarios we could assume a limited amount of training and evaluation speech data. One reason for that could be that it is often not feasible to have large amounts of training data from all the speakers. Let us consider for example, the realistic scenario where the entrance of a smart room in an ambient intelligence building [1] is equipped with a microphone recording the speech of every person who wants to access the room. Suppose now that this person appears for the first time in front of the entrance, and would like to have constant access in the future. Practically speaking and based on the aforementioned rule of thumb about the specific amount of training data we would like to have as much speech data as possible from that person. Nonetheless, it is somewhat frustrating for the speaker to be for a long period of time in front of the microphone while recording voice data, especially when the entrance is located outside of the building. Thus, we are interested in acquiring a limited amount of training data while keeping the recognition rate high. Secondly, we could notice that in order to speed up the recognition process, the evaluation data should be as short as possible. Especially in cases where the recognition procedure is performed under constrained computational resources (e.g. recognition performed using a mobile phone) it is a need for achieving low-latency response.

[1] The described scenario constitutes a practical research problem as a part of the AmI programme http://www.ics.forth.gr/ami/

Some interesting works have been carried out in the field of speaker recognition under short training and evaluation data assumptions with an emphasis given on speaker verification. In [70], the importance of speech detection process when applied in short duration speech data is highlighted and the limits of both GMM and support vector machine-based system with a GMM supervector linear kernel are examined under a maximum a posteriori (MAP) adapted mean parameters context. It is also indicated that eigenvoice modeling could increase the performance. A joint factor analysis (JFA) model [33] is extended in [71] such as to independently optimize the speaker and session variability subspaces, where it is shown that for speaker verification based on short utterances it is important for the session subspace to be trained with matched length utterances, while the speaker subspace should be trained using as much data as possible. JFA model is also used in [72] where i-vectors are combined with normalization techniques such as within-class covariance normalization, linear discriminant analysis, scatter difference nuisance attribute projection and Gaussian probabilistic linear discriminant analysis. A minimax strategy is used in [73] in order to estimate the first order statistics as a step towards increasing the robustness of the extracted i-vectors for solving the problem of i-vectors' uncertainty representation when computed using a small number of feature vectors.

A top-down bottom-up method using test token histograms is studied in [74] for the problem of in-set/out-of-set speaker recognition. The core idea is based on filling acoustic holes and fortifying the acoustic information using the claimed speaker's test token histogram adopting a modified scheme of GMM model. Additionally, a dimension-decoupled version of GMM is proposed in [75] to deal with the problem of small sets of training and evaluation voice data examined on speaker identification task. In particular, a novel way to reduce the number of necessary free parameters in the GMM is proposed in order to obtain more stable statistical estimates of model parameters and likelihoods using less amount of data. An exemplar-based sparse presentation approach is followed in [76], where sparse discriminant analysis and probabilistic linear discriminant analysis techniques are used to model the sparse exemplar activations for speaker identification. The work presented in [77] comes as an extension of [76], where a group sparsity constraint is introduced under a spectral factorization framework in order to limit the number of active speakers from multiple candidates and managing to narrow down the set of speakers to be active at a time.

According to the works briefly described above it is obvious that speaker recognition based on a small amount of training and evaluation voice data is a relatively modern research problem, gradually begun to be studied during the last few years. However, it is clear that there is fertile ground for further research, especially within the context of robust speaker recognition which constitutes the main goal of the current thesis.

1.3 Contributions of the thesis

In this study, *our aim is to examine the efficiency of techniques heavily based on sparse and low-rank assumptions targeted at noise robust text-independent speaker identification using a limited amount of training and testing speech data.* In specific, the current thesis can be distinguished into two parts. In the first part, we examine the efficiency of classification methods based on sparse representation of the available features. The focus is given on using short training and testing sessions in adverse noisy conditions. In the second part, we study the problem of recovering reliable speech features based on missing data imputation by exploiting the low-rank behaviour of the speech spectrotemporal representation. The target application task is the same as in the first part.

The main contributions of this thesis can be summarized as follows:

- The lack of an extensive research work of how sparsity-based classification behaves under the noisy speaker identification task using a limited amount of training and evaluation speech features led us to introduce the sparse representation classification (SRC) in order to examine its robustness efficiency. Speech features are extracted from all the training data and used to build a dictionary, where during the identification process each test feature vector can be represented as a linear combination of a few columns of the dictionary which belong to the same speaker. The optimally estimated sparse weights of the linear combinations are called sparse codes and computed via a solving an optimization problem based on ℓ_p-norms, where $p = 1$ or $p = 2$.

- Speaker identification is treated as multiple hypothesis problem based on a statistical modeling approach. We exploit the statistical property that the extracted mel-frequency cepstral coefficients (MFCCs) follow a generalized Gaussian distribution (GGD). After estimating the GGD parameters of all the training and testing feature vectors the Kullback-Leibler divergence (KLD) is adopted for computing the identity of the speaker.

- The raw data choice of dictionary elements in SRC context as well as the large size of the dictionary motivate us to use a discriminative dictionary learning technique. We aim at finding a smaller dictionary whose elements will be chosen in such a way in order to produce highly discriminative sparse codes which would lead in better classification results. This task is performed by jointly estimating a dictionary built by the training data and an appropriate linear classifier.

- We take advantage of the speech signal's low-rank property in the log-magnitude STFT domain in order to generate reliable speech features before the identification procedure. An ideal binary reliability mask is used to distinguish the speech-dominated spectrotemporal regions from the noise-dominated ones. The missing regions are completed through the application of Singular Value Thresholding (SVT) algorithm and thus a reliable STFT spectrogram is recovered. A great advantage of the SVT-based proposed reconstruction method is that it produces a reliable STFT spectrogram, which means that any type of speech features based on STFT representation can be extracted and further used as input to any classifier.

- We propose a supervised version of SVT which estimates low-rank representation and matrix completion in a joint fashion. SVT-based recovery algorithm acts in a unsupervised manner because it does not take into consideration the existing prior knowledge with respect to the available training data. This observation motivate us to propose an extension of the SVT-based method, which exploits the prior knowledge that the data matrix is low rank, as well as the knowledge that the data can be represented efficiently in terms of a dictionary which is built using the training data. The proposed algorithm is named J-SVT and estimates the low-rank representation of a data matrix in terms of a given dictionary, by employing a small number of observations from the original data matrix.

Part I

Sparsity-based techniques for speaker identification

Chapter

2

Sparse representation classification for speaker identification

No problem is too small or too trivial if we can really do something about it.

Richard Feynman (1918-1988)

2.1 Introduction

As it was mentioned in Chapter 1, speaker recognition systems are essential in a variety of security and commercial applications, such as information retrieval, control of financial transactions, control of entrance into safe or reserved areas and buildings, *etc.* [3]. Speaker recognition can be based on both the separate or combined use of several biometric features [78] (voice, face, fingerprints, *etc.*). In the current study, we focus on speaker identification using only voice patterns.

In order to correctly identify a person, each speaker in the database is usually assigned a specific speaker model consistently describing the extracted speech features. During the identification process, the system returns the speaker's identity based on the closest matching of the test utterance against all speaker models. This procedure has proven to be effective under acoustic conditions in matched training and testing [5]. However, in practical applications where speech signals are corrupted by noise due to either the environment in which the speaker is present (*e.g.* the user is crossing a busy street) or due to the voice transmission medium (*e.g.* the user is speaking through a cell-phone), robust identification is a challenging problem. Figure 2.1 shows the structure of a typical speaker identification system. It is distinguished into two phases. During the training (or enrollment) phase a model is built for each speaker in the database with respect to the available speech training data. In order to identify an unknown speaker, a speaker model is also built according to the testing data and the test speaker model is compared to all the trained speaker models. This comparison is appropriately evaluated using a matching rule and the result of this matching process provide us the estimated (or the most probable) identity.

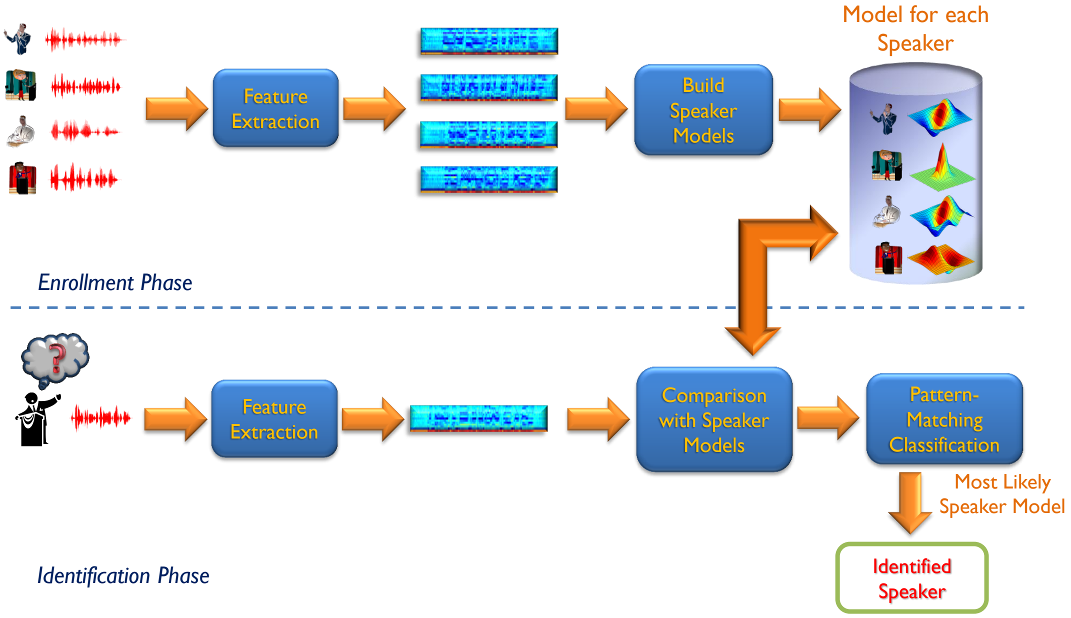


Figure 2.1: Block diagram of a speaker identification system.

The most popular approach for speaker identification is based on Gaussian Mixture Models (GMM) [5]. Other classifiers based on joint factor analysis (JFA) [33] and Support Vector Machines (SVM) [7] have also been used for this task. For a more detailed description can be found in Chapter 1 and especially Section 1.2 focus on recognition systems using a limited amount of speech data which is one of the main goals of the current work.

Recently, the focus of the speaker recognition research community has been given both on the study of features that are more robust in noise environments and on finding more robust and efficient identification algorithms. Specifically, in [6] robust features based on mel-frequency cepstral coefficients (MFCCs [79]) are proposed, in combination with a projection measure technique for speaker identification. In [80], the speech features are based on a harmonic decomposition of the signal where a reliable frame weighting method is adopted for noise compensation. In [10], the descriptors introduced are based on the AM-FM representation of the speech signal, while in [8] the proposed features are derived from auditory filtering and cepstral analysis (in both cases a GMM is used to model the feature space). In [9, 81] the noise robust speaker identification problem under mismatched testing and training conditions is studied. In [9], the identification is performed in the space of adapted GMMs where Bhattacharyya shape is used to measure the closeness of speaker models, while in [81] a multicondition model training and missing feature theory is adopted to deal with the training and testing mismatch, where this model is incorporated into a GMM for noise robust speaker identification.

An important aspect in speaker identification is that in real-time applications the system should be able to respond within a short time duration about the identity of the speaker. However, when the number of the enrolled speakers in the database grows significantly, it is quite difficult for the system to quickly assign the speaker with a specific identity. For addressing such real-time efficiency concerns, in [82] a method based on approximating GMM likelihood scoring with an approximated cross entropy is proposed. In [11], the GMM-based speaker models are clustered using a k-means algorithm so as to select only a small proportion of speaker models used in likelihood computations. These approaches achieve a more efficient operation compared to state-of-the-art, without degrading the identification performance in large population databases.

2.2 State-of-the-art identification methods

In the current section a description of the state-of-the-art methods used to perform speaker identification is given. For the feature extraction task it is assumed that the speech signal/utterance is segmented into overlapping frames, where MFCC features [79] are computed during the feature extraction process.

2.2.1 Gaussian Mixture Model

Gaussian Mixture Models (GMMs) have been applied with great success in the text-independent speaker identification problem [5]. The approach is to model the probability density function (PDF) of the feature space of each speaker in the dataset (training phase) as a sum of Gaussian functions, and then use the maximum a-posteriori rule to identify the speaker. A Gaussian mixture density is a weighted sum of M multidimensional Gaussian densities, where the mixture density can be represented as

$$\lambda_i = \left\{ p_m^i, \mu_m^i, \mathbf{\Sigma}_m^i \right\}, \; m = 1, \ldots, M, \tag{2.1}$$

where for the i^{th} speaker, p_m^i is the weight of the m^{th} mixture (prior probability), $\boldsymbol{\mu}_m^i$ is the corresponding mean vector, $\mathbf{\Sigma}_m^i$ is the covariance matrix, and M is the total number of Gaussian mixtures. Each speaker is represented by a GMM and the corresponding model λ, whose parameters are computed via the Expectation-Maximization (EM) algorithm applied on the training features. For the speaker identification task (testing phase), the estimated speaker identity (speaker index) is obtained based on the maximum a-posteriori probability for a given sequence of observations as follows

$$S_q = \arg \max_{1 \leq i \leq S} p(\lambda_i | \mathcal{V}) = \arg \max_{1 \leq i \leq S} \frac{p(\mathcal{V}|\lambda_i) p(\lambda_i)}{p(\mathcal{V})}. \tag{2.2}$$

In the above equation, $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_N \,|\, \mathbf{v}_i \in \mathcal{F},\ i = 1, \ldots, N\}$ denotes a sequence of N feature vectors, where $\mathcal{F}$ denotes the feature space and S is the total number of speakers. For equally likely speakers and since $p(\mathcal{V})$ is the same for all speaker models the above equation becomes

$$S_q = \arg \max_{1 \leq i \leq S} p(\mathcal{V}|\lambda_i). \tag{2.3}$$

For independent observations and using logarithms, the identification criterion becomes

$$S_q = \arg \max_{1 \leq i \leq S} \sum_{t=1}^{N} \log p(\mathbf{v}_t|\lambda_i), \tag{2.4}$$

where

$$p(\mathbf{v}_t|\lambda_i) = \sum_{m=1}^{M} \frac{p_m^i}{(2\pi)^{d/2}|\mathbf{\Sigma}_m^i|^{1/2}} \exp\Big\{ -\frac{1}{2}(\mathbf{v}_t - \boldsymbol{\mu}_m^i)^T {\mathbf{\Sigma}_m^i}^{-1} (\mathbf{v}_t - \boldsymbol{\mu}_m^i)\Big\}, \tag{2.5}$$

d being the dimension of each feature vector.

Maximum a-posteriori adaptation

An extended version of the GMM model named universal background model-GMM (UBM-GMM) was introduced in [39] in order diminish the drawback that the available speech samples from specific speakers are often not enough to efficiently estimate a GMM model. The core concept of the UBM-GMM technique is based on the fact that once a model has been trained using speaker-independent speech training data, this can be further utilised as a prior when training specific speaker-dependent models. This can be translated in turning the ML estimation process into a maximum a-posteriori adaptation (MAP) one, where the prior is represented by the UBM model. In other words, a UBM model is trained and then an estimation of the speaker GMMs is performed by adaptation of the UBM using the individual speaker data as the adaptation data.

In the case of speaker identification, the use of UBM-GMM is not necessary to be adopted since each speaker's estimated GMM model is sufficient to perform the identification in a typical manner. However, the use of UBM-GMM might be preferable in cases of little or insufficient speech data because it can model more accurate all the feature space across all speakers.

Given the set of feature vectors $\mathcal{V}$ and the UBM model λ_{UBM} the adapted mean new vectors are derived (the index i has been removed for the sake of simplicity in the equations below), as a trade-off between the UBM model means $\boldsymbol{\mu}_m$ and the new data in the form

$$\hat{\boldsymbol{\mu}}_m = \frac{n_m}{n_m + \tau} \bar{\boldsymbol{\mu}}_m + \frac{\tau}{n_m + \tau} \boldsymbol{\mu}_m, \tag{2.6}$$

where $\hat{\boldsymbol{\mu}}_m$ is the adapted mean for the m-th mixture, τ is a weighting MAP parameter controling the importance of training samples and the UBM during the adaptation process. The occupation likelihood of the adaptation data corresponding to each speaker is denoted by n_m, $\boldsymbol{\mu}_m$ is the speaker-independent UBM mean and $\bar{\boldsymbol{\mu}}_m$ is the mean of the observed individual speaker's adaptation data defined as

$$\bar{\boldsymbol{\mu}}_m = \frac{1}{n_m} \sum_t \frac{w_m p_m(\mathbf{v}_t)}{\sum_{m=1}^{M} w_m p_m(\mathbf{v}_t)} \mathbf{v}_t, \tag{2.7}$$

where $p_m(\mathbf{v}_t)$ is a multidimensional Gaussian density as in (2.5).

2.2.2 Joint factor analysis

Joint factor analysis (JFA) modeling is based on estimating the speaker space representing by the eigenvoice matrix and the session space defined by the eigensession matrix. An extension of JFA includes the estimation of only a single space referred to as total variability space which models both the speaker and session variabilities. The largest eigenvalues of the total variability covariance matrix are used to built the total variability matrix which in turn represents the total variability space. The factor analysis model is described as follows

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w}, \tag{2.8}$$

where $\mathbf{M} \in \mathbb{R}^{Md\times 1}$ represents the supervector (defined as the concatenation of the means of the GMMs for each speaker into a high-dimensional and fixed single vector, of dimension $Md\times 1$ with M denoting the number of Gaussian centres and d is the dimension of the features space) of a specific speaker or utterance, $\mathbf{m} \in \mathbb{R}^{Md\times 1}$ corresponds to the speaker-independent and channel-independent supervector of the UBM model, $\mathbf{T} \in \mathbb{R}^{Kd\times D}$ defines the total variability space and $\mathbf{w} \in \mathbb{R}^{D\times 1}$ is a random vector which is assumed to follow a standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The components of $\mathbf{w}$ are the total factors for a given speaker or utterance, also called as i-vectors. The matrix $\mathbf{T}$ is low-rank and its columns span the subspace where most of the speaker-specific information lives, along with channel variability.

After the definition of the total variability space in (2.8), the i-vector training is considered. Now, let us assume that each speaker's utterance corresponds to a sequence of feature vectors $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$ and the total variability space $\mathbf{T}$ is fixed. We want to estimate the maximum probability of a specific speaker denoted by its supervector $\mathbf{M}$ given the utterance.

The Baum-Welch statistics needed to estimate the i-vector for a given speech utterance are obtained by

$$\eta_m = \sum_{n=1}^{N} P(m|\mathbf{v}_n, \lambda_{\text{UBM}}) \tag{2.9}$$

$$\chi_m = \sum_{n=1}^{N} P(m|\mathbf{v}_n, \lambda_{\text{UBM}})\mathbf{v}_n, \tag{2.10}$$

where $m = 1, \ldots, M$ is the Gaussian index and λ_{UBM} denotes the UBM. The posterior probability of the m-th mixture component generating the feature vector $\mathbf{v}_n$ is denoted by $P(m|\mathbf{v}_n, \lambda_{\text{UBM}})$. The centralized first-order Baum-Welch statistics based on the UBM mean mixtures are also needed for i-vector estimation

$$\tilde{\chi}_m = \sum_{n=1}^{N} P(m|\mathbf{v}_n, \lambda_{\text{UBM}})(\mathbf{v}_n - \boldsymbol{\mu}_m), \tag{2.11}$$

where $\boldsymbol{\mu}_m$ is the mean of the m-th UBM mixture component.

The maximum likelihood estimation problem can be written as

$$\max_{\mathbf{M}} p(\mathbf{M}|\mathcal{V}) = \max_{\mathbf{M}} p(\mathcal{V}|\mathbf{M})p(\mathbf{M}) = \min_{\mathbf{M}} \big\{ -\log(p(\mathcal{V}|\mathbf{m} + \mathbf{T}\mathbf{w})) - \log(p(\mathbf{w}))\big\}, \tag{2.12}$$

where $p(\mathbf{w}) \sim \mathcal{N}(\mathbf{0}, \mathcal{I})$. The solution of problem (2.12) is given by the following equation

$$\mathbf{w} = (\boldsymbol{I} + \boldsymbol{T}^T\boldsymbol{\Sigma}^{-1}\eta(\mathcal{V})\boldsymbol{T})^{-1}\boldsymbol{T}^T\boldsymbol{\Sigma}^{-1}\tilde{\chi}(\mathcal{V}), \tag{2.13}$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{Md\times Md}$ is a diagonal covariance matrix modeling the residual variability not captured by the total variability matrix $\boldsymbol{T}$ which is estimated during the factor analysis training [83]. The diagonal matrix $\eta(\mathcal{V}) \in \mathbb{R}^{Md\times Md}$ contains blocks in its main diagonal of the form $\eta_m\boldsymbol{I}$ with $m = 1, \ldots, M$. The supervector $\tilde{\chi}(\mathcal{V}) \in \mathbb{R}^{Md\times 1}$ is obtained by concatenating all the first-order Baum-Welch statistics $\tilde{\chi}_m$ for a given utterance $\mathcal{V}$.

After the total variability space and i-vectors estimation, linear discriminant analysis (LDA) is applied to project the i-vectors into a lower dimensional space as $\mathbf{y} = \boldsymbol{\Psi}\mathbf{w}$. The goal in LDA is the maximization of between-class (or inter-speaker) covariance matrix and the minimization of the within-class (or intra-speaker) covariance matrix. The intra-speaker covariance for S speakers is computed as

$$\boldsymbol{\Sigma}_b = \sum_{s=1}^{S} (\mathbf{w}_s - \bar{\mathbf{w}})(\mathbf{w}_s - \bar{\mathbf{w}})^T \tag{2.14}$$

while the inter-speaker covariance matrix is given by

$$\mathbf{\Sigma}_w = \sum_{s=1}^{S} \frac{1}{u_s} \sum_{t=1}^{u_s} (\mathbf{w}_t^s - \bar{\mathbf{w}}_s)(\mathbf{w}_t^s - \bar{\mathbf{w}}_s)^T, \tag{2.15}$$

where u_s corresponds to the total number of utterances for each the s-th speaker and

$$\bar{\mathbf{w}}_s = \frac{1}{u_s} \sum_{t=1}^{u_s} \mathbf{w}_t^s \tag{2.16}$$

is the mean of i-vectors for each speaker. The speaker population mean $\bar{\mathbf{w}}$ is the mean of the total data set

$$\bar{\mathbf{w}} = \frac{1}{u_{tr}} \sum_{s=1}^{S} \left(\sum_{t=1}^{u_s} \mathbf{w}_t^s \right), \tag{2.17}$$

with $u_{tr} = u_1 + u_2 + \ldots + u_S$ denoting the total number of utterances.

