ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Πολλαπλών όψεων συνδυασμός ακουστικών χαρακτηριστικών με χαρακτηριστικά παραγωγής ομιλίας για αναγνώριση φωνημάτων στη βάση δεδομένων rtMRI-TIMIT

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Ιωάννη Κ. Δούρου

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Αθήνα, Μάιος 2017

Εθνικο Μετσοβιο Πολυτεχνειο
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Πολλαπλών όψεων συνδυασμός ακουστικών χαρακτηριστικών με χαρακτηριστικά παραγωγής ομιλίας για αναγνώριση φωνημάτων στη βάση δεδομένων rtMRI-TIMIT

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Ιωάννη Κ. Δούρου

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 29η Μαΐου 2017.

.....	.....	.....
Πέτρος Μαραγκός	Γεράσιμος Ποταμιάνος	Κωνσταντίνος Τζαφέστας
Καθηγητής	Αναπληρωτής Καθηγητής	Επίκουρος Καθηγητής
Ε.Μ.Π.	Παν/μίου Θεσσαλίας	Ε.Μ.Π.

Αθήνα, Μάιος 2017

.....

Ιωάννης Κ. Δούρος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον καθηγητή Πέτρο Μαραγκό για την εμπιστοσύνη που μου έδειξε, καθώς και για τη δυνατότητα που μου έδωσε να εκπονήσω την παρούσα διπλωματική εργασία κάτω από την καθοδήγησή του στο Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων.

Επιπροσθέτως, θα ήθελα να εκφράσω τις ευχαριστίες μου στο Νάσο Κατσαμάνη για τη διάθεση που επέδειξε, το χρόνο που αφιέρωσε και τις γνώσεις και την έμπνευση που μου μετέδωσε κατά την προσπάθεια μου.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου, τους φίλους μου, τους δασκάλους μου καθώς και όλους τους κοντινούς μου ανθρώπους για την αμέριστη στήριξή τους σε κάθε μου βήμα.

Περίληψη

Σε αυτήν τη διπλωματική εργασία ερευνούμε τη χρήση πληροφοριών άρθρωσης, και πιο συγκεκριμένα δεδομένων rt-MRI της φωνητικής οδού, για τη βελτίωση της απόδοσης αναγνώρισης ομιλίας. Για τον σκοπό των πειραμάτων μας χρησιμοποιούμε δεδομένα από τη βάση δεδομένων rtMRI-TIMIT. Αρχικά, τα χαρακτηριστικά SIFT εξάγονται για κάθε πλαίσιο του βίντεο. Έπειτα οι SIFT περιγραφείς του κάθε πλαισίου μετασχηματίζονται σε ένα μεμονωμένο ιστόγραμμα ανά εικόνα, με χρήση της μεθοδολογίας Bag of Visual Words. Εφόσον αυτό το είδος πληροφοριών άρθρωσης είναι δύσκολο να εξαχθεί σε μια τυπική εγκατάσταση αναγνώρισης ομιλίας, θεωρούμε ότι είναι διαθέσιμο μόνο στο στάδιο της εκπαίδευσης. Συνεπώς χρησιμοποιούμε μια προσέγγιση πολλαπλών όψεων με εφαρμογή canonical correlation analysis (CCA) σε οπτικά και ηχητικά δεδομένα. Με χρήση του πίνακα μετασχηματισμού που εξήχθη κατά τη διάρκεια του σταδίου εκπαίδευσης, μετασχηματίζουμε τα ηχητικά δεδομένα της εκπαίδευσης και της δοκιμής για να παράγουμε τα τελικά χαρακτηριστικά (συνδυασμός ακουστικών χαρακτηριστικών με χαρακτηριστικά του συστήματος παραγωγής ομιλίας) τα οποία αποτελούν την είσοδο του συστήματος αναγνώρισης. Τα πειραματικά αποτελέσματα επιδεικνύουν βελτιώσεις στην αναγνώριση φωνής συγκριτικά με την χρήση μόνο ακουστικών χαρακτηριστικών.

