ΠΡΟΣ

1) **Ολα τα μέλη ΔΕΠ του Τμήματος Επιστήμης Υπολογιστών**
2) **Τους εκπροσώπους των Μεταπτυχιακών φοιτητών του Τμήματος Επιστήμης Υπολογιστών**
3) **Την Επταμελή Εξεταστική Επιτροπή**
4) **Ολα τα μέλη της Πανεπιστημιακής Κοινότητας**

Πρόσκληση σε Δημόσια Παρουσίαση της Διδακτορικής Διατριβής του

κ. Τζαγκαράκη Χρήστου

Την Τρίτη, 1 Ιουλίου 2014 και ώρα 17:00 στην αίθουσα Τηλεδιάσκεψης Κ206 του Τμήματος Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης στο Ηράκλειο, θα γίνει η δημόσια παρουσίαση και υποστήριξη της Διδακτορικής Διατριβής του υποψηφίου διδάκτορος του Τμήματος Επιστήμης Υπολογιστών κ. Τζαγκαράκη με θέμα:

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

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

ΠΕΡΙΛΗΨΗ

Η *αναγνώριση ομιλητή* αποτελεί τη διαδικασία της αυτόματης αναγνώρισης του ατόμου που μιλάει, με βάση κάποια χαρακτηριστικά που εξάγονται από το σήμα φωνής. 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.

Στην προσπάθειά μας να αντιμετωπίσουμε τις παραπάνω δυσκολίες, 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 τη 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*. Τα βέλτιστα εκτιμώμενα αραιά βάρη των γραμμικών συνδυασμών, οι επονομαζόμενοι και *αραιοί κώδικες*, που προκύπτουν ως λύσεις ενός 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. Οι εκτενείς συγκρίσεις που πραγματοποιήθηκαν τόσο με πιθανοτικά μοντέλα, τα οποία βασίζονται στην υπόθεση ότι τα χαρακτηριστικά της φωνής ακολουθούν γενικευμένη Γκαουσιανή κατανομή, όσο και με μερικές εκ των κορυφαίων μεθόδων ταξινόμησης, όπως μοντέλα μίξης Γκαουσιανών κατανομών και κοινής παραγοντικής ανάλυσης, ανέδειξαν την υπεροχή της προτεινόμενης μεθόδου.

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. Η ανακατασκευή των μη αξιόπιστων φασματικών περιοχών πραγματοποιείται μέσω του Singular Value Thresholding (SVT) αλγορίθμου, 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. Κατά τη διάρκεια της πειραματικής αξιολόγησης η προτεινόμενη μέθοδος συγκρίνεται με την ευρέως χρησιμοποιούμενη τεχνική της αραιής συμπλήρωσης, αναδεικνύοντας την ισχύ της στον υπολογισμό αξιόπιστων χαρακτηριστικών.

Επίσης, 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.

Επόπτης Διδακτορικής Διατριβής: Επίκουρος Καθηγητής, Αθανάσιος Μουχτάρης

Abstract

Speaker recognition is the process of recognizing a speaker automatically, based on specific features extracted from the speech signal. 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.

Focusing on overcoming the above limitations, 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". The optimally estimated sparse weights of the linear combinations, the so-called *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.

Extending the 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. 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 the proposed method.

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 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.

Finally, 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.

Supervisor: Assistant Professor Athanasios Mouchtaris