The main goal of LDA is to maximize the between-speaker variation while minimizing the within-speaker variances, by adopting the Fisher criterion. More simply, the purpose of LDA is to maximize the Rayleigh quotient

$$\mathcal{J}(\mathbf{\Psi}) = \frac{\mathbf{\Psi}^T \mathbf{\Sigma}_b \mathbf{\Psi}}{\mathbf{\Psi}^T \mathbf{\Sigma}_w \mathbf{\Psi}}. \tag{2.18}$$

This maximization computes a projection matrix $\mathbf{\Psi}$ composed by the best eigenvectors (those with highest eigenvalues) of the general eigenvalue equation

$$\mathbf{\Sigma}_b \boldsymbol{q} = \boldsymbol{\lambda} \mathbf{\Sigma}_w \boldsymbol{q}, \tag{2.19}$$

where $\boldsymbol{\lambda}$ is a diagonal matrix of eigenvalues. The i-vectors are then submitted to the projection matrix $\mathbf{\Psi}$ obtained from LDA. The dimension of the new subspace $\mathbf{y}$ with $\mathbf{y} = \mathbf{\Psi}\mathbf{w}$, must be less than the number of speakers used during training.

For a speaker identification task, given the i-vector $\mathbf{w}_s$ corresponding to speaker s and the i-vector $\mathbf{w}_t$ of the speaker to be identified, we are interested in testing two hypotheses, i.e., H_1 that both $\mathbf{w}_s$ and $\mathbf{w}_t$ share the same speaker identity or H_0 that the i-vectors were generated from different speakers. The identification score can be computed as the log-likelihood ratio for this hypothesis test as

$$\text{llr score}_s = \log \frac{p(\mathbf{w}_s, \mathbf{w}_t | H_1)}{p(\mathbf{w}_s | H_0) p(\mathbf{w}_t | H_0)}. \tag{2.20}$$

The estimated identity of the speaker is given by the following rule

$$S_q = \arg\max_{1\leq s\leq S} \text{llr score}_s, \tag{2.21}$$

where a more detailed analysis regarding the log-likelihood scoring function can be found in [84].

2.3 Proposed identification methods

In the following, we describe the proposed classification methods for the speaker identification task under noisy conditions and using short training and testing utterances.

2.3.1 Statistical Modeling based on Generalized Gaussian Density

In this subsection, we describe a statistical approach which treats the speaker identification problem as a multiple hypothesis problem. Following the notation of the previous subsection, let us again assume that there are S speakers in total and the set of N independent feature vectors is defined as $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_N \,|\, \mathbf{v}_i \in \mathcal{F},\ i = 1, \ldots, N\}$, where $\mathcal{F}$ denotes the feature space. Each speaker is assigned a hypothesis H_i. The goal is to select one hypothesis out of S best describing the test speaker's data. Under the common assumption of equal prior probabilities of the hypotheses, the optimal rule resulting in the minimum probability of classification error is to select the hypothesis with the highest likelihood among the S. Thus, the correct identity is assigned to the speaker corresponding to the hypothesis H_j if

$$p(\mathcal{V}|H_j) \geq p(\mathcal{V}|H_i),\ i \neq j\,, \forall\ i = 1, ..., S. \tag{2.22}$$

For solving this problem, a parametric approach is adopted where each conditional probability density $p(\mathcal{V}|H_i)$ is modeled by a member of a family of PDFs, denoted by $p(\mathcal{V};\boldsymbol{\theta}_i)$, where $\boldsymbol{\theta}_i$ is a set of model parameters. Under this assumption, the extracted features for the i^{th} speaker are represented by the estimated model parameter $\hat{\boldsymbol{\theta}}_i$, computed in the feature extraction stage. For assigning the correct identity i^* to the closest speaker identity:

1. compute the Kullback-Leibler divergence (KLD) between the density of the speaker to be identified $p(\mathcal{V};\boldsymbol{\theta}_t)$ and the density $p(\mathcal{V};\boldsymbol{\theta}_i)$ associated with the i^{th} speaker identity in the database, $\forall\, i = 1, \ldots, S$

$$D(p(\mathcal{V};\boldsymbol{\theta}_t)\|p(\mathcal{V};\boldsymbol{\theta}_i)) = \int p(\mathbf{v};\boldsymbol{\theta}_t) \log \frac{p(\mathbf{v};\boldsymbol{\theta}_t)}{p(\mathbf{v};\boldsymbol{\theta}_i)}\, d\mathbf{v} \tag{2.23}$$

2. assign i^* to the identity corresponding to the smallest value of the KLD

$$i^* = \arg\min_i D(p(\mathcal{V};\boldsymbol{\theta}_t)\|p(\mathcal{V};\boldsymbol{\theta}_i))\,,\ i = 1,\ldots,S. \tag{2.24}$$

A chain rule holds for the KLD and is applied in order to combine the KLDs from multiple data sets or dataset dimensions. This rule states that the KLD between two joint PDFs, $p(\mathcal{V},\mathcal{W})$ and $q(\mathcal{V},\mathcal{W})$, where $\mathcal{V},\mathcal{W}$ are assumed to be independent data sets, is given by

$$D(p(\mathcal{V},\mathcal{W})\|q(\mathcal{V},\mathcal{W})) = D(p(\mathcal{V})\|q(\mathcal{V})) + D(p(\mathcal{W})\|q(\mathcal{W})). \tag{2.25}$$

The proposed method is based on fitting a Generalized Gaussian density (GGD) on the PDF of the features set. In fact, independence among MFCC vector components is assumed, thus a GGD for each scalar component is estimated. This task can be achieved by estimating the two parameters of the GGD (α, β), which is defined as

$$p(v;\,\alpha,\beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)}e^{-(|v|/\alpha)^{\beta}}, \tag{2.26}$$

where $\Gamma(\cdot)$ is the Gamma function, and the GGD parameters are computed using Maximum Likelihood (ML) estimation. Substitution of (2.26) into (2.23) gives the following closed form for the KLD between two GGDs

$$D(p_{\alpha_1,\beta_1}||p_{\alpha_2,\beta_2}) = \log\Big(\frac{\beta_1\alpha_2\Gamma(1/\beta_2)}{\beta_2\alpha_1\Gamma(1/\beta_1)}\Big) + \Big(\frac{\alpha_1}{\alpha_2}\Big)^{\beta_2}\frac{\Gamma(\frac{\beta_2+1}{\beta_1})}{\Gamma(\frac{1}{\beta_1})} - \frac{1}{\beta_1}. \tag{2.27}$$

In the current work, mel-frequency coefficients are used as feature vectors. Based on the independence assumption (2.25) and the KLD between two GGDs (2.27), the overall mean distance between two feature sets $\mathcal{V}_1$, $\mathcal{V}_2$ is as follows

$$D(\mathcal{V}_1\|\mathcal{V}_2) = \frac{1}{d}\sum_{k=1}^{d} D\Big(p^{\mathcal{V}_1}_{\alpha_k,\beta_k}\|p^{\mathcal{V}_2}_{\alpha_k,\beta_k}\Big), \tag{2.28}$$

where d is the dimension of the feature space $\mathcal{F}$ (i.e., the order of the mel-cepstral coefficients). The PDF $p^{\mathcal{V}_1}_{\alpha_k,\beta_k}$ and $p^{\mathcal{V}_2}_{\alpha_k,\beta_k}$ denote the GGD of the k^{th} mel-frequency coefficient of the feature set $\mathcal{V}_1$ and $\mathcal{V}_2$, respectively.

2.3.2 Sparse Representation Classification

The approach of classification based on sparse representation is described in this subsection. This approach was initially applied in face recognition in [85], and is first applied here for noise robust speaker identification under short test and training sessions.

Let us assume that the n_i training samples corresponding to the feature vectors of the i^{th} speaker are arranged as columns of a matrix

$$\mathbf{V}_i = [\mathbf{v}_{i,1}|\mathbf{v}_{i,2}|\ldots|\mathbf{v}_{i,n_i}] \in \mathbb{R}^{d\times n_i} \tag{2.29}$$

dubbed as dictionary, where the column vector $\mathbf{v}_{i,j}$ denotes the j^{th} d-dimensional feature vector of the i^{th} speaker, and n_i is the number of training feature vectors for the i^{th} speaker. The total number of training feature vectors in our database equals $N_{tr} = n_1 + \ldots + n_S$.

In a speaker identification application, the goal is to infer correctly the identity of an unknown speaker, given a new test sample (feature vector) $\mathbf{x}_t \in \mathbb{R}^{d\times 1}$. In the following, let $\mathbf{x}_t$ be a feature vector, which is extracted from the i^{th} speaker. Then, it can be expressed as a linear combination of the training samples associated with this speaker as follows

$$\mathbf{x}_t = c_{i,1}\mathbf{v}_{i,1} + c_{i,2}\mathbf{v}_{i,2} + \cdots + c_{i,n_i}\mathbf{v}_{i,n_i} = \mathbf{V}_i\,\mathbf{c}_i, \tag{2.30}$$

where $\mathbf{c}_i = \{c_{i,j}\}_{j=1}^{n_i}$ is the vector of coefficients of the representation of $\mathbf{x}_t$ in terms of the columns of $\mathbf{V}_i$.

The overall training data matrix $\mathbf{V}$ is formed by concatenating all the training data matrices $\mathbf{V}_i$, $i = 1, \ldots, S$,

$$\begin{aligned} \mathbf{V} &= [\mathbf{v}_{1,1}|\cdots|\mathbf{v}_{1,n_1}|\mathbf{v}_{2,1}|\cdots|\mathbf{v}_{2,n_2}|\cdots|\mathbf{v}_{S,1}|\cdots|\mathbf{v}_{S,n_S}] \\ &= [\mathbf{V}_1|\mathbf{V}_2|\cdots|\mathbf{V}_S] \in \mathbb{R}^{d\times N_{tr}}\ . \end{aligned} \tag{2.31}$$

By combining (2.30) and (2.31), $\mathbf{x}_t$ can be expressed in terms of the overall training data matrix $\mathbf{V}$, namely, $\mathbf{x}_t = \mathbf{V}\mathbf{c}$, where

$$\mathbf{c} = [0,\,\ldots,\,0,\,c_{i,1},\,c_{i,2},\,\ldots,\,c_{i,n_i},\,0,\,\ldots,\,0] \in \mathbb{R}^{N_{tr}\times 1} \tag{2.32}$$

denotes the coefficients vector, hereafter called the *sparse code*, whose elements are all zero except for those associated with the i^{th} speaker. Notice that, the larger the number of speakers S is, the sparser the sparse code $\mathbf{c}$ will be. This observation motivates us to solve the following optimization problem for a sparse solution

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \|\mathbf{c}\|_0\,,\ \text{s.t.}\ \ \mathbf{x}_t = \mathbf{V}\mathbf{c}, \tag{2.33}$$

where $\|\cdot\|_0$ denotes the ℓ_0 norm, which counts the number of non-zero elements in a vector.

The optimization problem in (2.33) is an NP-hard problem. However, an approximate solution can be obtained if the ℓ_0 norm is substituted by the ℓ_1 norm as follows

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \|\mathbf{c}\|_1 \,, \text{ s.t. } \mathbf{x}_t = \mathbf{V}\mathbf{c}, \tag{2.34}$$

where $\|\cdot\|_1$ denotes the ℓ_1 norm of a vector. The efficient solution of the optimization problem in (2.34) has been studied extensively.

Given the training data matrix $\mathbf{V}$ and the new feature vector (test sample) $\mathbf{x}_t$, the following optimization problem can be practically solved through the orthogonal matching pursuit (OMP) [86] algorithm in order to obtain an estimate of $\mathbf{c}$,

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \|\mathbf{x}_t - \mathbf{V}\mathbf{c}\|_2 \,, \text{ s.t. } \|\mathbf{c}\|_0 = K, \tag{2.35}$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm, $\|\cdot\|_0$ is the ℓ_0 (pseudo)norm, which is defined as the number of non-zero elements of a given vector and K denotes the number of iterations of the algorithm or, equivalently, the number of non-zero elements in $\hat{\mathbf{c}}$.

Algorithm 1: Orthogonal matching pursuit (OMP) algorithm

```
Input: $\mathbf{x}_t$, $\mathbf{V}$, maximum iterations $j_{\max}$, tolerance $\epsilon$
Output: estimated sparse code $\mathbf{c}$
Initialization:
$j = 1$
$\mathbf{r}_j = \mathbf{x}_t$
$\Lambda_j = \emptyset$
while $j \leq j_{\max}$ or $\|\mathbf{r}_j\|_2 \leq \epsilon$ do
    $\mathbf{b} = (\mathbf{V}_{\Lambda_j^c})^T \mathbf{r}_j$
    $I^* = \arg\min_{I} |b_I|$
    $\Lambda_{j+1} = \Lambda_j \cup I^*$
    $\mathbf{c}_{\Lambda_{j+1}} = \mathbf{V}^{\dagger}_{\Lambda_{j+1}} \mathbf{x}_t$
    $\mathbf{r}_{j+1} = \mathbf{x}_t - \mathbf{V}_{\Lambda_{j+1}} \mathbf{c}_{\Lambda_{j+1}}$
    $j = j + 1$
end
```

The OMP algorithm has been proposed within the context of greedy sparse approximation algorithms, where each column selection of $\mathbf{V}$, also known as atom selection, is not changed. However, the residual approximation error update is performed by projecting the current residual onto the subspace spanned by the atoms selected up to a certain iteration. Algorithm 1 implements the OMP sparse estimation process. In the OMP algorithm, the set of active indexes Λ is defined and initialized as the empty set. Inner products are calculated between the residual error and a sub-dictionary whose atom indexes are restricted to be in Λ^c that is the complement of the active set (i.e., only the inner products of unused atoms are evaluated at each iteration). The atom achieving the larger absolute inner product value is selected and then the set Λ is updated to include the chosen index as well as the residual error update is performed by calculating the vector of coefficients $\mathbf{c}_\Lambda$ resulted from the signal's projection onto the subspace spanned by the active atoms. This is achieved by computing the Moore-Penrose pseudo-inverse $\mathbf{V}_\Lambda^\dagger$ of the sub-dictionary $\mathbf{V}_\Lambda^\dagger := (\mathbf{V}_\Lambda^T\mathbf{V}_\Lambda)^{-1}\mathbf{V}_\Lambda^T$ that contains the active atoms.

During OMP all the inner product operations are computed only for the atoms that do not belong to the active set because the residual error at each iteration is orthogonal to the space spanned by the atoms belonging to the active set. This means that, at each iteration, the inner products $\langle \mathbf{r}, \mathbf{v}_k \rangle = 0 \; \forall \; k \in \Lambda$ and the same atom cannot be selected twice. Moreover, if the dictionary is a basis that spans the space $\mathbb{R}^d$, the algorithm converges to a representation with zero residual error after at most d iterations. The advantage of using the OMP algorithm in terms of convergence comes at the expense of computing one pseudo-inverse matrix per iteration

$$\hat{\mathbf{c}}_\Lambda = \arg\min_{\mathbf{c}_\Lambda} \|\mathbf{x}_t - \mathbf{V}\mathbf{c}_\Lambda\|_2. \tag{2.36}$$

In the ideal case, the indices of the non-zero entries of the estimated sparse code $\hat{\mathbf{c}}$ will correspond to those columns of $\mathbf{V}$ associated with the i^{th} speaker, and thus, the test sample $\mathbf{x}_t$ will be assigned correctly to that speaker. However, due to potential modeling errors and/or noise-corrupted data, in practice there may be also several non-zero entries of small amplitude in $\hat{\mathbf{c}}$, which correspond to multiple speakers. To overcome this drawback, we define for each speaker i an indicator function $\delta_i : \mathbb{R}^{N_{tr}} \to \mathbb{R}^{N_{tr}}$ such that the only non-zero entries of vector $\delta_i(\hat{\mathbf{c}}) \in \mathbb{R}^{N_{tr}}$ are from the i^{th} speaker, and this procedure is repeated S times for each speaker. As a result, for a given speaker i we can approximate $\hat{\mathbf{x}}_t^i = \mathbf{V}\delta_i(\hat{\mathbf{c}})$ and assign the test sample to the speaker with the minimum residual between $\mathbf{x}_t$ and $\hat{\mathbf{x}}_t^i$ as

$$i^* = \arg\min_i \|\mathbf{x}_t - \mathbf{V}\delta_i(\hat{\mathbf{c}})\|_2 \,, \; i = 1, \ldots, S. \tag{2.37}$$

This process is performed for each frame of the speech signal of the speaker to be identified, and the final class, that is, the speaker's identity, is estimated by means of a majority voting approach applied on a predefined set of frames. In other words, the unknown speaker is assigned the class to which most of the frames of his/her speech signal are classified in using (2.37).

2.4 Experimental results

In this section, we examine the identification performance of the three methods described in Section 2.2, regarding the correct speaker identification rate. For this purpose, several sim-

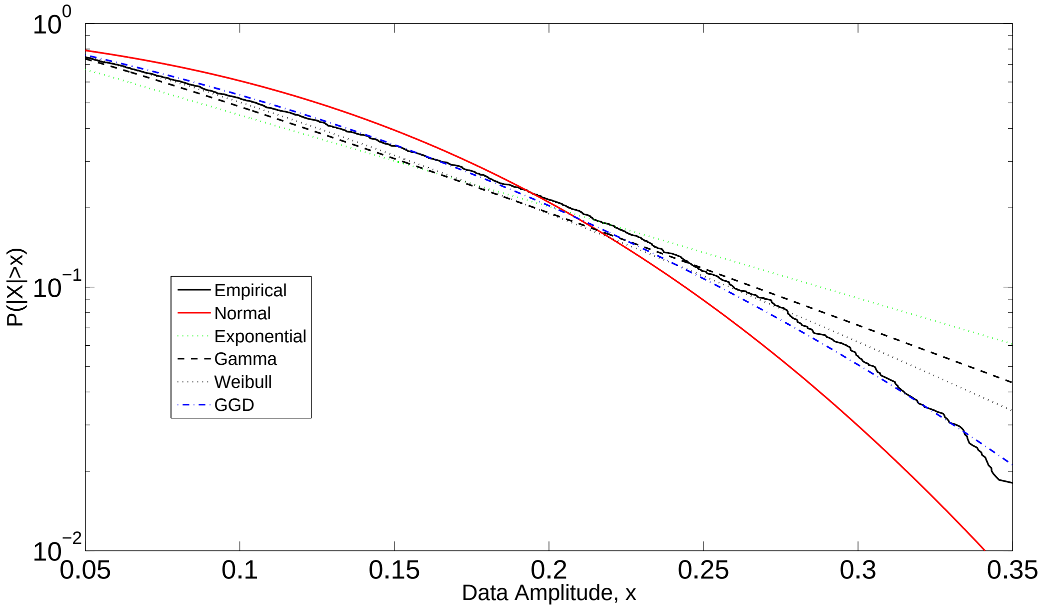


Figure 2.2: Example Amplitude Probability Density curves of the 8-th MFCC coefficient from the training data (20 sec) of the 10-th speaker.

ulations under noisy conditions were conducted. The speech signals used for the simulations were obtained from the VOICES corpus, available by OGI's CSLU [87], which consists of twelve speakers (seven male and five female speakers). For all simulations, 20-dimensional MFCC coefficients were extracted from the speech utterances in a segment-by-segment basis. The frame duration was kept at 20 msec with 10 msec of frame shift. Before the feature extraction task, the training as well as the test utterances were pre-filtered using a low-pass filter of the form $H(z) = 1 - 0.97z^{-1}$, and then a silence detector algorithm based on the short-term energy and zero-crossing measures of speech segments was applied [1]. All the speech signals in the corpus have a sampling rate of 22050 Hz. For the GMM-based identification results, a GMM with a diagonal covariance matrix was chosen for the simulations. The number of mixtures depended on the amount of training data (see description of Experiment 1 below).

For the GGD-based identification case, Amplitude Probability Density (APD) curves ($P(|X| > x)$) are adopted to show that the GGD best matches the actual density of the data. An example for a part of the VOICES corpus is given in Figure 2.2, where we compare the empirical APD (solid line) against the APD curves obtained for the GGD, Weibull, Gamma, Exponential and the Gaussian models. The results in the figure correspond to the 8^{th} MFCC coefficient of the training data (20 sec duration) corresponding to the 10^{th} speaker (independence among feature vector components is assumed). Clearly, the GGD follows more closely the empirical APD than the other densities. This trend was observed in the majority of the training utterances used

[1] http://www.mathworks.com/matlabcentral/fileexchange/19298-speechcore

in our experiments. Thus, the GGD model is expected to give better results than the other densities when applied directly to the MFCC coefficients of the twelve speakers.

The performance evaluation follows the philosophy as described in [5], where each sequence of feature vectors $\{\mathbf{x}_t\}$ is divided into overlapping segments of Q feature vectors, where the segments have the following form

$$\begin{gathered}
\underbrace{\mathbf{x}_1, \mathbf{x}_1, \mathbf{x}_3, \ldots, \mathbf{x}_Q}_{1^{\text{st}}\text{ segment}}\, \mathbf{x}_{Q+1}, \ldots, \mathbf{x}_{P-1}, \mathbf{x}_P \\
\mathbf{x}_1, \underbrace{\mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_Q, \mathbf{x}_{Q+1}}_{2^{\text{nd}}\text{ segment}}, \ldots, \mathbf{x}_{P-1}, \mathbf{x}_P \\
\vdots \\
\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_Q, \mathbf{x}_{Q+1}, \ldots, \mathbf{x}_{P-Q}, \underbrace{\mathbf{x}_{P-Q+1}, \ldots, \mathbf{x}_{P-1}, \mathbf{x}_P}_{\text{P}-\text{Q}+1^{\text{th}}\text{ segment}}
\end{gathered} \tag{2.38}$$

The correct identification rate of the j^{th} speaker is computed as the percentage of the correctly identified segments of length Q over the total number of segments

$$\text{correct ident. rate } (\text{CIR}_j) = \frac{\#\text{ correctly identified segments}}{\text{total }\#\text{ of segments}} \cdot 100\%, \tag{2.39}$$

where in the current work the total number of segments equals $P - Q + 1$. The total mean correct identification rate is used as an evaluation metric during the test simulations defined as

$$\text{mean CIR} = \frac{1}{S} \sum_{j=1}^{S} \text{CIR}_j, \tag{2.40}$$

where S denote the total number of speakers.

In the previous sections, it was mentioned that in the current work the focus is given on noise robust speaker identification using short training and testing sessions. Towards this direction, white Gaussian noise is added on the test utterances, the SNR taking the values of 10, 15, 20, 25 dB. In addition, the test segment lengths Q vary from 10 to 500 with a step size of length 40. Length $Q = 10$ corresponds to 0.1 sec, length $Q = 50$ corresponds to 0.5 sec, and so forth. The training utterances have a duration of 5, 10, 15 and 20 seconds, corresponding to a quite short training session. The training for all methods is performed using the clean speech data.

2.4.1 Experiment 1 – Identification using GMM

In this experiment, during the training process the MFCC coefficients for each speaker are collected. For each speaker, the corresponding MFCC data are modeled using a diagonal GMM. The number of mixtures was chosen to be 4, for the 5 and 10 sec training data, and 8 for the 15 and 20 sec training data. These choices of parameters were found experimentally to produce the best performance for the GMM-based identification. Clearly, the number of mixtures is small due to the small size of the training dataset. During the identification process, the identification rule (2.4) is used, and the correct identification rate is computed as in (2.40).

2.4.2 Experiment 2 – Identification using KLD based on GGD

The same experimental steps as in Experiment 1 are also followed here. Thus, for each speaker the MFCC vectors are collected during the training process. We estimate the GGD parameters (α, β) for each vector component, assuming independence among the MFCC components. During the identification process, a test utterance contains multiple MFCC vectors as explained. For each MFCC component of the test vectors, the GGD parameters (α, β) are estimated. In order to identify a speaker, we compute the KLD between the GGD model of the test data and each of the GGD models of the speakers in the dataset (per vector component). This procedure results in 20 distance values (since each MFCC vector contains 20 components). The final step is to compute the mean of these distances, as in (2.28). The identity of the speaker whose data result in the minimum distance is identified as the final result. The correct identification rate is computed as in (2.40).