Λέξεις κλειδιά: χαρακτηριστικά παραγωγής ομιλίας, προσέγγιση πολλαπλών όψεων, αναγνώριση φωνημάτων, κρυφά Μαρκοβιανά μοντέλα, ανάλυση κανονικής συσχέτισης, μηχανές διανυσματικής υποστήριξης, χαρακτηριστικά SIFT, βάση δεδομένων rtMRI-TIMIT, Bag of Visual Words, SMOTE

Abstract

In this thesis, we investigate the use of articulatory information, and more specifically rt-MRI data of the vocal tract, to improve speech recognition performance. For the purpose of our experiments, we use data from the rtMRI-TIMIT database. Firstly, SIFT features are extracted for each video frame. Afterwards, the SIFT descriptors of each frame are transformed to a single histogram per picture, by using the Bag of Visual Words methodology. Since this kind of articulatory information is difficult to acquire in typical speech recognition setups we only consider it to be available in the training phase. Thus, we use a multi-view setup approach by applying canonical correlation analysis (CCA) to visual and audio data. By using the transformation matrix, acquired during the training stage, we transform both train and test audio data to produce MFCC-articulatory features, which form the input for the recognition system. Experimental results demonstrate improvements in phone recognition in comparison with the audio-based baseline.

Keywords: articulatory features, multi-view approach, phonetic recognition, hidden Markov models, canonical correlation analysis, support vector machines, SIFT features, rtMRI-TIMIT database, Bag of Visual Words, SMOTE

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

Τα αυτόματα συστήματα αναγνώρισης ομιλίας (ASR) έχουν πολλές εφαρμογές στην καθημερινή μας ζωή που ποικίλουν από την απομαγνητοφώνηση ηχογραφημένης ομιλίας έως την εκτέλεση φωνητικών εντολών από συσκευές. Παρόλο που τα σύγχρονα συστήματα αναγνώρισης μπορούν να φτάσουν πολύ υψηλές επιδόσεις με ακρίβεια μεγαλύτερη του 95% η αποτελεσματικότητα τους μειώνεται δραματικά όταν το περιβάλλον ηχογράφησης δεν είναι ιδανικό, για παράδειγμα ένα δωμάτιο με αρκετό θόρυβο. Η μειωμένη απόδοση σε μέρη με αρκετό θόρυβο κάνει τα υπάρχοντα συστήματα αναγνώρισης ακατάλληλα για εφαρμογή στην καθημερινή μας ζωή. Στόχος μας είναι να βελτιώσουμε τις επιδόσεις και την ευρωστία ενός συστήματος αυτόματης αναγνώρισης ομιλίας αξιοποιώντας γνώση για το σύστημα παραγωγής της ανθρώπινης φωνής.

Ο λόγος του αυξημένου ενδιαφέροντος της επιστημονικής κοινότητας για τα ASR συστήματα είναι το εύρος των πεδίων που μπορούν να εφαρμοστούν. Για παράδειγμα:

- Άτομα με ειδικές ανάγκες: Ολικά η εν μέρει κωφοί μπορούν να ωφεληθούν παρά πολύ από τα προηγμένα συστήματα ASR. Για παράδειγμα, τους βοηθούν στην επικοινωνία σε περίπτωση που βρεθούν σε μέρη με πολύ κόσμο όπως μια αίθουσα συνεδριάσεων ή μια διάλεξη. Αυτό επιτυγχάνεται μετατρέποντας την ομιλία σε κείμενο έτσι ώστε το άτομο με ειδικές ανάγκες να μπορεί να διαβάσει και να καταλάβει τον ομιλητή. Επίσης, άτομα με αναπηρία που δεν μπορούν να χρησιμοποιήσουν τα χεριά τους, χρησιμοποιούν συστήματα ASR για να μπουν στο διαδίκτυο, να ελέγξουν έναν υπολογιστή ή ακόμα και να γράψουν.

- Στις εγκαταστάσεις Υγείας: Τα συστήματα ASR χρησιμοποιούνται κατά κόρον από τους γιατρούς για να κρατούν σημειώσεις σχετικές με τις παρατηρήσεις τους και να ενημερώνουν τους φακέλους των ασθενών. Συγκεκριμένα, υπάρχουν δυο περιπτώσεις. Στην πρώτη ο γιατρός ηχογραφεί την αναφορά του και ένα σύστημα ASR είναι υπεύθυνο να μετατρέψει την ηχογραφημένη αναφορά στον γραπτό φάκελο του ασθενή. Στην δεύτερη περίπτωση ο γιατρός μπορεί να εισάγει προκαθορισμένες επιλογές σε ένα ASR σύστημα έτσι ώστε αυτό να συμπληρώσει κάποιες φόρμες με προκαθορισμένες τιμές.