2.4.3 Experiment 3 – Identification using SRC

In this subsection, the experimental procedure for the SRC approach is described. First, consider that from the training speech data of each speaker a number of n_i of MFCC vectors is extracted. Consider a test utterance length of Q frames. Adopting the notations from the theory of SRC in Section 2.3.2, the training matrix $\mathbf{V}$ has dimension $20 \times (12 \cdot n_i)$ and the test sample (feature) vector $\mathbf{x}_t$ is a 20×1 vector. The test segment consists of Q distinct test samples $\mathbf{x}_t$. Thus, the optimization problem of the form

$$(P_q): \ \hat{\mathbf{c}_q} = \arg\min_{\mathbf{c}_q} \|\mathbf{c}_q\|_1\,, \ \text{s.t.} \ \mathbf{x}_{t,q} = \mathbf{V}\mathbf{c}_q, \ \text{for} \ q = 1, \dots, Q \tag{2.41}$$

is solved Q times for each different $\mathbf{x}_{t,q}$. The Orthogonal Matching Pursuit [86] is used to solve this problem. Each solution $\hat{\mathbf{c}_q}$ of the problem (P_q) is used to get an identity i (for $i = 1, \dots, 12$) of one of the 12 speakers in the dataset. Thus, a segment of length Q vectors will provide Q identification results. The predominant identity is found based on the majority of the decisions and the identification rate is computed as in (2.40).

2.4.4 Discussion

In this subsection, the main observations of the results in Figures (2.3.a)-(2.3.d) are discussed. The percentage of correct identification results is given as a function of the length of the test utterance. We are mainly interested to examine the performance of the described methods for short test sessions. The four figures correspond to training data of duration 5, 10, 15, and 20 sec respectively, so as to examine the effect of using a short training dataset. The correct identification rates as a function of the test utterances segment length L are depicted. The black, red and green curves correspond to the SRC, GMM and KLD-GGD method, respectively. There are twelve curves in total, where the first part of each legend name indicates the corresponding method and the last part indicates the SNR value used for this method, *e.g.* "SRC 10dB" means that the black solid curve depicts the identification performance of the SRC approach under noise conditions of 10dB. From the Figures (2.3.a)-(2.3.d) we notice that the SRC method is superior over the GMM and KLD-GGD approach, especially for short test and training sessions, and is quite robust to noise. The GMM performance improves as the training and test data duration increases because the large amount of feature vectors increases the accuracy of the GMM model, however its sensitivity to noise is clearly indicated. The KLD-GGD approach does not have high correct identification rates even in the case where the amount of training and test data is 20 and 5 sec, respectively. Based on the results, we can assume that the GGD parameters (α, β) are not well-estimated in the case where the test data have short duration.

The main point regarding the SRC method that has to be highlighted is that even in the case where the training data duration is 5 sec and the test utterance segments length is as low as 2 sec, the performance is greater than 80% for SNR values 15, 20 and 25 dB. Even in the extreme case of 10 dB SNR, the correct identification rate is above 70% for at least 2 sec test utterance segments length. Additionally, for lower test sessions than 2 sec the identification results for SRC are significantly better than the baseline method. For example, for 20 sec training data and 1.5 sec of test data, the SRC method gives correct identification above 70% for all SNR values. For the same case, for 10 dB SNR, GMM results in correct identification of slightly more than only 20%. This is important for applications where a decision must be made using a small amount of test data, without having enough training data for a given number of speakers, and the speaker is located in a noisy environment.

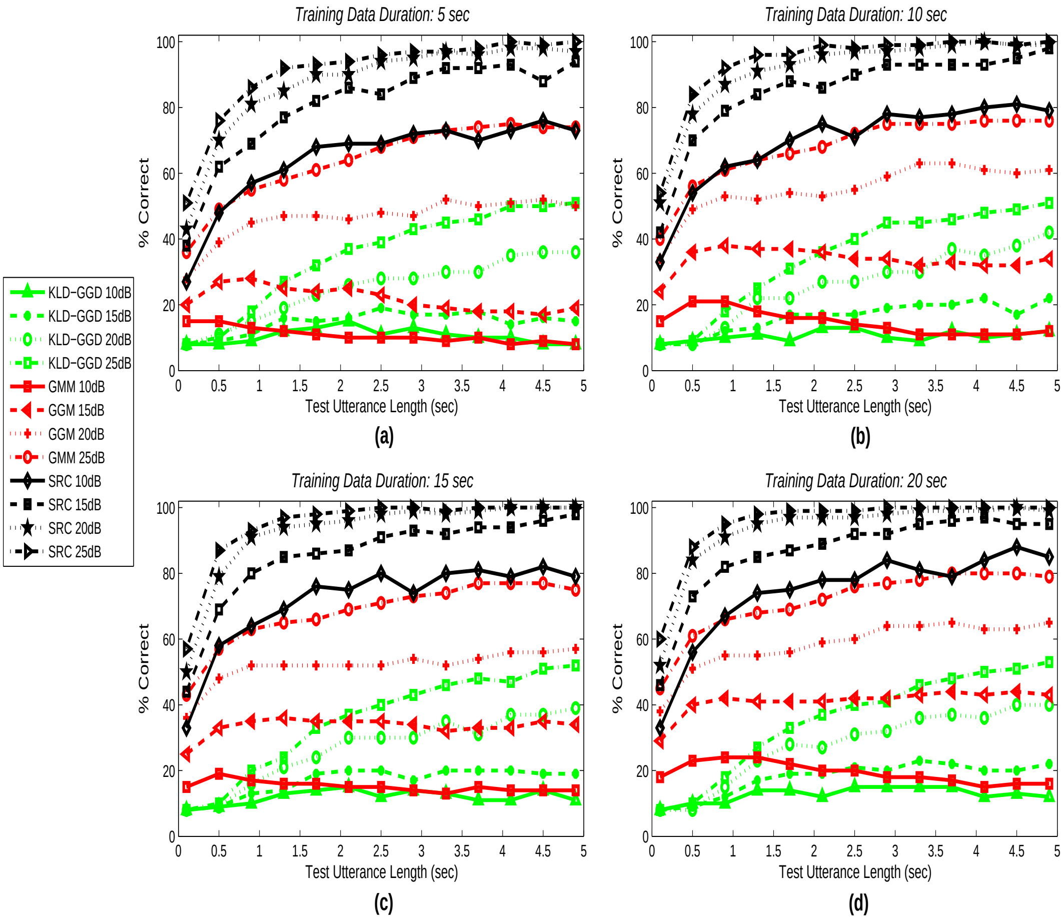


Figure 2.3: Speaker identification performance as a function of the test data duration for different number of SNR values. The duration of the training data is: (a) 5 sec, (b) 10 sec, (c) 15 sec and (d) 20 sec.

2.5 Experimental results: a multicondition perspective

In practical applications speech signals are contaminated with noise due to either the noisy environment in which the speaker is present (e.g., car, restaurant) or the voice transmission medium (*e.g.,* cell-phone, voice over IP communication). To deal with such problems and achieve accurate identification, multicondition GMMs have been proposed (e.g., [81]).

The idea behind the multicondition GMM is to enlarge the training set by corrupting the clean speech training data with simulated noise with different characteristics. As a result, the training set is increased to contain $Z+1$ different subsets $T_0, T_1, \ldots, T_Z$, i.e., clean data T_0 plus noisy data $T_1, \ldots, T_Z$ at Z different noisy conditions. To estimate the correct speaker during the identification process, a GMM of the typical form as described in Section 2.2.1 is applied on the augmented training set $T = T_0, \ldots, T_Z$ and the maximum a posteriori probability rule (2.4) is then used to estimate the identity of the speaker.

In this section, we examine the identification performance of the SRC compared with a multicondition GMM and a baseline GMM, regarding the correct speaker identification rate. For this purpose, several simulations under noisy conditions were conducted. The speech signals used for the simulations were obtained from the VOICES corpus, available by OGI's CSLU [87], which consists of twelve speakers (seven male and five female speakers). The speech signals, originally sampled at 22 kHz, were downsampled to 8kHz, with $N = 320$ samples per frame and 50% overlapping between frames. For all simulations, 22-dimensional LSF coefficients were extracted from the speech utterances in a segment-by-segment basis. For the GMM-based identification results, a GMM with a diagonal covariance matrix was chosen for the simulations. The number of mixtures depended on the amount of training data.

The performance evaluation follows the philosophy as described in Section 2.4 (see 2.38 and 2.40). In the following two subsections we describe the simulations conducted to examine the correct identification rates of the two proposed approaches.

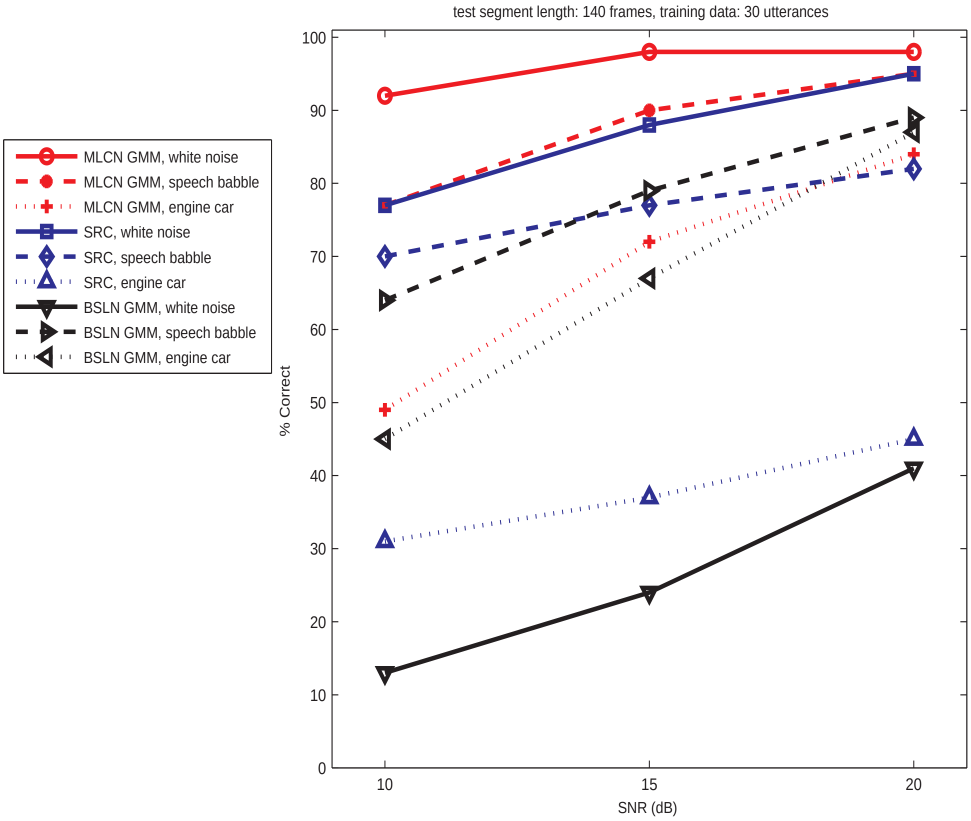


Figure 2.4: Speaker identification performance as a function of the noise SNR. The duration of the training data is 30 utterances (per speaker). Correct identification rates evaluated using three different types of noise: white, speech babble, car engine. The test segment length is 140 frames.

2.5.1 Speaker identification based on SRC

In the current proposed approach the focus is on noise robust speaker identification using short training and testing sessions. To explore this, three different types of noise are added to the test utterances: white noise, speech babble noise and car engine noise. The noise signals were taken from the NOISEX-92 database [88]. The SNR of the corrupted speech takes the values of 10, 15, 20 dB. In addition, the test segment lengths Q is chosen to be 100, 200, 300. Length $Q = 100$ corresponds to 2 sec, length $Q = 200$ corresponds to 4 sec, and length $Q = 300$ corresponds to 6 sec. The training utterances have a duration of 5, 10, 15 and 20 seconds, corresponding to a quite short training session. The testing data (over which the identification results per segment are averaged) have a duration of approximately 20 sec. The experimental results of the current section can be categorized as follows:

1. *baseline GMM*: train clean speech data only, where the number of mixtures was chosen to be 4, for the 5 and 10 sec training data, and 8 for the 15 and 20 sec training data. These choices of parameters were found experimentally to produce the best performance for the GMM-based identification. Clearly, the number of mixtures is small due to the small size of the training dataset. During the identification process, the identification rule (2.4) is used, and the correct identification rate is computed as in (4.21).

2. *multicondition GMM*: train clean plus noisy speech data (clean data are corrupted during training by *white* noise of SNR 10, 15 and 20 dB), where the number of mixtures was experimentally chosen to be 8, for the 5 and 10 sec training data, and 16 for the 15 and 20 sec training data.

3. *SRC*: consider that from the training speech data of each speaker a number of n_i of LSF vectors are extracted. Consider a test utterance length of Q frames. Adopting the notations from the theory of SRC in Section 2.3.2, the training matrix $\mathbf{V}$ has dimension $22 \times (12 \cdot n_i)$ (each matrix $\mathbf{V}_i$ contains clean plus noisy speech data, corrupted by white noise of SNR 10, 15 and 20 dB) and the test sample (feature) vector $\mathbf{x}_t$ is a 22×1 vector. The test segment consists of Q distinct test samples $\mathbf{x}_t$. Thus, the optimization problem of the form

$$\begin{aligned} (P_q):\ \hat{\mathbf{c}}_q &= \arg\min_{\mathbf{c}_q} \|\mathbf{c}_q\|_1 \\ \text{s.t. } \mathbf{x}_{t,q} &= \mathbf{V}\mathbf{c}_q, \ \text{ for } q = 1, \ldots, Q \end{aligned} \tag{2.42}$$

is solved Q times for each different $\mathbf{x}_{t,q}$. Orthogonal Matching Pursuit [86] is used to solve this problem. Each solution $\hat{\mathbf{c}}_q$ of the problem (P_q) is used to get an identity i (for

$i = 1, \ldots, 12$) of one of the 12 speakers in the dataset. Thus, a segment of length Q vectors will provide Q identification results. The predominant identity is found based on the majority of the decisions and the identification rate is computed as in (2.40).

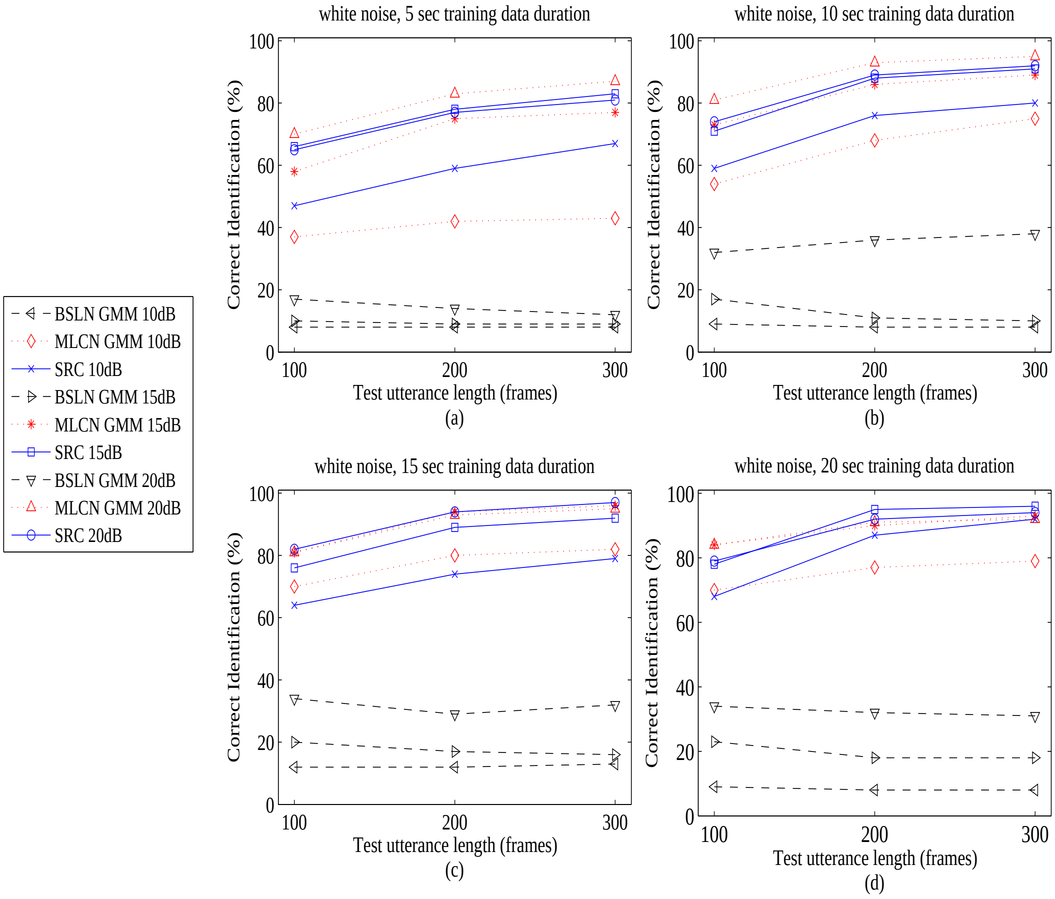


Figure 2.5: Speaker identification performance as a function of the test data duration for different number of white noise SNR values. The duration of the training data is: (a) 5 sec, (b) 10 sec, (c) 15 sec and (d) 20 sec.

Before analyzing the experimental results of the proposed method in terms of short training and testing sessions, we are interested to test the SRC approach using a larger dataset of training and testing vectors. These results depicted in Fig. 2.4. The correct identification rates as a function of the three types of noise SNR are depicted. The test utterance length is 140 frames, where 30 utterances per speaker were used for training (30 utterances gives about 150 sec amount of training data). The black, red and blue curves correspond to the baseline GMM (noted as “BSLN GMM”), multicondition GMM (noted as “MLCN GMM”) and the SRC method, respectively. In the BSLN GMM method, 32 mixtures per speaker were trained (diagonal covariance matrix),while in the BLSN GMM approach 64 diagonal mixtures per speaker (clean and noisy speech data) were trained. For this particular case, for the SRC method the $\mathbf{V}$ matrix is formed using the GMM centers taken from the MLCN GMM training process, so as to provide comparable performance results. In the remaining results, this matrix in fact contains the actual speech feature vectors. It can be seen in Fig. 2.4 that the SRC

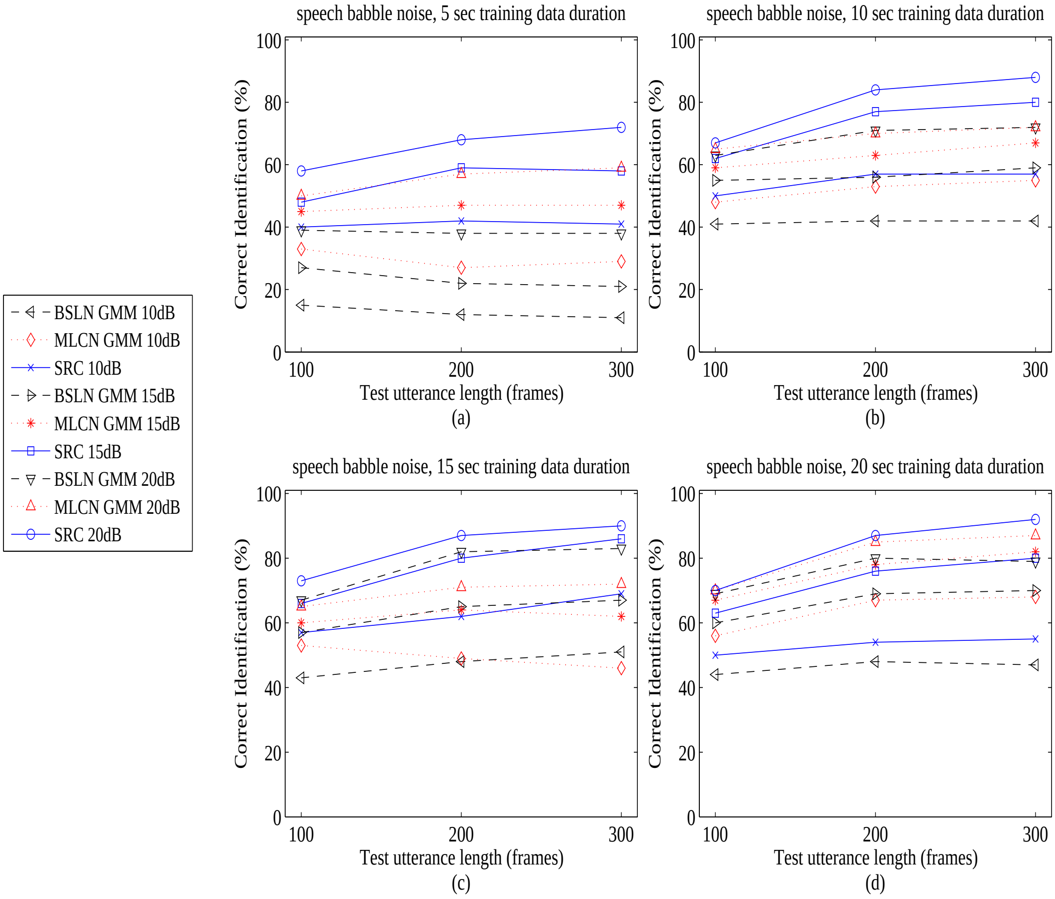


Figure 2.6: Speaker identification performance as a function of the test data duration for different number of speech babble noise SNR values. The duration of the training data is: (a) 5 sec, (b) 10 sec, (c) 15 sec and (d) 20 sec.

method, in which the performs worse than the BSLN GMM and MLCN GMM for the white and speech babble noise. The MLCN GMM is superior than the other two methods, which can be expected since it has been shown to provide very good results when using large training and testing sessions.

The performance evaluation results in terms of correct identification rates for short training and testing sessions corresponding to the white noise, speech babble and car engine noise are depicted in Figs. 2.5-2.7, respectively, where the identification results are given as a function of the length of the test utterance. We are mainly interested in examining the performance of the described SRC-based method for short test sessions. In each figure, the four subfigures correspond to training data of duration 5, 10, 15, and 20 sec respectively, so as to examine the effect of using a short training dataset. The correct identification rates as a function of the test utterances of segment length Q are depicted.

There are nine curves in total in each subfigure, where the first part of each legend name indicates the corresponding method and the last part indicates the SNR value used for this method, *e.g.*, “SRC 15dB” means that the blue solid curve depicts the identification performance of the SRC approach under noise conditions of 15 dB.

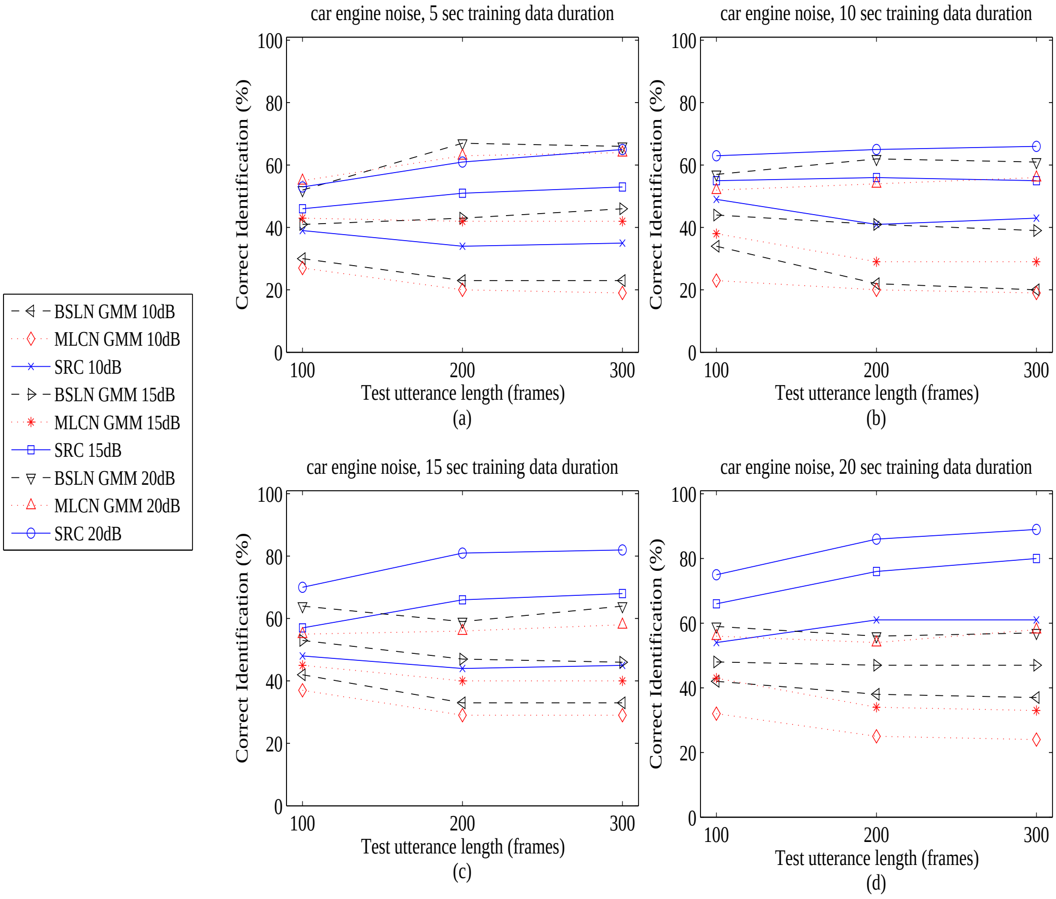


Figure 2.7: Speaker identification performance as a function of the test data duration for different number of car engine noise SNR values. The duration of the training data is: (a) 5 sec, (b) 10 sec, (c) 15 sec and (d) 20 sec.

From the Figs. 2.5-2.7 we notice that the SRC method appears to be quite robust to different noise conditions and it is superior to the BSLN GMM and MLCN GMM approaches in all training cases as well as in all noisy conditions, except for the case of 5 and 10 sec training data duration of 20 dB white noise where SRC appears to be slightly worse than the MLCN GMM method.

In Fig. 2.5 (white noise) the MLCN GMM approach appears to be better than the BSLN GMM because it is trained using clean plus noisy speech data contaminated with white noise and as a result it appropriately captures the characteristics of such a wideband noise during the GMM parameters estimation process. However, the performance of BSLN GMM seems to improve in the case of speech babble and car engine noise and achieves identification performance approximately similar to the MLCN GMM. This observation is more obvious in Fig. 2.7.b-2.7.d, where the performance of the BSLN GMM is slightly better in all cases than the MLCN GMM, because the small amount of training data is not enough for the MLCN GMM to capture the statistical properties (via the GMM parameters estimation) of noise with different spectral characteristics compared to the white noise. The main point regarding the SRC method that has to be highlighted is that the performance is approximately greater than 80% for SNR values 15 and 20 dB in the case where the training data duration is 20 sec and the test utterance segments length is 200 frames.

2.6 Experimental results: beyond state-of-the-art

In this section, we perform an extra set of simulations in order to verify the effectiveness of the proposed SRC-based speaker identification approach against the state-of-the-art methods of UBM-GMM and JFA. We used the VOICES corpus as in the previous section. The original signals are sampled at 22 kHz, and downsampled to 16 kHz. During the feature extraction step, an analysis window of 640 samples, i.e., 40 ms at 16000 samples per second with 50% overlapping between two consecutive frames, is employed to compute a mel-frequency spectrogram of $\Omega = 30$ bands, where a silence detector algorithm based on the short-term energy and zero-crossings measure of speech segments is applied[2]. A cepstral mean and variance normalization process

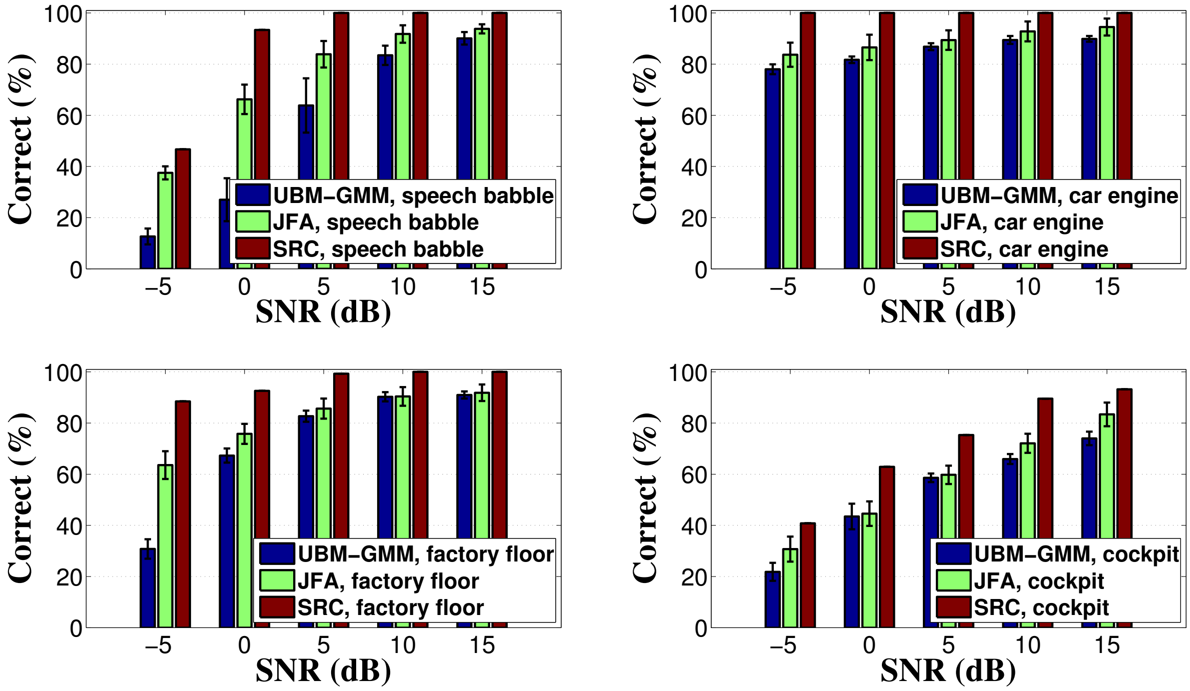


Figure 2.8: Correct identification rates as a function of the SNR. The proposed SRC is compared against the state-of-the-art UBM-GMM and JFA methods for four noise types.

[2] http://www.mathworks.com/matlabcentral/fileexchange/19298-speechcore

followed by feature warping is also applied during the training and testing feature extraction process.

The resulting $\Omega \times T$ mel-spectrogram, where T is the total number of frames on which mel-frequency analysis was performed, is of size 30×600 with $T = 600$ corresponding to approximately 12.02 sec uttered training data per speaker. For the UBM-GMM framework a diagonal covariance matrix was chosen during the simulations. We pooled all the target speakers training data using the mel-scale frequency coefficients of order $\Omega = 30$, where after experimentation we found that best results on average obtained when used 16 number of mixtures. The dimension of the total variability space was set to 12, which equals the number of speakers of VOICES corpus.

The average identification error rate is computed as the percentage of the erroneously identified segments over the total number of test segments. For each speaker, the total number of test utterances used for the evaluation is equal to 4, where the segment length is set to 400 frames (corresponding around to 8.02 sec). The test utterances are corrupted by four different types of additive noise, namely, speech babble noise, car engine noise, factory floor noise and F-16 cockpit noise, where the SNR of the corrupted speech takes the values of -5, 0, 5, 10 and 15 dB. The noise signals were taken from the NOISEX-92 database [88].

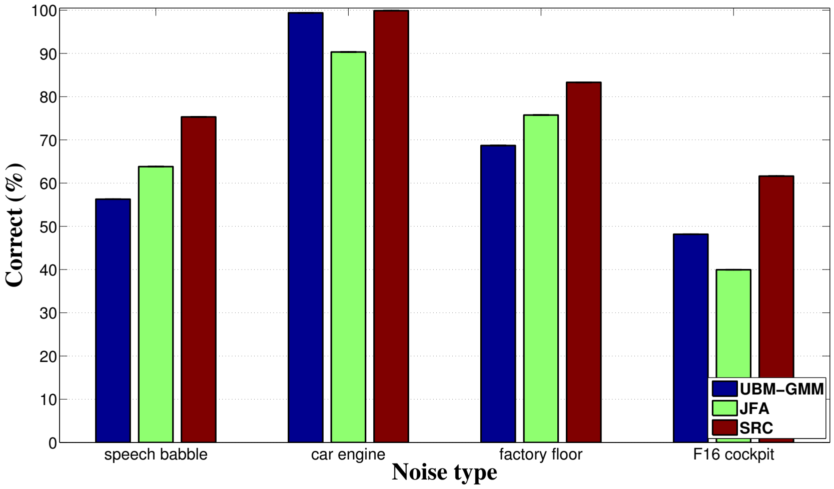


Figure 2.9: Average correct identification rates across all noise types comparing SRC vs. JFA vs. UBM-GMM.

In Fig. 2.9 are depicted the average correct identification rates for all the compared methods across all noise types. It is easy to verify that SRC is better than JFA and UBM-GMM in all noisy conditions. In specific, the achieved SRC correct rates are greater than 76% in the case of speech babble, car engine and factory floor noise types. Additionally, JFA is better than UBM-GMM by approximately 7.6% for the speech babble noise and about 7% for the factory floor noise. Contrary, UBM-GMM appears to better as compared to JFA in the case of car engine noise by 8.9%, while for the F16 cockpit noise UBM-GMM achieves an average correct identification error about 8.2% more than JFA.

Chapter 3

Discriminative dictionary sparse coding for speaker identification

It is all right to make mistakes; nothing is perfect because with perfection, we would not exist.

Stephen Hawking (1942-)

3.1 Introduction

Following the goal of achieving robust identification results under limited amount of training and testing speech data, the focus is given on enhancing the discriminative ability of the estimated sparse codes. Towards this direction, a discriminative learning approach is introduced. The problem is faced under a joint learning perspective, where an overcomplete dictionary is learned, resulting in highly discriminative sparse codes, along with a linear classifier. A speech corpus of twelve speakers is used for the identification evaluation towards the direction of examining applications consisting of a moderate number of speakers

3.2 Prior work on sparsity based classification for speech signals

The concept of sparse representation (or sparse coding) comes as an alternative solution to the universal data models, which do not generalize well for limited training data. Prior work on classification of speech signals has been already described in Chapter 1. The main focus is given on representing an input test sample as a sparse linear combination of an overcomplete matrix, the so-called *dictionary*, whose columns consist of a set of basis functions, usually referred to as atoms. Next, we will mention prior work on classification of speech signals based in this kind of assumptions some of them already mentioned in Section 1.2.

In [89], robust speech recognition is achieved by modeling noisy speech signals as a sparse linear combination of speech and noise *exemplars* (spectro-temporal representations spanning multiple time-frames of the speech signal). A similar approach is followed in [90], where a combination of large vocabulary continuous speech recognition techniques with small vocabulary tasks results in low phonetic error rates. Sparse codes may also serve as a new type of feature vectors to be given as input in a typical classifier. More specifically, a gradient descent-based dictionary learning approach is adopted in [91] to learn the redundant matrix related with the training data. This comes in combination with a multilayer perceptron classifier, which is applied on the generated sparse codes for phoneme recognition. The same task is also studied in [92]. An orthogonal matching pursuit-based (OMP) dictionary learning technique is applied and the obtained sparse codes are further used for classification by means of a support vector machine (SVM) classifier. A phone recognition approach employing hidden Markov models (HMM) is examined in [93], using sparse codes which take advantage of the phonetic labels information as additional features during the recognition process. Moreover, the sparse codes feature extraction is followed by sparse discriminant analysis to perform speaker recognition in [76], while in [94] SRC is used for the same task using GMM mean supervectors as feature vectors on clean speech data taken from TIMIT speech corpus.

Dictionary learning techniques can be applied for learning the best dictionary that gives the most discriminative sparse codes for classification. The work in [95] showed that a satisfactory speaker verification performance can be achieved by applying a supervised K-SVD algorithm for learning an appropriate discriminative dictionary. Motivated by the successful application of K-SVD for face and object categorization [96], our proposed method addresses the problem of text-independent speaker identification by extending our previous work [97]. Here, we adopt a discriminative dictionary learning approach, which is applied on noise robust speaker identification under the assumption of short training speech utterances.

Here, the proposed method learns an overcomplete dictionary, resulting in highly discriminative sparse codes, along with a linear classifier. This estimation is performed in a joint fashion by imposing additional constraints on the associated objective function in order to produce similar sparse codes for those training samples belonging to the same speaker. This is in contrast to recently introduced sparsity-based methods [89, 90, 91, 92, 93, 76, 94], which do not treat jointly the estimation of the dictionary, the sparse codes, and the classifier parameters. On the other hand, in [95], a method was suggested to learn jointly only the dictionary and the sparse codes. To the best of our knowledge, this is the first study on noise robust speaker identification, which tackles the problem from such a threefold joint learning perspective.

3.3 Reconstructive dictionary sparse coding

Before proceeding with the description of the discriminative sparse coding technique, let us first mention the main points related with the reconstructive dictionary sparse coding task.

A reconstructive dictionary learning method aims at learning an overcomplete dictionary for sparse coding approximation. Following the notation of Section 2.3.2, let

$$\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^{N_{tr}} \in \mathbb{R}^{d\times 1} \tag{3.1}$$

be a set of N_{tr} input signals building the corresponding data matrix

$$\mathbf{V} = [\mathbf{v}_1|\mathbf{v}_2|\cdots|\mathbf{v}_{N_{tr}}]. \tag{3.2}$$

The goal of dictionary sparse coding (or dictionary learning) is to find a dictionary $\mathbf{D} \in \mathbb{R}^{d\times Z}$ with a fixed number of Z columns or atoms such that

$$\mathbf{V} \approx \mathbf{DC}, \tag{3.3}$$

where the columns of matrix $\mathbf{C} \in \mathbb{R}^{Z\times N_{tr}}$ contain the sparse representation coefficients. In other words, each input signal $\mathbf{v}_i$ can be sparsely represented on the estimated dictionary $\mathbf{D}$ and it is associated with a sparse representation vector $\mathbf{c}_i$ (i.e., it contains a small number of nonzero coefficients).

The estimation of dictionary $\mathbf{D}$ and sparse representation matrix $\mathbf{C}$ can be formalized in a similar way as in (2.35) as follows

$$\begin{aligned}\hat{\mathbf{D}}, \hat{\mathbf{C}} = \arg\min_{\mathbf{D},\mathbf{C}} \|\mathbf{V} - \mathbf{DC}\|_F^2 \,, \\ \text{s.t. } \|\mathbf{c}_j\|_0 = K \,,\ \forall j = 1, \ldots, N_{tr} \ ,\end{aligned} \tag{3.4}$$

where the objective function is the Frobenius norm $\|\cdot\|_F$ of the residual error and the sparsity of the representation coefficients is enforced in the approximation of every input signal. The most popular strategy in handling the optimization problem (3.4) is to solve alternatively, starting from an initial guess $\mathbf{D}_0$ of the dictionary and solving the two following steps iteratively

> *Sparse coding update*: during the t-th iteration given a fixed dictionary $\mathbf{D}_t$ the matrix of sparse representation coefficients $\mathbf{C}_t$ can be estimated as a typical sparse coding problem using any solver that is suitable to the particular K-sparse approximation problem.
>
> *Dictionary update*: during the $(t+1)$-th iteration given a fixed matrix of sparse representation coefficients $\mathbf{C}_t$, the dictionary $\mathbf{D}_{t+1}$ is updated in order to improve the objective of the dictionary learning optimization, where each dictionary atom is normalized to have unit norm.

The K-SVD algorithm [98] can be adopted to solve the problem (3.4) following the iterative strategy described above. In specific, we assume that the objective function $\mathcal{J}(\mathbf{D},\mathbf{C}) = \|\mathbf{V}-\mathbf{D}\mathbf{C}\|_F$ can be written as the sum of rank-1 matrices as follows

$$\mathcal{J}(\mathbf{D},\mathbf{C}) = \|\mathbf{V}-\mathbf{D}\mathbf{C}\|_F = \left\|\mathbf{V} - \sum_{p=1}^{Z} \mathbf{d}_p\mathbf{c}_p^T\right\|_F = \left\|\left(\mathbf{V} - \sum_{q\neq p} \mathbf{d}_q\mathbf{c}_q^T\right) - \mathbf{d}_p\mathbf{c}_p^T\right\|_F . \tag{3.5}$$

Now, we can observe that the sparse representation coefficients' vector $\mathbf{c}_p^T$ and the atom $\mathbf{d}_p$ can be optimized in a joint fashion by minimizing the cost function (3.5), i.e., computing the best rank-1 approximation of the partial residual matrix

$$\mathbf{E}_p = \mathbf{V} - \sum_{q\neq p} \mathbf{d}_q\mathbf{c}_q^T. \tag{3.6}$$

The partial residual matrix $\mathbf{E}_p$ and its rank-1 approximation are restricted to the columns

```
Algorithm 2: K-SVD reconstructive dictionary learning
Input: $\mathbf{V}$, $\mathbf{D}_0$, maximum iterations $j_{\max}$, sparsity threshold $K$
Output: estimated dictionary $\mathbf{D}$ and sparse representation matrix $\mathbf{C}$
Initialization:
while $j \leq j_{\max}$ do
    for $n = 1$ to $N_{tr}$ do
        $\hat{\mathbf{c}}_n = \arg\min_{\mathbf{c}_n} \|\mathbf{v}_n - \mathbf{D}_j\mathbf{c}_n\|_2$ , s.t. $\|\mathbf{c}_n\|_0 = K$,
    end
    for $p = 1$ to $Z$ do
        $\Lambda_p = j \subseteq \{1,\ldots,N_{tr}\}$ when $c_{p,j} \neq 0$ (for each atom $\mathbf{d}_p$ the set $\Lambda_p$ of zero
        elements of the $p$-th row of $\mathbf{C}$, i.e., the set of training data that use the $p$-th atom
        in their sparse approximation)
        $\mathbf{E}_p = \left(\mathbf{V} - \sum_{m\neq p} \mathbf{d}_m\mathbf{c}_p^T\right)_{\Lambda_p}$ (calculate a partial residual matrix and restrict its
        columns to the active set of signals that use the $p$-th atom for their sparse
        approximation)
        $[\boldsymbol{A},\boldsymbol{\Sigma},\boldsymbol{\Gamma}] = \texttt{SVD}(\mathbf{E}_p)$ (descending order of the singular values $\{\sigma\}$)
        $\mathbf{d}_p = \mathbf{a}_1$
        $\mathbf{c}_{\Lambda_p} = \sigma_{1,1}\boldsymbol{\gamma}_1^T$ (the representation coefficients $(\mathbf{c}_p^T)_{\Lambda_p}$ and the atoms $\mathbf{d}_p$ are
        updated using the best rank-1 approximation of the partial residual matrix $\mathbf{E}_p$
        which is computed using SVD decomposition)
        $\mathbf{d}_p = \frac{\mathbf{d}_p}{\|\mathbf{d}_p\|}$
    end
    $j = j + 1$
end
```

In addition, the support of the sparse representation coefficients must not be changed during the dictionary update step and thus, the partial residual matrix $\mathbf{E}_p$ and its corresponding rank-1 approximation of are restricted to the columns corresponding to the signals that use the p-th atom in their sparse approximation, i.e., the indexes corresponding to the non-zero elements of the vector $\mathbf{c}_p$.

3.4 Discriminative dictionary sparse coding based on K-SVD

In the previous section, the dictionary learning problem is introduced within the framework of minimizing the reconstruction error. Moving a step further, we are interested in enhancing the discriminativeness of the estimated sparse representation coefficients (or sparse codes). We aim at solving a dictionary sparse coding optimization problem which will incorporate extra optimization terms associated with the discriminative constraints. Here, a method of discriminative dictionary sparse coding based on a (class) label-consistent K-SVD is analyzed, which constitutes the key component of the proposed approach. This method, which was introduced in the framework of face and object recognition [96] and to our knowledge is now applied for a first time in the field of speaker identification. We apply the method in the context of *noisy* conditions using *small training* data sessions.

The sparse coding optimization problem expressed by (2.35) can be extended to the following *dictionary learning* optimization problem:

$$\begin{aligned} \hat{\mathbf{D}}, \hat{\mathbf{C}} &= \arg\min_{\mathbf{D},\mathbf{C}} \|\mathbf{V} - \mathbf{D}\mathbf{C}\|_F^2 \\ &\text{s.t. } \|\mathbf{c}_j\|_0 = K\,,\ \forall j = 1, \ldots, N_{tr}\ , \end{aligned} \tag{3.7}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix, $\mathbf{D} \in \mathbb{R}^{d\times Z}$ is the learned dictionary, $\mathbf{C} \in \mathbb{R}^{Z\times N_{tr}}$ is the matrix of sparse codes, where $\mathbf{c}_j$ denotes the j^{th} column of $\mathbf{C}$, and Z is the dictionary size. We emphasize at this point that the sparse codes $\{\mathbf{c}_j\}_{j=1}^{N_{tr}} \in \mathbb{R}^{Z\times 1}$ are of different dimensionality compared with the sparse code vectors introduced in the first part of the current section. However, the same symbol is used for notational convenience.