- Στον στρατό: Τα ASR συστήματα χρησιμοποιούνται κυρίως στα μαχητικά αεροπλάνα όπου τα χεριά του πιλότου είναι απασχολημένα και δεν μπορούν να γράψουν, να ρυθμίσουν συχνότητες στον ασύρματο, να ορίσουν στόχο ή συντεταγμένες. Επίσης μπορεί να χρησιμοποιηθεί και για την εκτέλεση πολλών άλλων εντολών όπως το να πυροβολήσει ένα όπλο. Υπάρχουν όμως και αρκετές δυσκολίες στην χρήση ASR στα μαχητικά αεροσκάφη. Το πιο σοβαρό από αυτά είναι ο εκκωφαντικός θόρυβος που μειώνει σημαντικά την αποτελεσματικότητα της αναγνώρισης ομιλίας κυρίως στα ελικόπτερα, οπού ο πιλότος δεν φορά ειδικό κράνος για την μείωση του θορύβου από το εξωτερικό περιβάλλον.

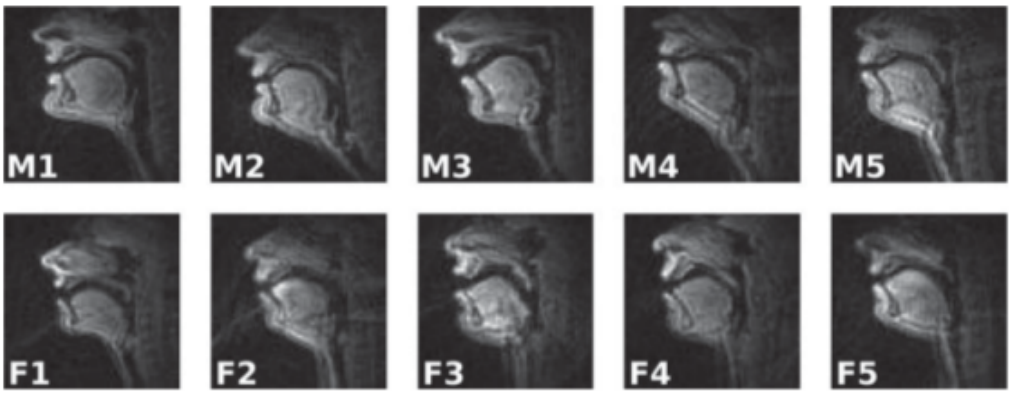
- Χρήση στην καθημερινότητα: Η αναγνώριση ομιλίας είναι ένα από τα βασικά χαρακτηριστικά των συσκευών του μέλλοντος. Ξεκινώντας από τα smart-phones τα οποία μπορούν να δεχθούν φωνητικές εντολές ώστε να πραγματοποιήσουν κλήσεις ή να χρησιμοποιήσουν την φωνητική αναζήτηση όρων/φράσεων, η ιδέα αυτή έχει επεκταθεί και σε αλλά πεδία όπως, για παράδειγμα, στα αυτοκίνητα στα οποία ο οδηγός μπορεί να εισάγει και να προβάλλει χάρτες, διαδρομές, βενζινάδικα κτλ. Αυτή η ιδέα μπορεί να επεκταθεί ακόμα παραπάνω και να χρησιμοποιηθεί και σε αλλά αντικείμενα μέσα σε ένα σπίτι όπως για τον έλεγχο των φώτων ή της τηλεόρασης.

- Μετάφραση: Αναγνώριση ομιλίας μπορεί να χρησιμοποιηθεί και σε συσκευές μετάφρασης από ομιλία σε ομιλία. Τέτοιες συσκευές θα καθιστούσαν δυνατή την επικοινωνία μεταξύ ανθρώπων από διαφορετικές χώρες. Η χρήση τέτοιων συσκευών θα μπορούσε να γίνεται από απλούς τουρίστες ως και σε αίθουσες συνεδριάσεων.

Ο βασικός στόχος αυτής της διπλωματικής είναι να διερευνήσει τρόπους βελτίωσης ενός συστήματος αναγνώρισης φωνής με την αξιοποίηση πληροφορίας και γνώσης που έρχεται από το ανθρώπινο σύστημα παραγωγής ομιλίας.