In order to enhance the discrimininative capability of the estimated sparse codes, an additional constraint is embedded in the objective function (3.7) as follows,

$$\begin{aligned} \hat{\mathbf{D}}, \hat{\mathbf{C}}, \hat{\mathbf{M}} &= \arg\min_{\mathbf{D},\mathbf{C},\mathbf{M}} \|\mathbf{V} - \mathbf{D}\mathbf{C}\|_F^2 + \lambda_1\|\mathbf{P} - \mathbf{M}\mathbf{C}\|_F^2 \\ &\text{s.t. } \|\mathbf{c}_j\|_0 = K\,,\ \forall j = 1, \ldots, N_{tr}\ , \end{aligned} \tag{3.8}$$

where λ_1 is a regularization parameter controlling the trade-off between the reconstruction error $\|\mathbf{V} - \mathbf{D}\mathbf{C}\|_F^2$ and the discriminative sparse-code error $\|\mathbf{P} - \mathbf{M}\mathbf{C}\|_F^2$. The columns of $\mathbf{P} = [\mathbf{p}_1|\cdots|\mathbf{p}_{N_{tr}}] \in \mathbb{R}^{Z\times N_{tr}}$ contain the discriminative sparse codes of the training features $\mathbf{V}$, while $\mathbf{M} \in \mathbb{R}^{Z\times Z}$ is a linear transformation matrix. In particular, $\mathbf{P}$ has a block-diagonal structure, where each one of the S blocks is an $m_i \times n_i$ matrix of ones, $\mathbf{J}_{m_i \times n_i}$, with m_i and n_i denoting the number of training feature vectors and dictionary items, respectively, which share the same class label (that is, correspond to the same speaker). For example, assuming $\mathbf{D} = [\mathbf{d}_1|\ldots|\mathbf{d}_6]$ and $\mathbf{V} = [\mathbf{v}_1|\ldots|\mathbf{v}_9]$, where $\mathbf{d}_1$, $\mathbf{d}_2$, $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ are from class 1, $\mathbf{d}_3$, $\mathbf{v}_4$, $\mathbf{v}_5$, and $\mathbf{v}_6$ are from class 2, and $\mathbf{d}_4$, $\mathbf{d}_5$, $\mathbf{d}_6$, $\mathbf{v}_7$, $\mathbf{v}_8$ and $\mathbf{v}_9$ are from class 3, $\mathbf{P}$ can be defined as

$$\mathbf{P} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \tag{3.9}$$

In addition, $\mathbf{M}$ transforms the original sparse codes $\mathbf{C}$ so as to increase their discriminative power in the new (sparse features) space $\mathbb{R}^Z$. As a result, the discriminative sparse-code error promotes (class) label consistency in the new (transformed) sparse codes by enforcing the features from the same speaker to have similar sparse representation.

In the following, let $\mathbf{Bc}$ define a linear classifier, where $\mathbf{B} \in \mathbb{R}^{S\times Z}$ denotes the classifier parameters, and $\mathbf{c}$ is a column of the sparse code matrix $\mathbf{C}$. The output of the linear classifier will be an $S \times 1$ vector, whose largest element corresponds to the index i if the sparse code $\mathbf{c}$ is related with speaker i. Thus, in order to estimate the linear classifier parameters $\mathbf{B}$, we incorporate the classification error $\|\mathbf{H} - \mathbf{BC}\|_F^2$, related with all the sparse codes contained in $\mathbf{C}$, into the objective function (3.8) as follows,

$$\begin{aligned} &\hat{\mathbf{D}}, \hat{\mathbf{C}}, \hat{\mathbf{M}}, \hat{\mathbf{B}} = \arg \min_{\mathbf{D},\mathbf{C},\mathbf{M},\mathbf{B}} \|\mathbf{V} - \mathbf{DC}\|_F^2 + \lambda_1\|\mathbf{P} - \mathbf{MC}\|_F^2 + \lambda_2\|\mathbf{H} - \mathbf{BC}\|_F^2 \\ &\text{s.t. } \|\mathbf{c}_j\|_0 = K\ ,\ \forall j = 1,\ldots,N_{tr}\ , \end{aligned} \tag{3.10}$$

where λ_1 and λ_2 are regularization parameters controlling the trade-off between the reconstruction error $\|\mathbf{V} - \mathbf{DC}\|_F^2$, the discriminative sparse-code error $\|\mathbf{P} - \mathbf{MC}\|_F^2$, and the classification error $\|\mathbf{H} - \mathbf{BC}\|_F^2$. Matrix $\mathbf{H} = [\mathbf{h_1}|\cdots|\mathbf{h}_{N_{tr}}] \in \mathbb{R}^{S\times N_{tr}}$ contains the class labels (or speaker index) of the training features $\mathbf{V}$. The column $\mathbf{h}_j \in \mathbb{R}^{S\times 1}$, which corresponds to the training feature vector $\mathbf{v}_j \in \mathbf{V}$ of the i^{th} speaker, is defined as an all-zeros vector except for the index corresponding to the true speaker label $i \in \{1,\ldots,S\}$. For example, a label vector

$$\mathbf{h}_i = (0,\ldots,0,1,0,\ldots,0)^T \in \mathbb{R}^{S\times 1}$$

corresponding to a training vector $\mathbf{v}_{i,n}$, $i \in \{1,\ldots,S\}$ and $n \in \{n_{i-1},\ldots,n_{i-1}+n_i-1\}$ with

$n_0 = 1$, where the non-zero element indicates that the training feature vector belongs to speaker j.

The K-SVD Algorithm 2 is adopted in the proposed scheme to estimate simultaneously the unknown parameters by solving the reformulated optimization problem (3.10) of the form

$$\begin{aligned} \hat{\mathbf{D}}, \hat{\mathbf{C}}, \hat{\mathbf{M}}, \hat{\mathbf{B}} &= \arg \min_{\mathbf{D},\mathbf{C},\mathbf{M},\mathbf{B}} \left\| \begin{pmatrix} \mathbf{V} \\ \sqrt{\lambda_1}\mathbf{P} \\ \sqrt{\lambda_2}\mathbf{H} \end{pmatrix} - \begin{pmatrix} \mathbf{D} \\ \sqrt{\lambda_1}\mathbf{M} \\ \sqrt{\lambda_2}\mathbf{B} \end{pmatrix} \mathbf{C} \right\|_F^2 \\ \text{s.t. } \|\mathbf{c}_j\|_0 &= K \,, \ \forall j = 1, \ldots, N_{tr} \,. \end{aligned} \tag{3.11}$$

After the solution of the optimization problem (3.11), the estimated dictionary $\hat{\mathbf{D}}$ and classifier parameters' matrix $\hat{\mathbf{B}}$ are exploited for the final classification process. Given a test sample $\mathbf{x}_t$ we first compute its sparse representation by solving

$$\hat{\boldsymbol{\gamma}} = \arg \min_{\boldsymbol{\gamma}} \|\mathbf{x}_t - \hat{\mathbf{D}}\boldsymbol{\gamma}\|_2 \,, \ \text{s.t. } \ \|\boldsymbol{\gamma}\|_0 = K \tag{3.12}$$

through the OMP algorithm. Finally, the estimated linear classifier $\hat{\mathbf{B}}$ is applied to estimate the class (or the speaker identity) of the test sample by finding the index of the maximum value of the class label vector

$$\begin{aligned} &\boldsymbol{\tau} = \hat{B}\hat{\boldsymbol{\gamma}} \\ &i^* = \arg \max_i \boldsymbol{\tau}(i) \,, \ i = 1, \ldots, S \ , \end{aligned} \tag{3.13}$$

where $\boldsymbol{\tau} \in \mathbb{R}^{S \times 1}$. As in SRC, this classification process is followed for each speech signal's frame, where finally majority voting is performed for a predefined set of frames to find the unknown speaker's identity.

3.5 Experimental results

In this section, the identification performance of the proposed discriminative K-SVD approach, described in Section 3.4, is evaluated in terms of the correct identification rate, and is compared with the SRC approach (discussed in Section 2.3.2) constituting the key part of the recent classification approaches for speech signals mentioned in Section 3.2. We also use the UBM-GMM [39] as the second method for comparison. The speech signals used in the subsequent experimental evaluations are obtained from the VOICES corpus, which is available from OGI's CSLU [87], consisting of 12 speakers (7 male and 5 female).

The original signals are sampled at 22 kHz, and downsampled to 8 kHz. During the fea-

Table 3.1: Average correct identification rates (%) for the discriminative K-SVD, SRC and UBM-GMM for five different number of SNR values and four noise types: white, speech babble, car engine and factory floor. The duration of the training data is 10 sec.

Noise	SNR	K-SVD		SRC	UBM-GMM
	(dB)	25	50		
White	20	89.11	96.32	92.89	97.41
	15	86.24	97.43	87.15	98.42
	10	82.92	86.77	83.45	96.41
	5	74.75	71.96	58.71	47.70
	0	57.04	51.57	31.50	34.67
Avg.		*78.01*	*80.81*	*70.74*	*74.92*
Speech babble	20	83.95	80.41	89.88	73.90
	15	88.05	81.18	86.28	55.99
	10	80.23	83.25	70.06	30.76
	5	65.33	71.62	20.76	15.16
	0	46.43	47.55	9.46	13.77
Avg.		*72.79*	*72.80*	*55.28*	*37.91*
Engine car	20	85.52	86.45	83.29	61.55
	15	76.69	82.12	69.32	49.53
	10	50.92	64.84	65.74	34.75
	5	24.75	42.55	33.82	26.80
	0	13.55	27.65	17.36	17.85
Avg.		*50.28*	*60.72*	*53.90*	*38.09*
Factory floor	20	84.10	80.39	84.84	66.09
	15	78.32	79.92	73.16	49.39
	10	73.10	75.64	63.92	11.69
	5	45.83	59.41	16.87	8.34
	0	18.12	44.18	8.33	8.33
Avg.		*59.89*	*67.90*	*49.42*	*28.76*

ture extraction step, an analysis window of 320 samples, with 50% overlapping between two consecutive frames, is employed to compute a mel-frequency spectrogram of $\Omega = 40$ bands, where a silence detector algorithm based on the short-term energy and zero-crossings measure of speech segments is applied[1]. The resulting $\Omega \times T$ mel-spectrogram, where T is the total number of frames on which mel-frequency analysis was performed, is reshaped by vectorizing every ϕ consecutive columns, and thus the new matrix is of size $\phi\Omega \times \lfloor T/\phi \rfloor = \tilde{\Omega} \times \tilde{T}$. For the UBM-GMM framework a diagonal covariance matrix was chosen during the simulations. We pooled all the target speakers training data using the mel-scale frequency coefficients of order $\Omega = 40$, where after experimentation we found that best results on average obtained when used 64 number of mixtures.

It is also important to point out that for the K-SVD and SRC-based simulations $\phi = 13$

[1] http://www.mathworks.com/matlabcentral/fileexchange/19298-speechcore

following the same vectorizing strategy as in exemplar-based techniques (ref. Section 3.2). In addition, $\phi = 1$ during the UBM-GMM evaluation process as a consequence of a more stable behaviour in capturing the discriminative statistics of lower dimensional features corresponding to short training data as in our study.

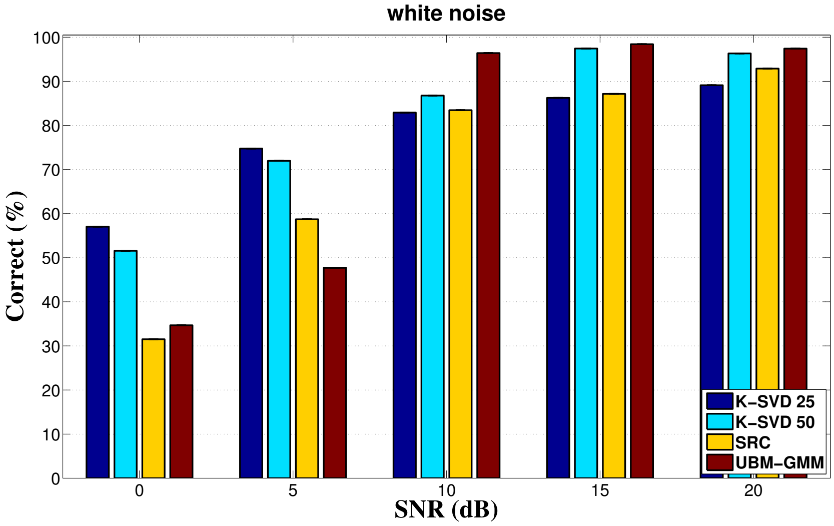


Figure 3.1: Speaker identification performance as a function of the white noise SNR.

The duration of the training data was around 10 sec per speaker. The average correct identification rate is computed as the percentage of the correctly identified segments over the total number of test segments. For each speaker, the total number of test segments used for the evaluation is approximately equal to 70, obtained by sliding a window of 15.6 sec over the time interval of the last 10 utterances, whose duration is about 60 sec.

The test utterances are corrupted by four different types of additive noise: white noise, speech babble noise, car engine noise and factory floor noise, where the SNR of the corrupted speech takes the values of 0, 5, 10, 15 and 20 dB. The noise signals were taken from the NOISEX-92 database [88]. In all cases, the data were trained under the multicondition framework [81], where the training dataset is enlarged by corrupting the clean speech training data with simulated noise of different characteristics. Here, the clean speech data are corrupted by white noise of SNR 10, 15 and 20 dB. The sparsity threshold K mentioned in Sections 2.3.2 and 3.4 was chosen experimentally to be 10 during the SRC evaluation procedure, while for K-SVD a sparsity threshold equal to 25 was found to give the best performance. Besides, the regularization parameters λ_1 and λ_2 of optimization problem (3.11) set equal to 0.25 and 2.25 on average, respectively.

As we can see from the experimental results in Table 3.1 (a visualization of the table can be found in Figures 3.1- 3.4), SRC achieves at least 15% higher average identification rates compared with the UBM-GMM with an exception in the case of white noise, where UBM-GMM

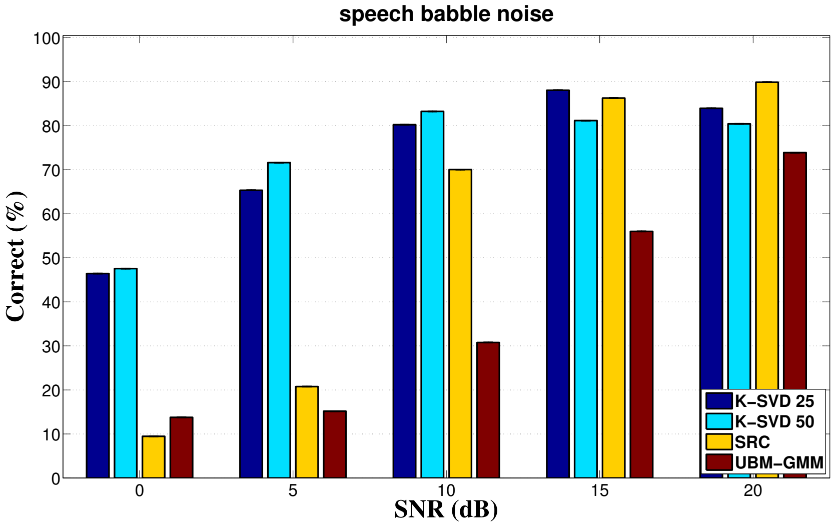


Figure 3.2: Speaker identification performance as a function of the speech babble noise SNR.

is about 4% better. The third and fourth column correspond to the identification rates obtained using a learned K-SVD dictionary of size 25% and 50% (termed as KSVD-25 and KSVD-50) of the initial training data matrix size, respectively. It is obvious that the proposed discriminative K-SVD approach is on average far better than that of the two methods used for comparison in both dictionary size schemes. A correct identification rate of at least 60% is on average achieved with the KSVD-25 in the case of the three out of the four noise types. In addition, KSVD-50 accomplishes at least approximately 70% in three of the four noisy conditions, where in noisy conditions such as 0 and 5 dB SNR is quite robust compared with the two methods used for comparison that completely fail to achieve acceptable identification rates.

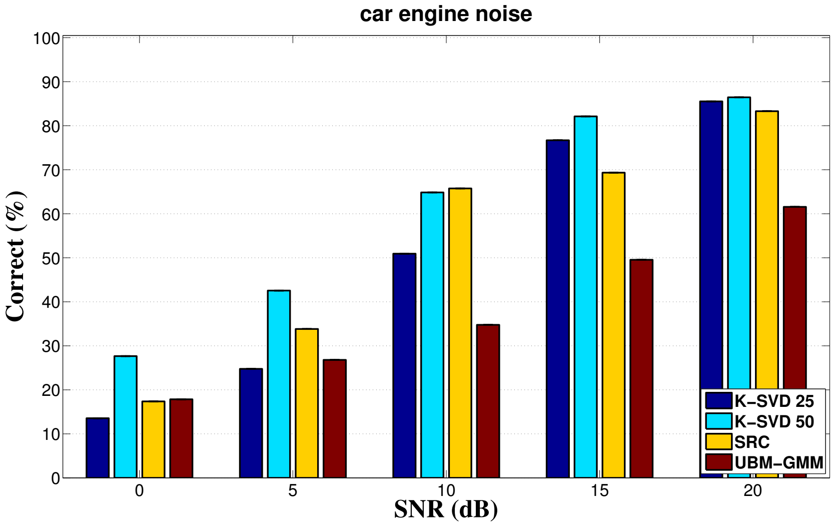


Figure 3.3: Speaker identification performance as a function of the car engine noise SNR.

It is also important to notice how the identification rates are compared between KSVD-50 and KSVD-25. In particular, we note that KSVD-25 achieves almost similar identification rates in the case of white and speech babble noise compared to KSVD-50 and it performs lower than KSVD-50 (approximately 10% lower rates) in the case of car engine and factory floor noise. Computational cost is very crucial in real-time applications of speaker identification. In such applications we would like to achieve as high as possible correct identification rates using small amount of data. Towards this direction, KSVD-25 could be applied on 25% of the initial training data in order to achieve robust identification rates under adverse noisy conditions. Figure 3.5 shows the average correct identification rates (where the mean value across all SNR values per noise type is computed) of all the methods for all types of noise. It is obvious that both discriminative dictionary sparse coding techniques, i.e., K-SVD 25 and K-SVD 50, are superior

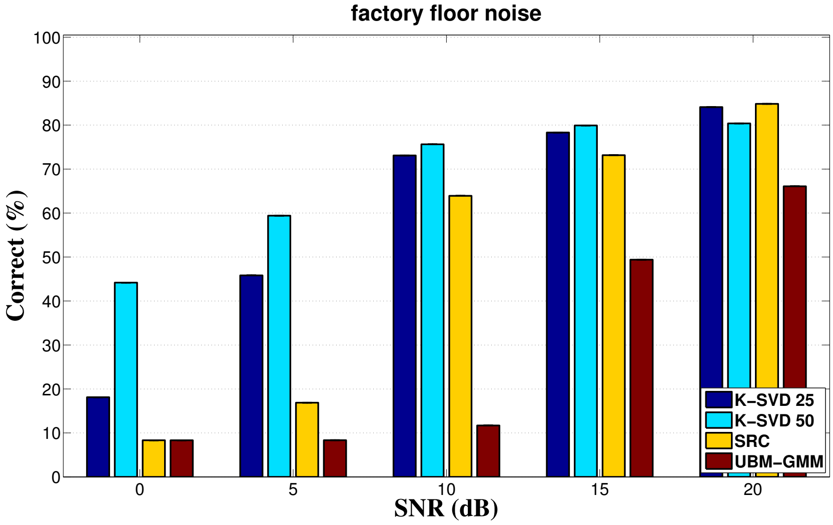


Figure 3.4: Speaker identification performance as a function of the factory floor noise SNR.

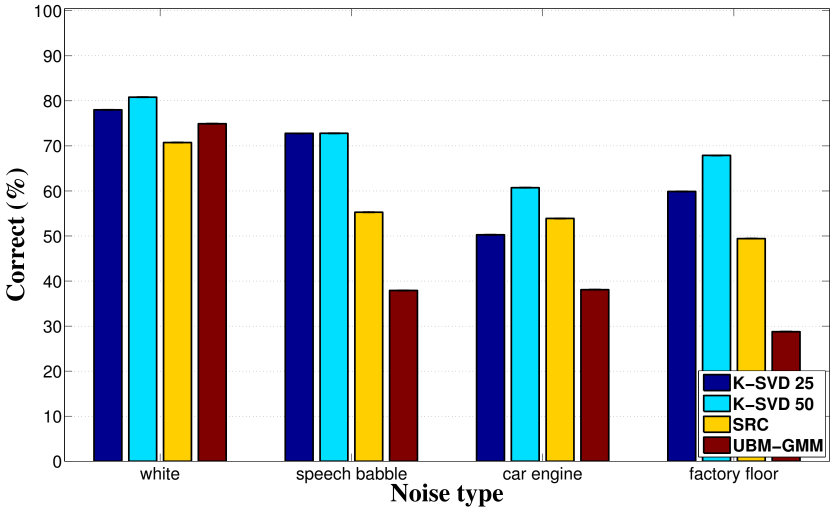


Figure 3.5: Average correct identification rates across all noise types comparing K-SVD 25 vs. K-SVD 50 vs. SRC vs. UBM-GMM.

to SRC and UBM-GMM except car engine noise where SRC is slightly better than K-SVD 25.

Part II

Low-rank techniques for recovery of missing features

Chapter

4

Missing features reconstruction based on a low-rank assumption

> Errors using inadequate data are much less than those using no data at all.
>
> Charles Babbage (1791-1871)

4.1 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. A literature review on MDT methods can be found in Section 1.1.2. For the sake of completeness we briefly mention below some of the basic works in the field. 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 [56, 58, 59], 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 [49, 50, 51, 52, 53, 54, 55] 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 [47].

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 [60, 61, 62], while in [64, 65] speaker verification is studied in the light of missing feature theory for improvement of recognition performance, while in [66] 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 this thesis, a novel imputation scheme based on matrix completion [99] 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 as examined in the previous sections. 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 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 [53].

4.2 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 $\mathcal{O}(nr\nu \ln^2 n)$ entries selected uniformly at random (with ν corresponding to the so-called degree of incoherence) [100], where $n = \max\{n_1, n_2\}$ and $r = \text{rank}(\boldsymbol{M})$. We 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

$$\begin{aligned} \min_{\boldsymbol{X}} \quad & \|\boldsymbol{X}\|_* \\ \text{s.t.} \quad & X_{ij} = M_{ij}\,, \ \ (i,j) \in \mathcal{I} \subset \{1,\dots,n_1\} \times \{1,\dots,n_2\}, \end{aligned} \tag{4.1}$$

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 $\sigma_1, \dots, \sigma_{\min\{n_1,n_2\}} \geq 0$ corresponding to the singular values of $\boldsymbol{X}$.

In the following, let the standard matrix completion *linear map* $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \rightarrow \mathbb{R}^k$. The constraints $X_{ij} = M_{ij}\,, \ \forall\,(i,j) \in \mathcal{I}$ in (4.1) can be represented by using the linear map $\mathcal{A}_{\mathcal{I}}$ as follows

$$\min_{\boldsymbol{X}} \ \|\boldsymbol{X}\|_* \ \text{ s.t. } \mathcal{A}_{\mathcal{I}}(\boldsymbol{X}) = \boldsymbol{b}, \tag{4.2}$$

where $\boldsymbol{b} := \mathcal{A}_{\mathcal{I}}(\boldsymbol{M})$ 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 (4.2) can also be written in matrix form

$$\mathcal{A}_{\mathcal{I}}(\boldsymbol{X}) \equiv \boldsymbol{A}\boldsymbol{x}, \ \boldsymbol{x} := \operatorname{vec}(\boldsymbol{X}) \ \ \forall \boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}, \tag{4.3}$$

where $\boldsymbol{A} \in \mathbb{R}^{k \times n_1 n_2}$ and $\operatorname{vec}(\cdot) : \mathbb{R}^{n_1 \times n_2} \rightarrow \mathbb{R}^{n_1 n_2 \times 1}$ 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,

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

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) [101] 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

$$\min_{\boldsymbol{X}} \tau \|\boldsymbol{X}\|_* + \frac{1}{2} \|\boldsymbol{X}\|_F^2 \ \text{ s.t. } \mathcal{A}_{\mathcal{I}}(\boldsymbol{X}) = \mathcal{A}_{\mathcal{I}}(\boldsymbol{M}), \tag{4.4}$$

where the positive constant τ is a trade off between the nuclear and Frobenius norm. The solution to problem (4.4) converges to that of (4.1) as $\tau \to \infty$. SVT comprises the two following iterative steps

$$\begin{cases} \boldsymbol{X}_t = \mathcal{D}_\tau(\mathcal{A}_\mathcal{I}^*(\boldsymbol{y}_{t-1})) \\ \boldsymbol{y}_t = \boldsymbol{y}_{t-1} - \delta(\mathcal{A}_\mathcal{I}(\boldsymbol{X}_t) - \boldsymbol{b}). \end{cases} \tag{4.5}$$

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 $\boldsymbol{\Sigma}_\tau = \text{diag}(\max\{\sigma_i - \tau, 0\})$ 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.

4.3 Missing-features recovery using low-rank matrix completion

As it was mentioned in the introduction, in the current part of our work the main 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

$$W(f,\rho) = \begin{cases} 1 := \text{reliable}, & 10\log_{10}\left(\frac{|S(f,\rho)|}{|N(f,\rho)|}\right) > \lambda \\ 0 := \text{unreliable}, & \text{otherwise} \end{cases} \tag{4.6}$$

where $\boldsymbol{W} \in \mathcal{B}^{F\times P}$ with $\mathcal{B} = \{0,1\}$ and λ is a pre-defined 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 (4.2) as follows

$$\hat{\boldsymbol{Y}} = \arg\min_{\boldsymbol{X}} \ \|\boldsymbol{X}\|_* \quad \text{s.t. } \mathcal{A}_\mathcal{I}(\boldsymbol{X}) = \mathcal{A}_\mathcal{I}(\boldsymbol{W} \odot \boldsymbol{Y}). \tag{4.7}$$

The linear map $\mathcal{A}_\mathcal{I}$ in (4.7) is related with matrix $\boldsymbol{A}$ as defined in (4.3), where the set of indices

$\mathcal{I}$ corresponds to the non-zero entries of the binary mask $\boldsymbol{W}$

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

Optimization problem (4.7) can be rewritten as

$$\begin{aligned} \hat{\boldsymbol{Y}} = \arg\min_{\boldsymbol{X}} \quad & \tau \left\|\boldsymbol{X}\right\|_* + \frac{1}{2}\left\|\boldsymbol{X}\right\|_F^2 \\ \text{s.t.} \quad & \mathcal{A}_{\mathcal{I}}(\boldsymbol{X}) = \mathcal{A}_{\mathcal{I}}(\boldsymbol{W} \odot \boldsymbol{Y}) \end{aligned} \tag{4.8}$$

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 OGI's CSLU [87]. 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. 4.1. 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 log-magnitude STFT matrix $\hat{\boldsymbol{Y}}$ is

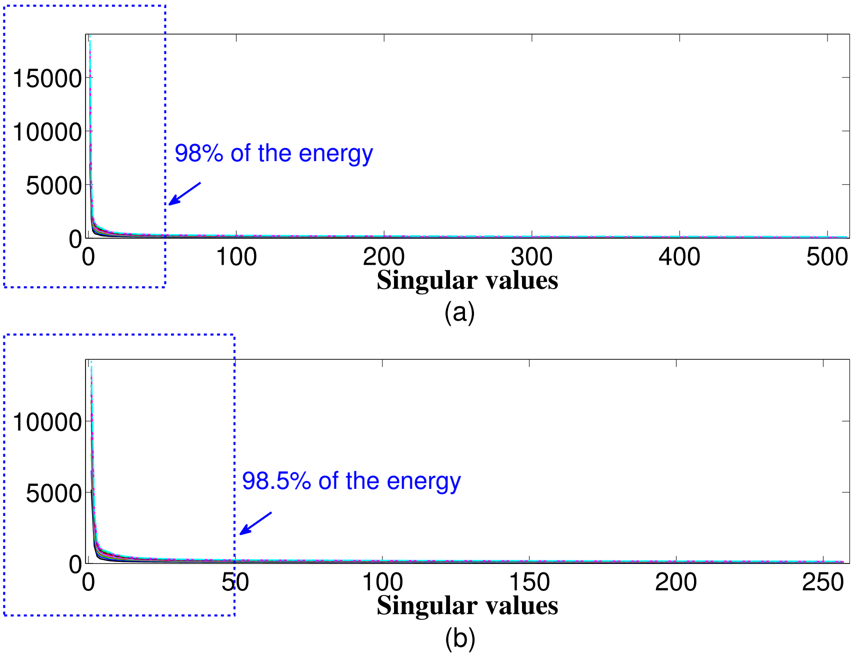


Figure 4.1: Ordered singular values spectra of the log-magnitude STFT spectrograms. The concentration of 98% of the energy is around 50.

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

$$\boldsymbol{Q} = 10 \cdot \log_{10}\left(\boldsymbol{B} \cdot 10^{\hat{\mathbf{Y}}/10}\right) \in \mathbb{R}^{d \times P}, \tag{4.9}$$

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

$$\mathbf{D} = \mathbf{\Psi}\boldsymbol{Q}, \tag{4.10}$$

where $\mathbf{\Psi}$ 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. A schematic representation of missing-feature recovery based on missing data imputation can be found in Figure 4.2.

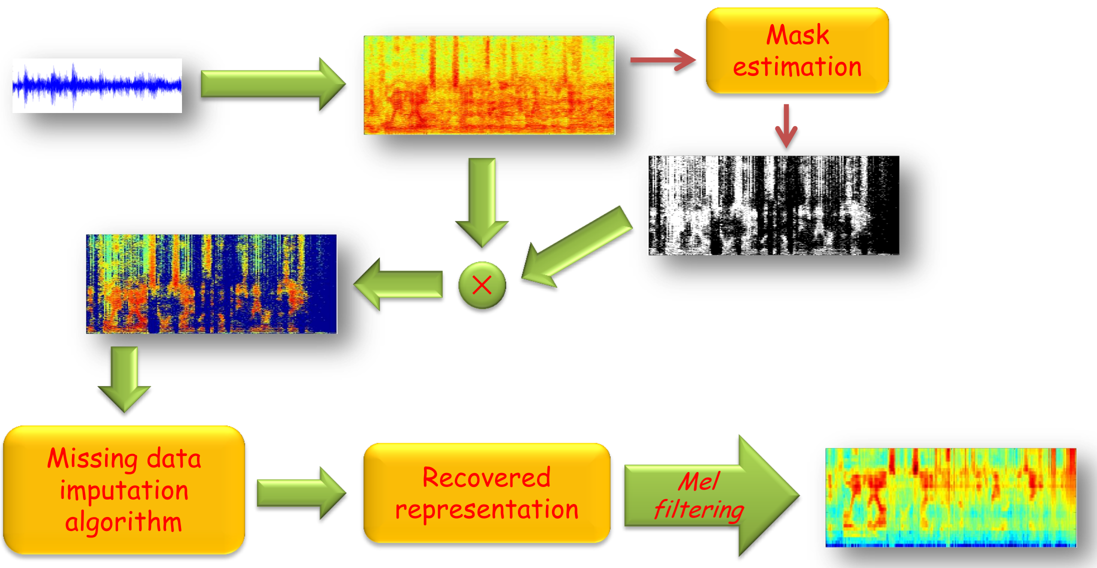


Figure 4.2: Flow diagram depicting the procedure of missing data imputation based on missing data imputation

4.4 Missing-feature recovery based on sparse imputation

In this section, we briefly describe the sparse imputation (SI) method [53] previously applied in the context of missing data imputation for robust speech recognition. The core idea in SI is

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

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 (4.9) and (4.10) 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

$$\tilde{\boldsymbol{d}}_{S,i} = \sum_{l=1}^{\beta} \alpha_{l,i} \boldsymbol{g}_l = \boldsymbol{G} \boldsymbol{\alpha}_i, \tag{4.11}$$

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 $\boldsymbol{G}_i$, $i = 1, \ldots, J$,

$$\begin{aligned} \boldsymbol{G} &= [\boldsymbol{g}_{1,1} | \cdots | \boldsymbol{g}_{1,m_1} | \boldsymbol{g}_{2,1} | \cdots | \boldsymbol{g}_{2,m_2} | \cdots | \boldsymbol{g}_{J,1} | \cdots | \boldsymbol{g}_{J,m_J}] \\ &= [\boldsymbol{G}_1 | \boldsymbol{G}_2 | \cdots | \boldsymbol{G}_J] \in \mathbb{R}^{dT \times \beta}, \end{aligned} \tag{4.12}$$

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

$$\hat{\boldsymbol{\alpha}}_i = \arg\min_{\boldsymbol{a}} \; \|\boldsymbol{a}\|_1 \quad \text{s.t. } \tilde{\boldsymbol{d}}_{S,i} = \boldsymbol{G}\boldsymbol{a}. \tag{4.13}$$

gives a unique solution to (4.11). Efficient ways to solve the convex optimization problem in (4.13) have been studied extensively. One way is to recast (4.13) as an ℓ_1 norm constrained least squares problem of the form

$$\hat{\boldsymbol{\alpha}}_i = \arg\min_{\boldsymbol{a}} \; \left\| \boldsymbol{G}\boldsymbol{\alpha} - \tilde{\boldsymbol{d}}_{S,i} \right\|_2 + \lambda \left\| \boldsymbol{\alpha} \right\|_1, \tag{4.14}$$

where the least absolute shrinkage and selection operator (LASSO) algorithm [102] 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 $\tilde{\mathbf{D}}^r_Y = \tilde{\boldsymbol{W}} \odot \tilde{\mathbf{D}}_Y$ gives a rough estimation of the reliable features. The reliable elements $\tilde{\boldsymbol{d}}^r_{Y,i}$ of the i-th column can be used to approximate the corresponding elements of $\tilde{\boldsymbol{d}}_{S,i}$ by solving the problem

$$\hat{\boldsymbol{\alpha}}_i = \arg\min_{\boldsymbol{a}} \left\| \boldsymbol{G}_r\boldsymbol{\alpha} - \tilde{\boldsymbol{d}}^r_{Y,i} \right\|_2 + \lambda \left\|\boldsymbol{\alpha}\right\|_1 , \tag{4.15}$$

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

$$\hat{\tilde{\boldsymbol{d}}}_{S,i} = \boldsymbol{G}\hat{\boldsymbol{\alpha}}_i. \tag{4.16}$$

It is important to note that by solving (4.15) the reconstruction error will not be zero in general, thus we only impute the unreliable elements

$$\hat{\tilde{\boldsymbol{d}}}_{S,i} = \begin{cases} \hat{\tilde{\boldsymbol{d}}}^r_{S,i} = \tilde{\boldsymbol{d}}^r_{Y,i} \\ \\ \hat{\tilde{\boldsymbol{d}}}^u_{S,i} = \boldsymbol{G}_u\hat{\boldsymbol{\alpha}}_i, \end{cases} \tag{4.17}$$

where $\boldsymbol{G}_u$ and $\hat{\tilde{\boldsymbol{d}}}^u_{S,i}$ 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 (4.15)-(4.17) for all columns of the features matrix $\tilde{\mathbf{D}}^r_Y$ 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.

4.5 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.

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

In this section, we compare the reconstruction performance of the SVT [101] algorithm with the performance obtained by reconstructing the missing data matrix using LMaFit [103] and ScGrassMC [104]. The experimental set-up, also used in our previous work [105], 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 [53] towards reducing the effects of noise in order to enhance the speaker identification accuracy. In the subsequent experimental evaluations we use UBM-GMM[2] [39] 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 [88]. For each combination of noise type and SNR level, the sampling ratio of the observed matrix $\boldsymbol{W} \odot \boldsymbol{Y}$ is defined as

$$\text{Sampling ratio} = \frac{\text{number of observed values } (k)}{\text{matrix size } (F \times P)}. \tag{4.18}$$

We note that the sampling ratio (4.18) is inversely proportional to the number of zeros in the binary mask $\boldsymbol{W}$ as defined in (4.6), 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. As a result, we can define the missing values ratio as follows

$$\text{Missing values ratio} = 1 - \text{Sampling ratio} = 1 - \frac{\text{number of observed values } (k)}{\text{matrix size } (F \times P)}. \tag{4.19}$$

The performance evaluation follows the strategy described in [97]. In particular, having solved (4.7) each completed matrix $\hat{\boldsymbol{Y}}$ corresponds to a sequence of feature vectors (columns)

[2] Universal Background Model for Gaussian Mixture Model

$\{\hat{\boldsymbol{y}}_t \in \mathbb{R}^{F\times 1}\}_{t=1}^{P}$ 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

$$\begin{gathered}
\underbrace{\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{y}}_2, \hat{\boldsymbol{y}}_3, \ldots, \hat{\boldsymbol{y}}_Q}_{1^{\text{st}}\text{ segment}} \hat{\boldsymbol{y}}_{Q+1}, \ldots, \hat{\boldsymbol{y}}_{P-1}, \hat{\boldsymbol{y}}_P \\
\hat{\boldsymbol{y}}_1, \underbrace{\hat{\boldsymbol{y}}_2, \hat{\boldsymbol{y}}_3, \ldots, \hat{\boldsymbol{y}}_Q, \hat{\boldsymbol{y}}_{Q+1}}_{2^{\text{nd}}\text{ segment}}, \ldots, \hat{\boldsymbol{y}}_{P-1}, \hat{\boldsymbol{y}}_P \\
\vdots \\
\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{y}}_2, \hat{\boldsymbol{y}}_3, \ldots, \hat{\boldsymbol{y}}_Q, \hat{\boldsymbol{y}}_{Q+1}, \ldots, \hat{\boldsymbol{y}}_{P-Q}, \underbrace{\hat{\boldsymbol{y}}_{P-Q+1}, \ldots, \hat{\boldsymbol{y}}_{P-1}, \hat{\boldsymbol{y}}_P}_{\text{P}-\text{Q}+1^{\text{th}}\text{ segment}}
\end{gathered} \tag{4.20}$$

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

$$\text{CIR}_j = \frac{\#\text{ cor. identified segments}}{\text{total}\,\#\text{ of segments}} \cdot 100\%, \tag{4.21}$$

where the total number of segments equals $P - Q + 1$. The total mean correct identification rate is used as an evaluation metric during the test simulations, which is given by

$$\text{mean CIR} = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{1}{J}\sum_{j=1}^{J}\text{CIR}_j^r\right), \tag{4.22}$$

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 (4.21).

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 4.3 and 4.4. 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. 4.3 shows the results corresponding to the speech babble noise, while Fig. 4.4 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

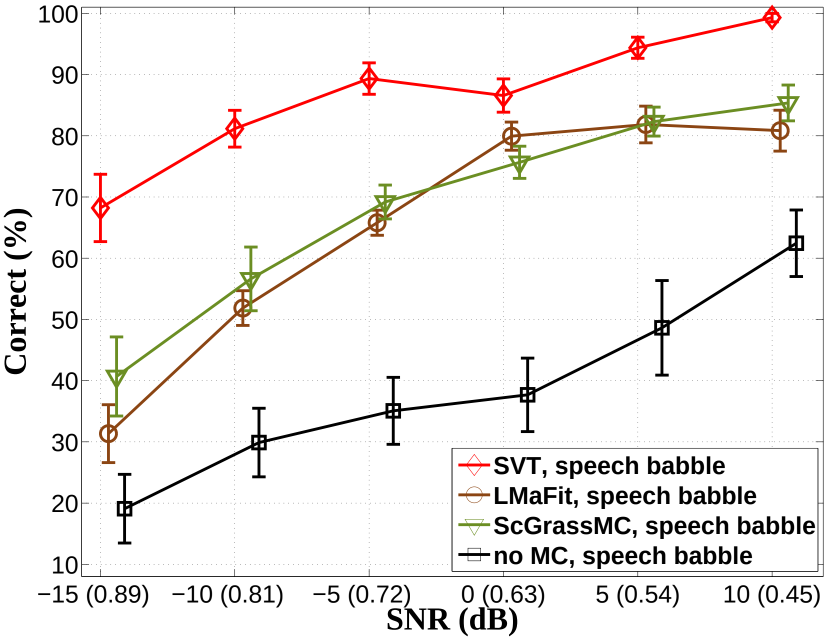


Figure 4.3: Mean correct identification rates (%) for the SVT, LMaFit, ScGrassMC and no MC for six different number of SNR values, where speech babble noise is added. The numbers inside the parentheses represent the missing values ratios (4.19).

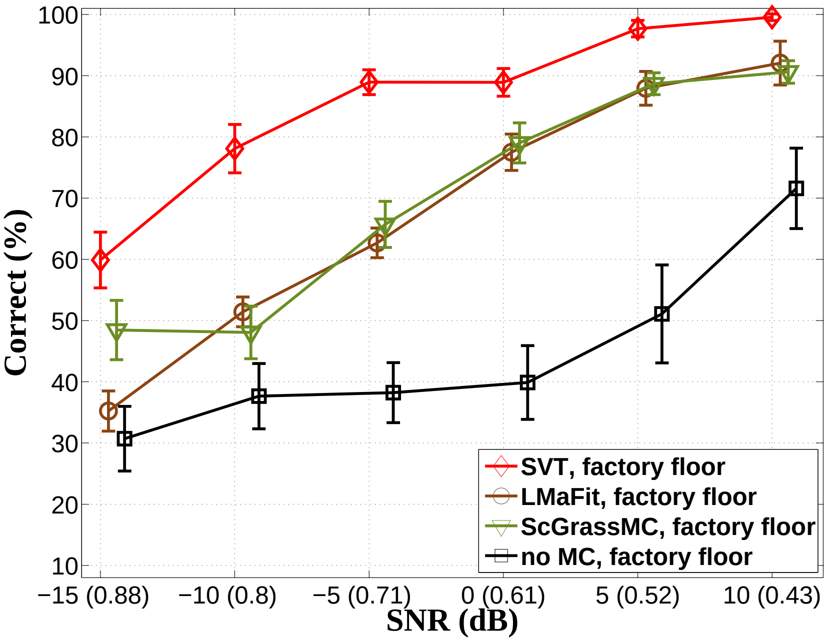


Figure 4.4: Mean correct identification rates (%) for the SVT, LMaFit, ScGrassMC and no MC for six different number of SNR values, where factory floor noise is added. The numbers inside the parentheses represent the missing values ratios (4.19).

-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%.

4.5.2 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 4.2 and 4.3, with respect to the resulting correct identification rates compared with the SI approach overviewed in Section 4.4. Fig. 4.5 and Fig. 4.6 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.

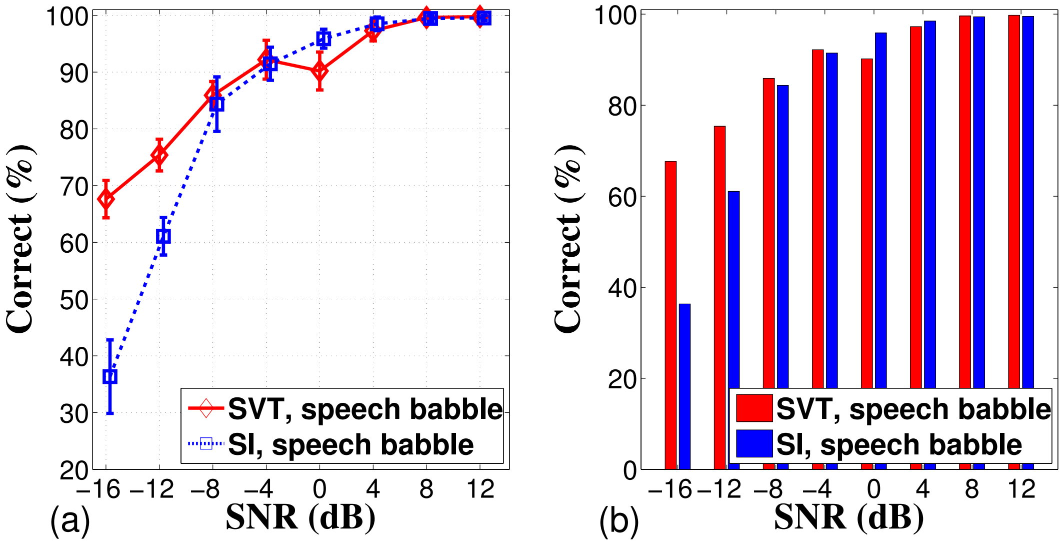


Figure 4.5: Mean correct identification rates (%) for the SVT vs. SI for eight different number of SNR values, where speech babble noise is added.

In Fig. 4.5.(a) and Fig. 4.6.(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 in low SNR values appear more clearly in the bar plots as depicted in Fig. 4.5.(b) and Fig. 4.6.(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

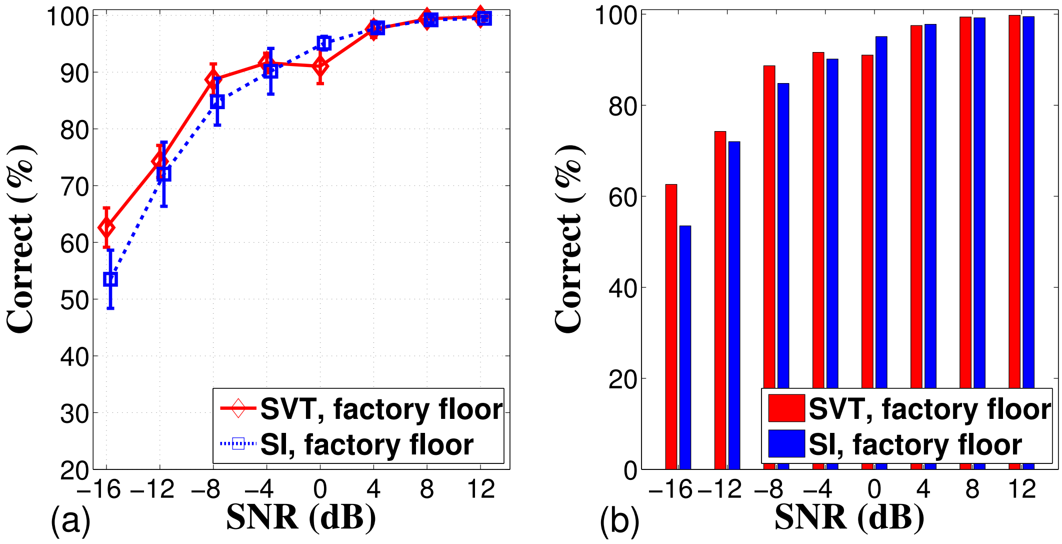


Figure 4.6: Mean correct identification rates (%) for the SVT vs. SI for eight different number of SNR values, where factory floor noise is added.

identification performance.

Chapter

5

Joint Low-Rank Representation and Matrix Completion Based on SVT

Try not to become a man of success, but rather try to become a man of value.

Albert Einstein (1879-1955)

5.1 Introduction

Many real-world problems often require the estimation of a matrix with missing entries. In general, the *matrix completion* problem involves the computation of the missing entries in a partially observed data matrix by imposing high data redundancy constraints through a *low-rank model.* The seminal papers [99, 106] prove that in many cases, the matrix can be correctly estimated with high probability from a number of observed entries greater than or equal to a certain constant value. The estimation is in the form of a rank minimization problem, where the *nuclear norm* [107], i.e., sum of the singular values, is used as the convex surrogate of the rank function.