Στα πλαίσια της παρούσας εργασίας επιλέξαμε να χρησιμοποιήσουμε δυναμικές Μαγνητικές τομογραφίες (rtMRI scans) για να συγκεντρώσουμε πληροφορίες παράγωγης ομιλίας. Σε αντίθεση με τις στατικές MRI εικόνες στις οποίες ο ομιλητής κρατάει σε σταθερή διάταξη τα μέρη της φωνητικής του οδού μέχρι να γίνει η σάρωση, στις δυναμικές MRI εικόνες ο ομιλητής μιλάει κανονικά κατά τη διάρκεια της καταγραφής. Ένα ακόμα πλεονέκτημα των (rtMRI scans) έναντι άλλων μεθόδων απόκτησης μαγνητικών τομογραφιών (π. χ. cineMRI scans), είναι ότι ο ομιλητής απαιτείται να κάνει τις εκφωνήσεις μόνο μία φορά καθ΄ όλη τη διάρκεια της καταγραφής. Καθώς όλη η εργασία έγινε με χρήση δυναμικών μαγνητικών τομογραφιών, όπου αναφέρουμε από εδώ και στο εξής μαγνητικές τομογραφίες εννοούμε ότι είναι δυναμικές.

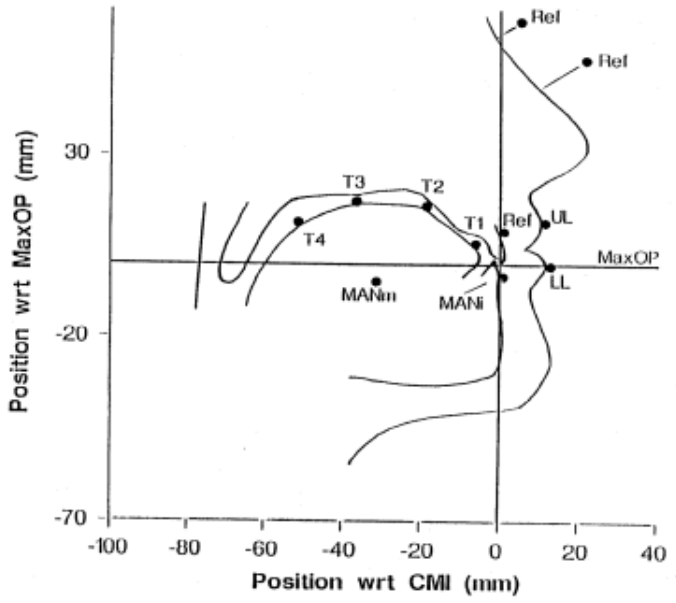
Ο λόγος που επιλέξαμε να χρησιμοποιήσουμε μαγνητικές τομογραφίες είναι γιατί οι μαγνητικές τομογραφίες (MRI scans) έχουν περισσότερα πλεονεκτήματα συγκριτικά με άλλες τεχνικές απεικόνισης (π. χ. Ακτινογραφίες (X-Ray scans)) όπως είναι, για παράδειγμα, η αυξημένη διακριτική ικανότητα των μαλακών ιστών [1]. Επιπλέον οι Μαγνητικές Τομογραφίες λαμβάνουν πληροφορίες για όλο τον λάρυγγα με περισσότερη ακρίβεια κάτι το οποίο άλλες μέθοδοι απεικόνισης δεν καταφέρνουν.



Σχήμα 1.1: Εικόνες rtMRI από τους δέκα ομιλητές της USC-TIMIT database (πάνω σειρά: άνδρες, κάτω σειρά: γυναίκες). [1]