Many algorithms have been proposed to solve the matrix completion (MC) problem. They can be summarized into two main categories with respect to the nature of the optimization problem. The first group of algorithms employs nuclear norm minimization such as in singular value thresholding (SVT) [101], templates for first-order conic solvers (TFOCS) [108], accelerated proximal gradient (APGL) [109] and augmented Lagrange multiplier (ALM) [110]. The second class of MC algorithms minimizes an approximation error objective function on a Grassmann manifold as examined in OPTSPACE [111], subspace evolution and transfer (SET) [112], Grassmanian rank-one update subspace estimation (GROUSE) [113], scaled gradients on Grassmann manifolds (ScGrassMC) [104], etc. Additionally, the low-rank matrix fitting algorithm (LMaFit) [103] optimizes an approximation error objective function based on the nuclear norm minimization framework, while in [114] MC is studied from a Bayesian point of view.

Over the last few years, MC has been tested in a wide range of practical applications including robust video denoising [115], bearing estimation of narrowband sources in sensor arrays [116], received signal-strength fingerprint based indoor localization in wireless local area networks [117] and audio bandwidth expansion [118]. It has also been utilized for other scientific problems such as position calibration in circular ultrasound tomography devices [119], high-quality reconstructions for large scale seismic interpolation problems [120], etc.

Nuclear norm minimization for subspace segmentation has been developed in parallel with MC since the germinal work introduced in [121]. The described *low-rank representation* idea looks for the lowest rank estimate of a data matrix with respect to a collection of data drawn from a union of multiple subspaces. Specifically, a learned dictionary or the data matrix itself can be exploited for seeking the low-rank representation (LRR) of the data. LRR seems to be very promising especially for classification tasks. For example, [122, 123] show that minimizing a nuclear norm based objective function coupled with sparsity constraints and a discriminative (or supervised) term enhances the power to discriminate features in image recognition. In [124] LRR is also adopted for music tagging, while in [125] is extended to the case of multiple dictionaries for music and singing voice separation.

Here, we propose a joint LRR and MC approach in the light of SVT framework. Especially, we are interested in studying the effect of *estimating the lowest rank representation of a data matrix with respect to a given basis or dictionary connected with a partially observed version of it under an SVT scheme.* A dictionary based MC method has been recently proposed in [126], where a similar optimization problem is examined for reconstruction and classification of simulated sensor network data using the CVX software package [127]. This method can potentially solve problems of very small size, however, the computational time is prohibitive for practical applications even for data matrices of moderate size. The novelty of the proposed approach is twofold. Firstly, in the current work a more rigorous mathematical formulation of the joint LRR and MC problem is presented by restating the optimization problem and giving a detailed algorithmic process for the estimation of the data matrix. Secondly, we employ an SVT algorithmic solution especially targeted for medium scale data, where an experimental evaluation is performed on synthetic data proving the efficacy of the proposed method. To the best of our knowledge, this is the first time that LRR is connected with MC under an SVT algorithmic process. Our proposed approach can be regarded as an enhanced version of SVT in the case that we have knowledge of the data generation process via a dictionary or basis. Therefore, we are strongly interested in examining the performance of the proposed algorithm versus the performance of the typical SVT algorithm under these conditions.

The rest of the chapter is organized as follows: Section 5.2 describes the proposed joint LRR and MC approach along with an SVT-based solution. An experimental evaluation of the proposed technique compared with typical SVT algorithm is described in Section 5.3.

5.2 Joint low rank representation and matrix completion using SVT

Singular value decomposition (SVD) followed by soft-thresholding on the computed singular values constitutes the core of the SVT algorithm described in Section 4.2. Any procurable information of the underlying procedure that generated the data matrix $\boldsymbol{M}$ is not taken into account by MC. Sometimes this property is considered as an asset since it does not require the explicit knowledge of such a generation procedure. In other cases, however, extra information about the data matrix is available and exploiting this knowledge can lead to more accurate solutions of different tasks at hand.

As mentioned in Section 5.1, the low-rank representation (LRR) approach has been recently introduced as an alternative to typical subspace-based methods like the SVD. The goal is to find the lowest rank representation of a data matrix by solving the following convex optimization problem

$$\min_{\boldsymbol{L}} \ \|\boldsymbol{L}\|_* \ \text{ s.t. } \boldsymbol{M} = \boldsymbol{M}\boldsymbol{L}, \tag{5.1}$$

where $\boldsymbol{M}$ is the data matrix and $\boldsymbol{L}$ is a low-rank matrix. Adopting the LRR formulation, let us assume that the additional information of the data matrix $\boldsymbol{M}$ can be modelled according to a specific matrix decomposition of the form $\boldsymbol{M} = \boldsymbol{G}\boldsymbol{L}$, where $\boldsymbol{G}$ is a known dictionary and $\boldsymbol{L}$ is a low-rank matrix containing the corresponding representation coefficients. Thus, problem (5.1) can be formulated as

$$\min_{\boldsymbol{L}} \ \|\boldsymbol{L}\|_* \ \text{ s.t. } \boldsymbol{M} = \boldsymbol{G}\boldsymbol{L}. \tag{5.2}$$

To apply the LRR scheme on matrices with missing data, we use the linear sampling operator $\mathcal{A}_{\mathcal{I}}$. The proposed sampling scheme is a combination of MC and LRR and seeks a low-rank coefficient matrix $\boldsymbol{L}$ from a small number of measurements $\mathcal{A}_{\mathcal{I}}(\boldsymbol{M})$. Thus, the convex optimization problem takes the form below

$$\min_{\boldsymbol{L}} \ \|\boldsymbol{L}\|_* \ \text{ s.t. } \mathcal{A}_{\mathcal{I}}(\boldsymbol{X}) = \mathcal{A}_{\mathcal{I}}(\boldsymbol{M}) \text{ and } \boldsymbol{X} = \boldsymbol{G}\boldsymbol{L}. \tag{5.3}$$

The goal is to efficiently solve problem (5.3) in the context of the SVT algorithm so that we can solve large-scale problems. Hence, combining (4.4) and (5.3) we get the joint LRR and MC version of SVT dubbed J-SVT defined as follows

$$\min_{\boldsymbol{L}} \ \tau \|\boldsymbol{L}\|_* + \frac{1}{2}\|\boldsymbol{L}\|_F^2 \ \text{ s.t. } \mathcal{A}_{\mathcal{I}}(\boldsymbol{X}) = \boldsymbol{b} \text{ and } \boldsymbol{X} = \boldsymbol{G}\boldsymbol{L}, \tag{5.4}$$

where $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$, $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$, $\boldsymbol{G} \in \mathbb{R}^{n_1 \times K}$, $\boldsymbol{L} \in \mathbb{R}^{K \times n_2}$ and K denotes the size of the dictionary. In the J-SVT problem (5.4), we consider the additional constraint that $\boldsymbol{X}$ must

Algorithm 3: J-SVT algorithm

```
Input: $\mathcal{A}_\mathcal{I}$, observed values $\boldsymbol{b}$, dictionary $\boldsymbol{G}$, step size $\delta$, tolerance $\epsilon$, parameter $\tau > 0$,
       maximum iterations $t_{\max}$
Output: estimated matrix $\boldsymbol{X} = \boldsymbol{G}\boldsymbol{L}_T$
Initialization: $\boldsymbol{y}_1 = \tau\boldsymbol{b}/\|\boldsymbol{G}^T\mathcal{A}_\mathcal{I}^*(\boldsymbol{b})\|$
for $t = 1$ to $t_{\max}$ do
    $[\boldsymbol{U}_t, \boldsymbol{\Sigma}_t, \boldsymbol{V}_t, s_t] =$ SVDshrink($\boldsymbol{G}^T\mathcal{A}_\mathcal{I}^*(\boldsymbol{y}_t), \tau, s_{t-1}$)
    $\boldsymbol{L}_t = \boldsymbol{U}_t\boldsymbol{\Sigma}_t\boldsymbol{V}_t^T$
    if $\|\mathcal{A}_\mathcal{I}(\boldsymbol{G}\boldsymbol{L}_t) - \boldsymbol{b}\|_2 \leq \epsilon\,\|\boldsymbol{b}\|_2$ then
        break
    end
    $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t - \delta(\mathcal{A}_\mathcal{I}(\boldsymbol{G}\boldsymbol{L}_t) - \boldsymbol{b})$
end
```

be in the form $\boldsymbol{X} = \boldsymbol{G}\boldsymbol{L}$ for a fixed dictionary $\boldsymbol{G}$. This constraint only amounts to changing the linear operator, and that does not affect the convergence proofs of SVT under a correctly scaled δ. Recall that SVT converges with $\delta < 2\|\mathcal{A}_\mathcal{I}\|^{-2}$. We have the following similar result:

Theorem 5.1. *With step-size $\delta < 2\|\mathcal{A}_\mathcal{I} \circ \boldsymbol{G}\|^{-2}$, J-SVT produces a sequence $\boldsymbol{L}_t$ that converges to the unique minimizer of* (5.4).

Proof. The proof of convergence for the SVT algorithm only uses the fact that $\mathcal{A}_\mathcal{I}$ is a linear operator and can be extended to handle a generic linear operator $\mathcal{A}$. By letting $\mathcal{A} = \mathcal{A}_\mathcal{I} \circ \boldsymbol{G}$ and $\mathcal{A}^* = \boldsymbol{G}^T\mathcal{A}_\mathcal{I}^*$ we arrive at J-SVT. The step-size must satisfy $\delta < 2\|\mathcal{A}\|^{-2} = 2\|\mathcal{A}_\mathcal{I} \circ \boldsymbol{G}\|^{-2}$. □

Since $\|\mathcal{A}_\mathcal{I} \circ \boldsymbol{G}\| \leq \|\boldsymbol{G}\|$, the step-size can best estimated using any upper bound on the spectral norm of $\boldsymbol{G}$.

Algorithm 4 implements the `SVDshrink` operation. The `partialSVD`($\boldsymbol{Z}, s$) algorithm returns the top s singular values and singular vectors. The most common computational approach is the Lanczos method. Here, we use the implementation in PROPACK, which re-orthogonalizes the singular vectors as needed in order to improve numerical stability. These Lanczos methods only require matrix-vector multiplies of the form $\boldsymbol{Z}\boldsymbol{u}$ and $\boldsymbol{Z}^T\boldsymbol{v}$, and thus we take advantage of sparsity in $\boldsymbol{Z}$. If $\boldsymbol{G}^T$ has a fast transform, we can also take advantage of this, and never even need to explicitly form the $\boldsymbol{G}$ or $\boldsymbol{G}^T$ matrix (e.g., if $\boldsymbol{G}$ is the FFT or FFT-based).

In another improvement on regular SVT, we introduce the Nesterov accelerated [128] version, which applies to both MC and LRR-MC problems.

Theorem 5.2. *Algorithm 5 produces a sequence $\boldsymbol{L}_t$ that converges to the unique minimizer of* (5.4) *if $\delta \leq \|\mathcal{A}_\mathcal{I} \circ G\|^{-2}$.*

Proof. This is a special case of the framework in [108] and the strong convexity of the objective. □

Algorithm 4: `SVDshrink` algorithm

```
Input: internal integer parameter $\ell$
function SVDshrink($\boldsymbol{Z}, \tau, s_0$)
    $s \leftarrow s_0 + 1$
    repeat
        $[\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V}] =$ partialSVD($\boldsymbol{Z}, s$)
        $s \leftarrow s + \ell$
    until $\boldsymbol{\Sigma}_{s,s} \leq \tau$
    return $[\boldsymbol{U}, \mathcal{D}_\tau(\boldsymbol{\Sigma}), \boldsymbol{V}, s]$
end function
```

Algorithm 5: Accelerated J-SVT algorithm: identical to J-SVT except replace line 8 in J-SVT with the following and initialize $\boldsymbol{z}_1 = \boldsymbol{y}_1$.

```
8 $\boldsymbol{z}_{t+1} = \boldsymbol{y}_t - \delta(\mathcal{A}_\mathcal{I}(\boldsymbol{G}\boldsymbol{L}_t) - \boldsymbol{b})$
9 $\boldsymbol{y}_{t+1} = \boldsymbol{z}_{t+1} + \frac{t}{t+3}(\boldsymbol{z}_{t+1} - \boldsymbol{z}_t)$
```

Note that we have lost a factor of 2 in the step-size bound in the accelerated version, which is because we can no longer over-relax (see [129]). Despite the smaller step-size, it has faster convergence rate guarantees and typically works faster in practice.

5.3 Experimental results

In this section, we compare the reconstruction performance of the proposed J-SVT scheme with the performance obtained by reconstructing the missing data matrix using the SVT algorithm. For this purpose, we perform simulations on synthetic data, where the dictionary $\boldsymbol{G}$ and the low-rank representation matrix $\boldsymbol{L}$ are generated from normally distributed random samples. As an evaluation metric, we employ the relative error, which is defined as follows:

$$\text{Relative error} = \frac{\left\|\hat{\boldsymbol{X}} - \boldsymbol{M}\right\|_F}{\|\boldsymbol{M}\|_F},$$

where $\hat{\boldsymbol{X}}$ is the recovered matrix and $\boldsymbol{M}$ is the original full data matrix. In the present case study, the size of the original data matrix $\boldsymbol{M}$ is set equal to $n_1 \times n_2 = 300 \times 500$. The maximum number of iterations $t_{\max}$, the tolerance ϵ and the parameter τ are set equal to 100, 10^{-5} and $5\sqrt{n_1 n_2}$, respectively. The step size δ is set equal to 1.9 in the case of SVT, while for the accelerated version of J-SVT we use $\delta = \|G\|^{-2}$. In the subsequent experimental evaluation, the reconstruction performance of both the J-SVT and SVT algorithms is also examined as a function of the sampling ratio, which is given by

$$\text{Sampling ratio} = \frac{\text{number of observed values (k)}}{\text{matrix size } (\mathrm{n}_1 \times \mathrm{n}_2)} \; .$$

Based on 10 Monte Carlo runs for each scenario, the total average were computed to show the overall relative errors for each algorithm.

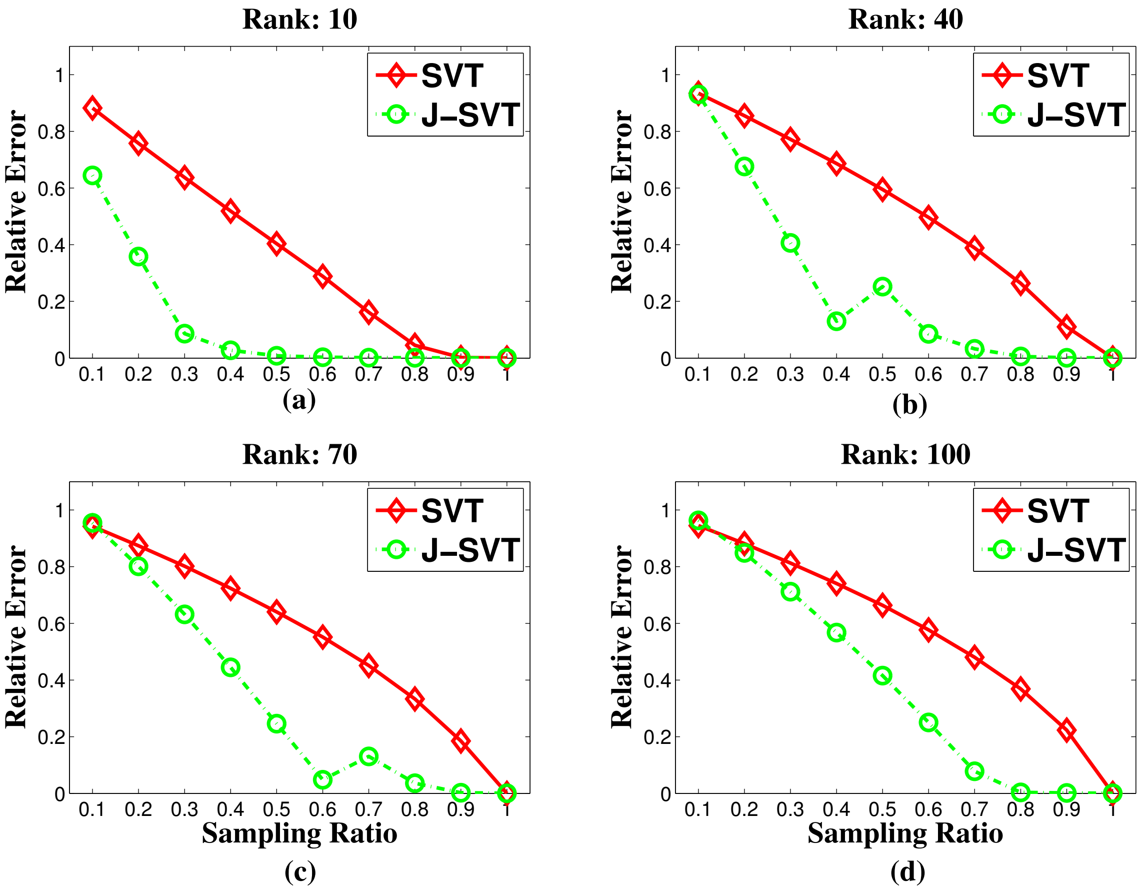
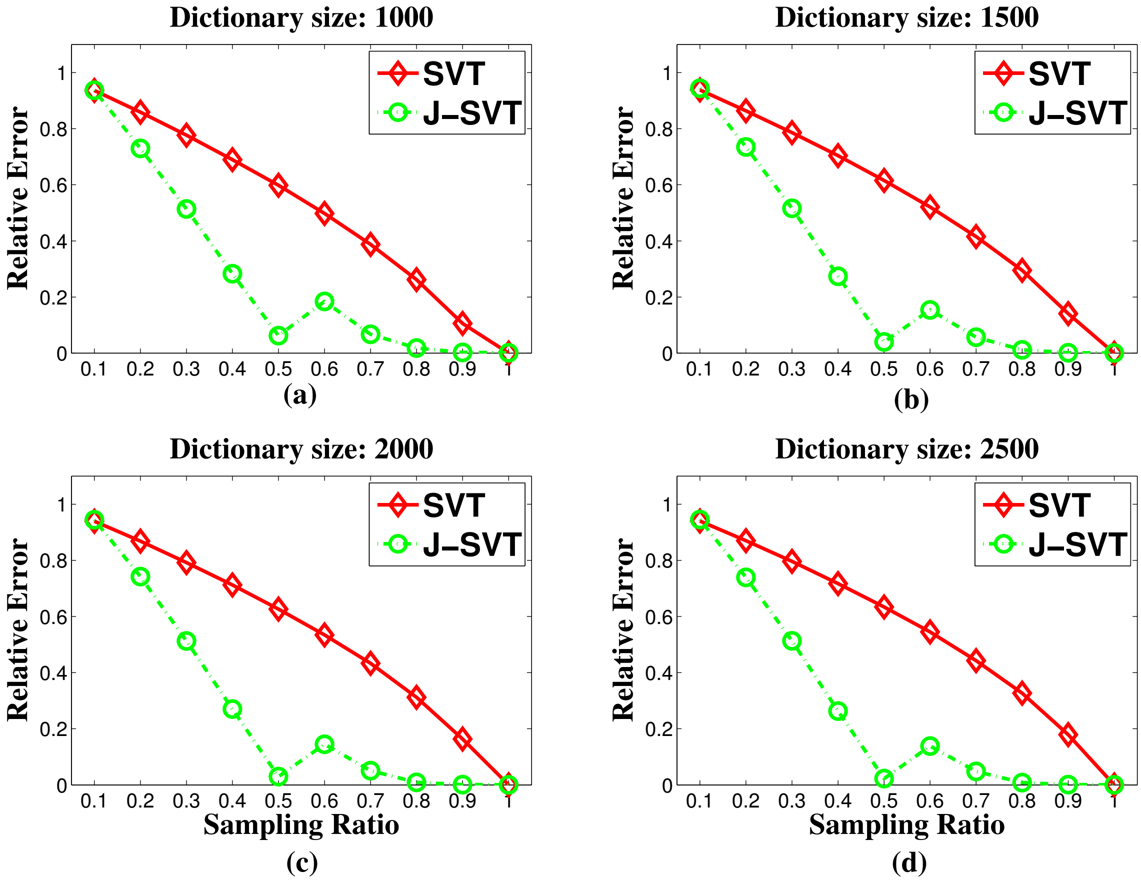


Figure 5.1: Relative error as a function of sampling ratio. The size of dictionary $\boldsymbol{G}$ is 300×1500. The rank of matrix $\boldsymbol{L}$ is: (a) 10, (b) 40, (c) 70 and (d) 100.

As a first set of experiments, we examine the reconstruction performance of J-SVT for a varying matrix rank. Figure 5.1 reveals that our proposed J-SVT algorithm outperforms clearly the SVT counterpart in case of a dictionary size 300×1500. More specifically, Figure 5.1.(a) shows that the relative error achieved by J-SVT is almost zero for a sampling ratio (SR) > 0.3, while the relative error achieved by SVT approaches zero for a significantly higher sampling ratio SR > 0.7. The effect of a varying matrix rank is shown in Figures 5.1.(b)-(d), which depict the reconstruction performance for matrix ranks equal to 40, 70 and 100, respectively. As it can be seen, the relative error corresponding to J-SVT is close to zero for SR ≈ 0.7, whereas the relative error of SVT approaches zero only for an almost full sampling (SR ≈ 0.9).

The second set of experiments concerns the performance evaluation of the two algorithms by varying the dictionary size. In Figure 5.2, the reconstruction accuracy of J-SVT is compared with the performance of SVT for dictionary sizes of 300×1000, 300×1500, 300×2000 and 300×2500, by fixing $\text{rank}(\boldsymbol{L}) = 50$. Clearly, J-SVT outperforms again SVT, while we highlight the approximately constant recovery behaviour of J-SVT regardless of the dictionary size. This



(a)

(b)

(c)

(d)

Figure 5.2: Relative error as a function of sampling ratio. The rank of matrix $\boldsymbol{L}$ is 50. The size of dictionary $\boldsymbol{G}$ is: (a) 300×1000, (b) 300×1500, (c) 300×2000 and (d) 300×2500.

observation is very important, since it reveals that J-SVT is highly robust, in terms of achieving a low reconstruction error, even in case of small-sized dictionaries, which represent our data in a compact way. This comes also as a significant advantage of J-SVT towards its application in practical scenarios, where the size of the dictionary comes at the expense of an increased computational and memory complexity.

As a final experimental evaluation, we compare the robustness of J-SVT against SVT under noisy conditions. In particular, the relative error curves presented in Figure 5.3 correspond to observed data corrupted by additive white noise, with the signal-to-noise ratio (SNR) being equal to 10, 15, 20 and 25 dB. As it can be seen J-SVT achieves a significantly improved reconstruction quality in regard with SVT. Especially in Figure 5.3.(b)-(d), SVT has almost twice as high relative error on average for the same range of sampling ratio values. As expected, the performance of SVT converges to the performance of J-SVT for a full sampling ratio (= 1).