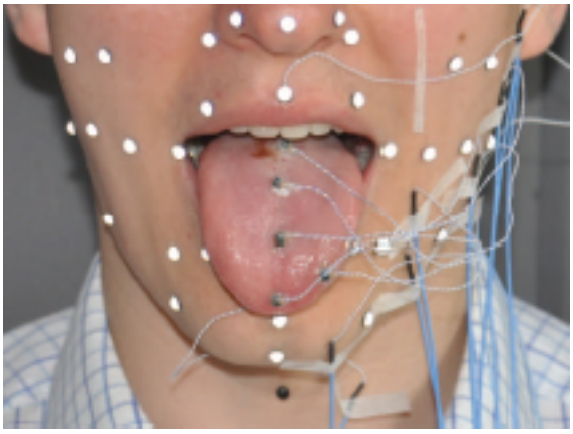
Συνεπώς, αξιοποιώντας τη γνώση από τις μαγνητικές τομογραφίες, μπορούμε να εξάγουμε χαρακτηριστικά που βασίζονται στη διαδικασία παραγωγής λόγου, λαμβάνοντας υπόψιν σε πολύ μεγάλο βαθμό, όχι μόνο τη δυναμική κατά τη διαδικασία ομιλίας, αλλά και το σύνολο των μερών της φωνητικής οδού. Επιπλέον, η ύπαρξη πολλών ομιλητών τόσο ανδρών όσο και γυναικών από διάφορες περιοχές, προσφέρουν τη δυνατότητα να αξιοποιήσουμε γνώση σχετικά με τη διακύμανση αφενός στα χωρικά χαρακτηριστικά της φωνητικής οδού των ομιλητών και αφετέρου, στις διάφορες ιδιαιτερότητες στην προφορά της γλώσσας (Αγγλική) που παρουσιάζεται σε διάφορες περιοχές.

Εκτός από τις μαγνητικές τομογραφίες, υπάρχουν και άλλες τεχνικές συλλογής δεδομένων παραγωγής ομιλίας. Μία από αυτές, είναι η τεχνική X-ray microbeam (XRMB), στην οποία μικρά σφαιρίδια από χρυσό προσαρτώνται σε διάφορα σημεία του προσώπου και της φωνητικής οδού (Σχήμα 1.3). Έπειτα, χρησιμοποιούνται δέσμες ακτίνων X προκειμένου να ανιχνευθούν οι κινήσεις των σφαιριδίων κατά τη διάρκεια που ο ομιλητής μιλάει [2].



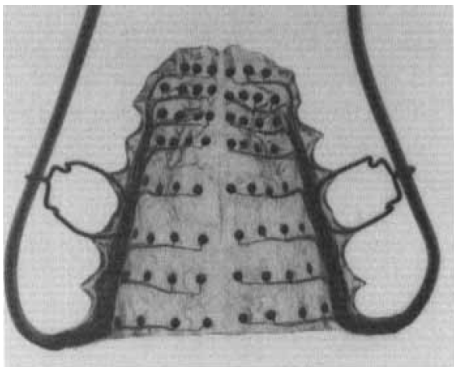
Σχήμα 1.2: Θέση των σφαιριδίων XRMB στο στόμα και το πρόσωπο του ομιλητή. [2]

Μία άλλη διαδεδομένη τεχνική είναι η Electromagnetic Articulography (EMA) κατά την οποία τοποθετούνται αισθητήρες στη γλώσσα και markers σε διάφορα σημεία του προσώπου. Στη συνέχεια χρησιμοποιείται ένα μαγνητικό πεδίο για να ανιχνεύσει την κίνηση των markers κατά τη διαδικασία παραγωγής ομιλίας. Η διάταξη των αισθητήρων στο πρόσωπο του ομιλητή φαίνεται στο Σχήμα 1.3 [3] [22].



Σχήμα 1.3: Θέση των αισθητήρων EMA και των markers στο πρόσωπο του ομιλητή. [3]

Ένας άλλος τρόπος καταγραφής δεδομένων παραγωγής ομιλίας είναι με χρήση της τεχνικής Electropalatography (EPG) όπου μία συσκευή με ηλεκτρόδια (Σχήμα 1.4) τοποθετείται στον ουρανίσκο του ομιλητή και καταγράφει τις επαφές που έχει η γλώσσα του με τα διάφορα σημεία της συσκευής καθώς μιλάει [4].



Σχήμα 1.4: Συσκευή για την καταγραφή πληροφορίας παραγωγής λόγου με χρήση της τεχνικής EPG [4]

Στην έρευνά μας εφαρμόζουμε ένα setup πολλαπλών όψεων (multi-view) για την αξιοποίηση της πληροφορίας από τις μαγνητικές τομογραφίες. Κάναμε πειράματα στη βάση δεδομένων rt-MRI TIMIT (αναλυτική περιγραφή της βάσης στο κεφάλαιο 4) και βελτιώσαμε το λάθος αναγνώρισης κατά περίπου 2% (από 67.06 και 63.74 σε 65.21 και 61.73 αντίστοιχα για τους δύο ομιλητές).

Συνοπτικά, χρησιμοποιήσαμε τα SIFT (Scale-Invariant Feature Transform) χαρακτηριστικά για την περιγραφή κάθε frame στο βίντεο. Εφαρμόζοντας τη μέθοδο Bag of Visual Words μετατρέπουμε αυτούς τους descriptors σε ένα ιστόγραμμα ανά

εικόνα. Εξάγουμε MFCCs τα οποία μαζί με τα articulatory- ακουστικά ιστογράμματα είναι οι δυο όψεις του πειράματος μας. Τέλος εφαρμόζουμε ένα multi- view setup χρησιμοποιώντας CCA. Τα πειραματικά αποτελέσματα παρουσιάζουν βελτιώσεις στην αναγνώριση φωνής σε σύγκριση με την κλασική audio-based προσέγγιση.

Η δομή αυτής της εργασίας είναι η ακόλουθη. Στο κεφάλαιο 2 παρουσιάζουμε την σχετική ερευνά που πραγματοποιήθηκε τα τελευταία χρόνια στον τομέα. Στο 3ο κεφάλαιο εξηγούμε το πως λειτουργεί ένα τυπικό σύστημα αναγνώρισης ομιλίας καθώς και τις μεθόδους που χρησιμοποιούνται. Στα κεφάλαια 4 και 5 περιγράφουμε τις μεθόδους που χρησιμοποιήσαμε στο σύστημα μας καθώς και τα αντίστοιχα πειράματα. Τέλος στο κεφάλαιο 6 συζητούμε τα αποτελέσματα της ερευνάς μας και δίνουμε μερικές κατευθύνσεις για μελλοντική έρευνα.

Κεφάλαιο 2

Σχετική έρευνα

Σε αυτό το κεφάλαιο θα αναφέρουμε τις διάφορες μεθόδους που έχουν αναπτυχθεί τα τελευταία χρόνια για τη εκπαίδευση των συστημάτων αναγνώρισης ομιλίας. Όπως έχουμε αναφέρει, τα συστήματα αναγνώρισης ομιλίας επιτυγχάνουν ιδιαίτερα υψηλά αποτελέσματα. Ωστόσο υπάρχει ακόμα χώρος για βελτίωση, ειδικά όταν το ακουστικό περιβάλλον δεν είναι ιδανικό, για παράδειγμα όταν υπάρχει πολύς θόρυβος ή αντήχηση.

Η βελτίωση των αποτελεσμάτων επιτυγχάνεται κυρίως με τη βελτίωση των χρησιμοποιούμενων χαρακτηριστικών ή με βελτίωση της αρχιτεκτονικής του συστήματος. Στο κεφάλαιο αυτό γίνεται μία προσπάθεια αρχικά να περιγράψουμε διάφορες προσεγγίσεις οπτικών χαρακτηριστικών και χαρακτηριστικών άρθρωσης και στη συνέχεια να περιγράψουμε δημοφιλείς αρχιτεκτονικές συστημάτων.

2.1 Οπτική πληροφορία και πληροφορία άρθρωσης

Παρακάτω γίνεται μία περιγραφή των διαφόρων ειδών πληροφορίας που μπορεί να χρησιμοποιήσει κάποιος έτσι ώστε, σε συνδυασμό με την ηχητική πληροφορία, να πετύχει βελτίωση του αποτελέσματος της αναγνώρισης.

Αρκετό ενδιαφέρον παρουσιάζεται γύρω από articulatory πληροφορίες υπό την μορφή π. χ. Electromagnetic Articulography (EMA), X-ray Microbeam (XRMB), and EPG του λάρυγγα και το πως μπορούν να ωφεληθούν οι τεχνολογίες ομιλίας [23], [24], [25]. Κάθε ένας από αυτούς τους τρόπους συλλογής δεδομένων έχει τα δικά του πλεονεκτήματα και μειονεκτήματα, τα οποία παρουσιάζουμε συνοπτικά παρακάτω (περιγραφή του τρόπου λειτουργίας κάθε μεθόδου γίνεται στο κεφάλαιο 1).