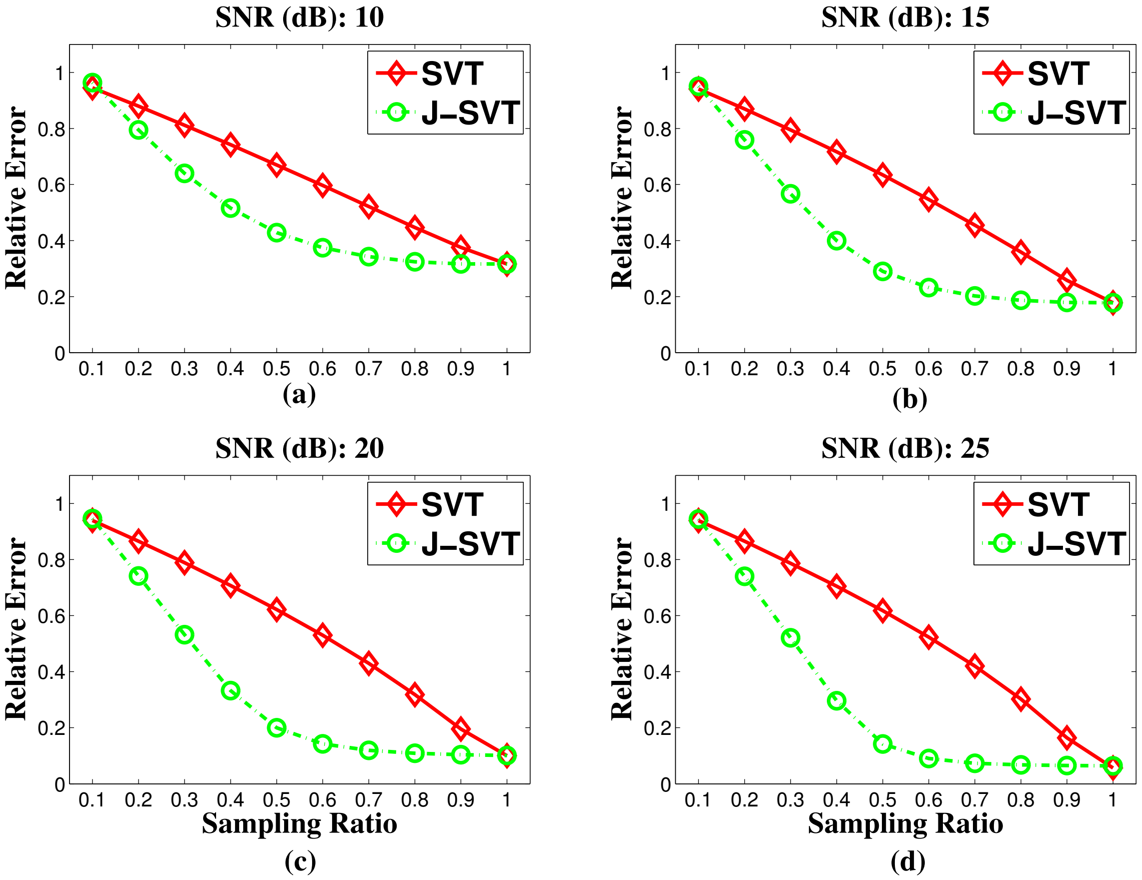


Figure 5.3: Relative error as a function of sampling ratio. The size of dictionary $\boldsymbol{G}$ is 300×1500 and the rank of matrix $\boldsymbol{L}$ is 50. The SNR level is set to: (a) 10 dB, (b) 15 dB, (c) 20 dB and (d) 25 dB.

Chapter 6

Conclusions and Future work

The future is always beginning now.

Mark Strand (1934-)

In this thesis, we studied the problem of robust speaker identification under the constraints of using a limited amount of training and evaluation speech data. In the first part of the current thesis, the focus is given on the problem of speaker identification using highly limited amounts of testing and training sessions, in noisy environments. A sparsity-based technique is proposed based on the assumption that the identified speech signal, and specifically the features that have been extracted from this signal, can be expressed as a sparse linear combination in terms of a dictionary. The optimally estimated sparse codes, which are obtained as the solutions of an optimization problem, are then employed for the final identification of the speaker based on a minimum reconstruction error criterion. An extension of the sparsity-based approach is then introduced to estimate jointly the dictionary comprising of the training data in conjunction with an appropriate linear classifier. The advantage of this approach is that it results in sparse codes, which are characterized by enhanced discriminative capability. Extensive experimental evaluations revealed the superiority of the proposed techniques compared to probabilistic approaches as well as compared to state-of-the-art speaker identification methods.

In the second part of this thesis, a technique for recovering a low-rank matrix is designed, which is employed for the reconstruction of those spectral regions of a speech signal, which are unreliable due to the presence of noise. The reconstruction of the unreliable spectral regions is performed by adopting the Singular Value Thresholding (SVT) algorithm, based on the assumption that the logarithmic magnitude representation of a speech signal in the time-frequency domain, obtained via the short-time Fourier transform (STFT), is of low rank. The comparison against the widely used method of sparse imputation, which is based on sparse representations, reveals the superiority of our proposed approach in terms of producing more reliable features. Then, an extended version of the matrix completion method, which exploits the prior knowledge that the data matrix is low rank, as well as the knowledge that the data can be represented efficiently in terms of a dictionary. In specific, a novel algorithm is proposed for joint low-rank representation and matrix completion (J-SVT), which is superior when compared with the standard SVT with respect to the computation of the low-rank representation of a data matrix in terms of a given dictionary, by employing a small number of observations from the original matrix. Through extensive simulations, we observed an improvement of the reconstruction error achieved by the J-SVT, in contrast to the typical SVT, for several distinct experimental scenarios.

There are still many open problems to be examined and future work to be done which will introduce further development on missing-feature reconstruction and discriminative sparse coding techniques. Some of them can be listed as below:

- Extend the experimental evaluation set-up in corpora containing more speakers and dealing with a broader range of noise types. Other types of applications could also be examined such speech stressed classification problems.

- Examine the discriminative properties of i-vectors compared to typical speech features such as MFCCs, especially under the discriminative dictionary learning framework.

- The combination of probabilistic models such as GGD with discriminative dictionary learning approaches, could lead in enhanced classification accuracy by taking advantage of the a-priori knowledge of the specific statistical behaviour of the speech features.

- The discriminative constraints modeled by the matrix $\mathbf{P}$ in (3.8) could be replaced by a distance metric (learning) constraint in order to further “push” the sparse codes from the same class to have very small distances.

- Another idea is related with the classification error in (3.10). In the current thesis, this error (i.e. matrix $\mathbf{B}$) is related with a linear classifier. A natural extension is to incorporate a non-linear classifier, which could lead to more robust classification performance.

- The unsupervised nature of SVT-based missing-feature reconstruction could be exploited to produce reliable log-magnitude STFT representations of noisy speech signals. This method is classifier-independent and thus, it could be straightforwardly used in automatic speech recognition applications, to deal with robustness issues.

- Practical estimated reliability masks could be applied to distinguish the reliable from the unreliable spectrotemporal regions. We expect that low-rank matrix completion will perform better than sparse imputation because it is more resistant to a large number of missing time-frequency bins.

- Experiments with real speech data using the J-SVT algorithm. Additionally, we could study the J-SVT approach in the light of L-BFGS approach as a tool to further decrease the computational time.

Bibliography

[1] J. Campbell, "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85(9), pp. 1437–1462, September 1997.

[2] D. Reynolds, "An overview of automatic speaker recognition technology," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Orlando, Florida, USA, May 2002, pp. 4072–4075.

[3] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. M. Chagnolleau, S. Meignier, T. Merlin, J. O. Garcia, D. P. Delacretaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 430–451, 2004.

[4] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[5] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Processing*, vol. 3(1), pp. 72–83, January 1995.

[6] K. H. Yuo, T. H. Hwang, and H. C. Wang, "Combination of autocorrelation-based features and projection measure technique for speaker identification," *IEEE Trans. on Speech and Audio Processing*, vol. 13(4), pp. 565–574, July 2005.

[7] J. C. Wang, C. H. Yang, J. F. Wang, and H. P. Lee, "Robust speaker identification and verification," *IEEE Comp. Intelligence Magazine*, vol. 2(2), pp. 52–59, May 2007.

[8] Y. Shao and D. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, April 2008.

[9] K. Kumar, Q. Wu, Y. Wang, and M. Savvides, "Noise robust speaker identification using Bhattacaryya distance in adapted gaussian models space," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Lausanne, Switzerland, August 2008.

[10] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16(6), pp. 1097–1111, August 2008.

[11] V. R. Apsingekar and P. L. D. Leon, "Speaker model clustering for efficient speaker identification in large population applications," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17(4), pp. 848–853, May 2009.

[12] A. Ariyaeeinia, J. Fortuna, P. Sivakumaran, and A. Malegaonkar, "Verification effectiveness in open-set speaker identification," *Vision, Image and Signal Processing, IEE Proceedings*, vol. 153(5), pp. 618–624, October 2006.

[13] P. Angkititrakul and J. Hansen, "Discriminative in-set/out-of-set speaker recognition," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15(2), pp. 498–508, February 2007.

[14] M. Zamalloa, L. Rodriguez, M. Penagarikano, G. Bordel, and J. Uribe, "Improving robustness in open set speaker identification by shallow source modelling," in *Proc. Odyssey: the Speaker and Language Recognition Workshop (Odyssey 2008)*, South Africa, January 2008.

[15] M. Hébert, *Text-dependent speaker recognition. In: Benesty, J., Sondhi, M., Huang, Y. (Eds.).* Springer-Verlag, Heidelberg, 2008.

[16] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15(4), pp. 1448–1460, May 2007.

[17] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," *Computer Speech and Language*, vol. 22(1), pp. 17–38, 2008.

[18] D. Sturim, W. Campbell, D. A. Reynolds, R. B. Dunn, and T. F. Quatieri, "Robust speaker recognition with cross-channel data: MIT-LL results on the 2006 NIST SRE auxiliary microphone task," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, Hawaii, USA, April 2007, pp. 49–52.

[19] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of Acoustical Society of America (JASA)*, vol. 55, pp. 1304–1312, 1974.

[20] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 29(2), pp. 254–272, April 1981.

[21] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Trans. on Speech and Audio Processing*, vol. 2(4), pp. 639–643, October 1994.

[22] R. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: A feature-based approach," *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 58–71, September 1996.

[23] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25(1-3), pp. 133–147, August 1998.

[24] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech and Audio Processing*, vol. 2(4), pp. 578–589, October 1994.

[25] J. Markel and A. Gray, Eds., *Linear Prediction of Speech.* Springer-Verlag, New York, 1982.

[26] M. S. Zilovic, R. P. Ramachandran, and R. J. Mammone, "Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer functions," *IEEE Trans. on Speech and Audio Processing*, vol. 6(3), pp. 260–267, May 1998.

[27] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, April 2003, pp. 53–56.

[28] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. 2001: A Speaker Odyssey - The Speaker Recognition Workshop (ISCA)*, Crete, Greece, June 2001, pp. 213–218.

[29] B. Xiang, U. Chaudhari, J. Navratil, G. Ramaswamy, and R. Gopinath, "Short-time Gaussianization for robust speaker verification," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Orlando, Florida, USA, May 2002, pp. 681–684.

[30] H. A. Murthy, F. Beaufays, L. P. Heck, and M. Weintraub, "Robust text-independent speaker identification over telephone channels," *IEEE Trans. on Speech and Audio Processing*, vol. 7(5), pp. 554–568, September 1999.

[31] R. Teunen, B. Shahshahani, and L. P. Heck, "A model-based transformational approach to robust speaker recognition," in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, Beijing, China, October 2000, pp. 495–498.

[32] P. Kenny and P. Dumouchel, "Disentangling speaker and channel effects in speaker verification," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, Quebec, Canada, May 2004, pp. 37–40.

[33] ——, "Experiments in speaker verification using factor analysis likelihood ratios," in *Proc. Odyssey: the Speaker and Language Recognition Workshop (Odyssey 2004)*, Toledo, Spain, June 2004, pp. 219–226.

[34] M. J. F. Gales and S. Young, "Hmm recognition in noise using parallel model combination," in *Proc. Third European Conf. on Speech Communication and Technology (Eurospeech)*, Berlin, Germany, September 1993, pp. 837–840.

[35] T. Matsui, T. Kanno, and S. Furui, "Speaker recognition using hmm composition in noisy environments," *Computer Speech and Language*, vol. 10, pp. 107–116, 1996.

[36] L. P. Wong and M. Russell, "Text-dependent speaker verification under noisy conditions using parallel model combination," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, Utah, USA, May 2001, pp. 457–460.

[37] S. Sagayama, Y. Yamaguchi, S. Takahashi, and J. Takahashi, "Jacobian approach to fast acoustic model adaptation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Munich, Germany, April 1997, pp. 835–838.

[38] C. Cerisara, L. Rigaziob, and J. C. Junqua, "α-Jacobian environmental adaptation,," *Speech Communication*, vol. 42, no. 1, pp. 25–41, 2004.

[39] 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.

[40] C. Barras and J. L. Gauvain, "Feature and score normalization for speaker verification of cellular data," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, Hong Kong, April 2003, pp. 49–52.

[41] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[42] S. Ahmed and V. Tresp, "Some solutions to the missing feature problem in vision," in *Advances in Neural Information Processing Systems (NIPS)*, San Mateo, California 1993, pp. 393–400.

[43] C. Guillemot and O. L. Meur, "Image inpainting: Overview and recent advances," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 127–144, January 2014.

[44] M. Cooke, P. Green, and M. Crawford, "Handling missing data in speech recognition," in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, Yokohama, Japan, September 1994, pp. 1555–1558.

[45] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound.* Cambridge, MA: MIT Press, 1990.

[46] M. Cooke and D. P. W. Ellis, "The auditory organization of speech and other sources in listeners and computational models," *Speech Communication*, vol. 35, no. 3-4, pp. 141–177, 2001.

[47] 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.

[48] B. Raj, R. Singh, and R. Stern, "Inference of missing spectrographic features for robust automatic speech recognition," in *Proc. Int. Conf. on Speech and Language Processing (ICSLP)*, Sydney, Australia, November 1998, pp. 1491–1494.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] J. Barker, M. Cooke, and P. Green, "Robust asr based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise," in *Proc. in European Conference on Speech Communication and Technology (Eurospeech)*, Aalborg, Denmark, September 2001, pp. 213–216.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] M. El-Maliki and A. Drygajlo, "Missing features detection and estimation for robust speaker verification," in *Proc. in European Conference on Speech Communication and Technology (Eurospeech)*, Budapest, Hungary, September 1999, pp. 975–978.

[64] 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.

[65] 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.

[66] 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.

[67] C. Cerisara, S. Demange, and J.-P. Haton, "On noise masking for automatic missing data speech recognition: A survey and discussion," *Computer Speech and Language*, vol. 21(3), pp. 443–457, 2007.

[68] D. Donoho, "Compressed sensing," *IEEE Trans. on Information Theory*, vol. 52(4), pp. 1289–1306, April 2006.

[69] E. Candés, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. on Information Theory*, vol. 52(2), pp. 489–509, February 2006.

[70] B. Fauve, N. Evans, and J. Mason, "Improving the performance of text-independent short duration SVM- and GMM-based speaker verification," in *Proc. Odyssey: the Speaker and Language Recognition Workshop (Odyssey 2008)*, South Africa, January 2008.

[71] R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," in *Proc. Int. Conf. on Spoken Language Processing (Interspeech)*, Brisbane, Australia, September 2008, pp. 853–856.

[72] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "i-vector based speaker recognition on short utterances," in *Proc. Int. Conf. on Spoken Language Processing (Interspeech)*, Florence, Italy, August 2011, pp. 2341–2344.

[73] V. Hautamäki, Y.-C. Cheng, P. Rajan, and C.-H. Lee, "Minimax i-vector extractor for short duration speaker verification," in *Proc. Int. Conf. on Spoken Language Processing (Interspeech)*, Lyon, France, August 2013, pp. 3708–3712.

[74] J.-P. Suh and J. H. L. Hansen, "Test token driven acoustic balancing for sparse enrollment data in cohort GMM speaker recognition," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aalborg, Denmark, August 2010, pp. 572–575.

[75] T. Stadelmann and B. Freisleben, "Dimension-decoupled Gaussian mixture model for short utterance speaker recognition," in *Proc. International Conf. on Pattern Recognition (ICPR)*, Istanbul, Turkey, August 2010, pp. 1602–1605.

[76] R. Saeidi, A. Hurmalainen, T. Virtanen, and D. A. van Leeuwen, "Exemplar-based sparse representation and sparse discrimination for noise robust speaker identification," in *Proc. Odyssey: the Speaker and Language Recognition Workshop (Odyssey 2012)*, Singapore, June 2012.

[77] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Group sparsity for speaker identity discrimination in factorisation-based speech recognition," in *in Proc. Int. Conf. on Spoken Language Proc. (INTERSPEECH 2013)*, Lyon, France, August 2013.

[78] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Trans. on Multimedia*, vol. 9(7), pp. 1396–1403, November 2007.

[79] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing.* Prentice-Hall, Englewood Cliffs, NJ, USA, 1989.

[80] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Speaker identification under noisy environments by using harmonic structure extraction and reliable frame weighting," in *in Proc. Int. Conf. on Spoken Language Proc. (INTERSPEECH)*, Pittsburgh, Pennsylvania, USA, September 2006.

[81] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15(5), pp. 1711–1723, July 2007.

[82] H. Aronowitz and D. Burshtein, "Efficient speaker recognition using approximated cross entropy (ACE)," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15(7), pp. 2033–2043, September 2007.

[83] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 13(4), pp. 345–354, May 2005.

[84] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *in Proc. Int. Conf. on Spoken Language Proc. (INTERSPEECH 2011)*, Florence, Italy, August 2011, pp. 249–252.

[85] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 31(2), pp. 210–227, February 2009.

[86] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. on Information Theory*, vol. 53(12), pp. 4655–4666, December 2007.

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

[88] 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.

[89] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19(7), pp. 2067–2080, September 2011.

[90] T. N. Sainath, B. Ramabhadran, M. Picheny, D. Nahamoo, and D. Kanevsky, "Exemplar-based sparse representation features: From TIMIT to LVCSR," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19(8), pp. 2598–2613, November 2011.

[91] G. S. V. S. Sivaram, S. K. Nemala, M. Elhilali, T. D. Tran, and H. Hermansky, "Sparse coding for speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, Texas, USA, March 2010, pp. 4346–4349.

[92] O. Vinyals and L. Deng, "Are sparse representations rich enough for acoustic modeling?" in *in Proc. Int. Conf. on Spoken Language Proc. (INTERSPEECH 2012)*, Portland, USA, September 2012.

[93] T. N. Sainath, D. Nahamoo, B. Ramabhadran, D. Kanevsky, V. Goel, and P. M. Shah, "Exemplar-based sparse representation phone identification features," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 4492–4495.

[94] I. Naseem, R. Togneri, and M. Bennamoun, "Sparse representation for speaker identification," in *Proc. International Conf. on Pattern Recognition (ICPR)*, Istanbul, Turkey, August 2010, pp. 4460–4463.

[95] B. C. Haris and R. Sinha, "Sparse representation over learned and discriminatively learned dictionaries for speaker verification," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, March 2012, pp. 4785–4788.

[96] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Maryland, USA, June 2011, pp. 1697–1704.

[97] 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.

[98] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. on Signal Processing*, vol. 54(11), pp. 4311–4322, November 2006.

[99] 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.

[100] 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.

[101] 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.

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

[103] 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.

[104] 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.

[105] 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.

[106] E. J. Candés and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98(6), pp. 925–936, June 2010.

[107] M. Fazel, H. Hindi, and S. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proc. American Control Conf.*, June 2001, pp. 4734–4739.

[108] S. Becker, E. J. Candés, and M. Grant, "Templates for convex cone problems with applications to sparse signal recovery," *Mathematical Programming Computation*, vol. 3, no. 3, pp. 165–218, 2011.

[109] K. C. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pacific Journal of Optimization*, pp. 615–640, 2010.

[110] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," UIUC Technical Report UILU-ENG-09-2215, October 2009.

[111] R. H. Keshavan and S. Oh, "A gradient descent algorithm on the Grassman manifold for matrix completion," arXiv:0910.5260, 2009.

[112] W. Dai, O. Milenkovic, and E. Kerman, "Subspace evolution and transfer (SET) for low-rank matrix completion," *IEEE Trans. on Signal Processing*, vol. 59, no. 7, pp. 3120–3132, July 2011.

[113] L. Balzano, R. Nowak, and B. Recht, "Online identification and tracking of subspaces from highly incomplete information," in *Proc. Annual Allerton Conf. on Communication, Control and Computing*, October 2010, pp. 704–711.

[114] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos, "Sparse bayesian methods for low-rank matrix estimation," *IEEE Trans. on Signal Processing*, vol. 60, no. 8, pp. 3964–3977, August 2012.

[115] H. Ji, C. Liu, Z. Shen, and Y. Xu, "Robust video denoising using low rank matrix completion," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, June 2010, pp. 1791–1798.

[116] A. Waters and V. Cevher, "Distributed bearing estimation via matrix completion," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, Texas, USA, March 2010, pp. 2590–2593.

[117] S. Nikitaki, G. Tsagkatakis, and P. Tsakalides, "Efficient training for fingerprint based positioning using matrix completion," in *Proc. European Signal Proc. Conf. (EUSIPCO)*, Bucharest, Romania, August 2012.

[118] D. L. Sun and R. Mazumder, "Non-negative matrix completion for bandwidth extension: A convex optimization approach," in *Proc. IEEE Conf. on Machine Learning for Signal Processing (MLSP)*, Southampton, UK, September 2013.

[119] R. Parhizkar, A. Karbasi, S. Oh, and M. Vetterli, "Calibration using matrix completion with application to ultrasound tomography," *IEEE Trans. on Signal Processing*, vol. 61, no. 20, pp. 4923–4933, October 2013.

[120] A. Y. Aravkin, R. Kumar, H. Mansour, B. Recht, and F. J. Herrmann, "A robust SVD-free approach to matrix completion, with applications to interpolation of large scale data," arXiv:1302.4886, 2013.

[121] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. Int. Conf. on Machine Learning*, Haifa, Israel, June 2010, pp. 663–670.

[122] Y. Zhang, Z. Jiang, and L. S. Davis, "Learning structured low-rank representations for image classification," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Portland, Oregon, USA, June 2013, pp. 676–683.

[123] D. Pei, F. Sun, and H. Liu, "Supervised low-rank matrix recovery for traffic sign recognition in image sequences," *IEEE Signal Proc. Letters*, vol. 20, no. 3, pp. 241–244, March 2013.

[124] Y. Panagakis and C. Kotropoulos, "Automatic music tagging by low-rank representation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, March 2012, pp. 497–500.

[125] Y. H. Yang, "Low-rank representation of both singing voice and music accompaniment via learned dictionaries," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, Curitiba, Brazil, November 2013.

[126] G. Tsagkatakis and P. Tsakalides, "Dictionary based reconstruction and classification of randomly sampled sensor network data," in *Proc. Sensor Array and Multichannel Signal Proc. Workshop*, Hoboken, NJ, June 2012, pp. 117–120.

[127] M. Grant and S. Boyd, "CVX: Matlab Software for Disciplined Convex Programming, version 1.21," 2011.

[128] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course.* Springer Netherlands, 2004.

[129] P. L. Combettes and V. R. Wajs, “Signal recovery by proximal forward-backward splitting,” *SIAM Multiscale Model. Simul.*, vol. 4, no. 4, pp. 1168–1200, 2005.