Χρησιμοποιώντας τις τεχνικές EMA, XRMB, μπορούμε να επιτύχουμε πολύ υψηλή συχνότητα δειγματοληψίας, περίπου $400Hz$ με την EMA και $160Hz$ με τηνXRMB, συνδυάζοντας ταυτόχρονα και υψηλή ακρίβεια στις ληφθέντες μετρήσεις. Το κύριο μειονέκτημα αυτών των μεθόδων είναι η αδυναμία καταγραφής πληροφοριών για την συμπεριφορά των υπολοίπων μερών της φωνητικής οδού, όπως ο λάρυγγας ή ο φάρυγ-

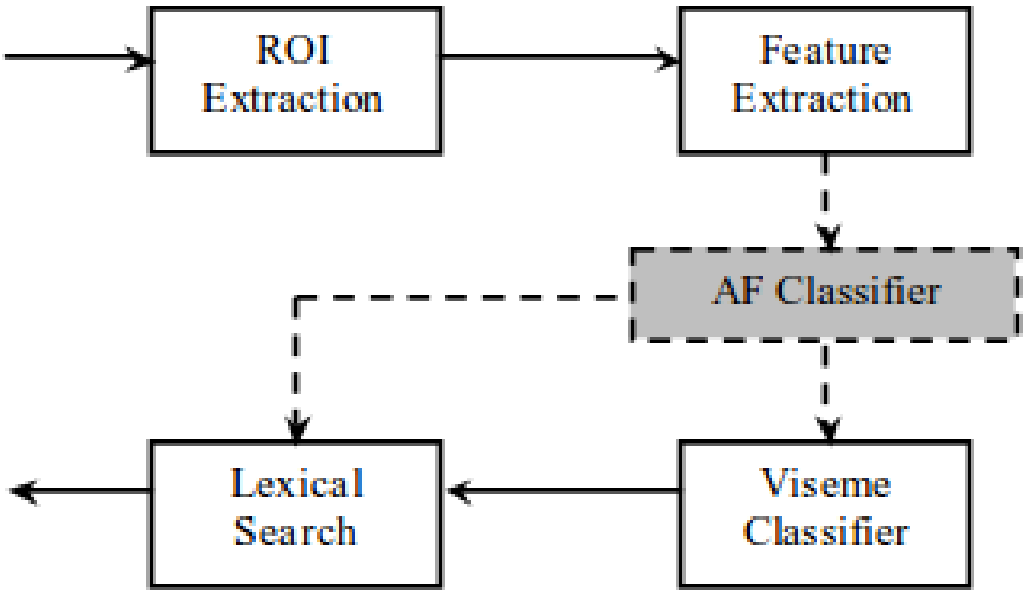
γας, τα οποία παίζουν και αυτά σημαντικό ρόλο στη διαδικασία παραγωγής λόγου [3], [2].

Τα ίδια με αυτά που περιγράφηκαν παραπάνω ισχύουν περίπου και για την τεχνική EPG. Η συχνότητα δειγματοληψίας είναι αρκετά υψηλή για επαρκής περιγραφή των φαινομένων παραγωγής λόγου ($200Hz$). Ωστόσο, εκτός από την αδυναμία καταγραφής όλων των μερών της φωνητικής οδού, αυτή η μέθοδος παρουσιάζει το επιπλέον μειονέκτημα ότι μπορεί να παρέχει μόνο πληροφορία για τα σημεία που αγγίζει η γλώσσα πάνω στην ειδική συσκευή και τίποτα περισσότερο [4].

Εκτός από τη χρήση articulatory χαρακτηριστικών, έχουν προταθεί και άλλες προσεγγίσεις για την αύξηση της ευρωστίας των συστημάτων αναγνώρισης ομιλίας. Αρκετές βασίζονται στην εξαγωγή διαφόρων ειδών οπτικών χαρακτηριστικών από το πρόσωπο του ομιλητή, όπου σε συνδυασμό με τον ήχο, παρέχουν στο προς εκπαίδευση σύστημα όλη την πληροφορία που έχει ένας ακροατής κατά τη διάρκεια μιας ομιλίας. Η ιδέα πίσω από αυτές τις προσεγγίσεις, είναι ότι αφού ο άνθρωπος είναι ικανός να αντιληφθεί τον συνομιλητή του με αυτές τις πληροφορίες αυτές είναι αρκετές για την εκπαίδευση των συστημάτων.

Οπτικά χαρακτηριστικά του προσώπου όπως Cosine Transform (DCT), Discrete Wavelet Transform (DWT), και Active Appearance Model coefficients σε συνδυασμό με ηχητικά χαρακτηριστικά χρησιμοποιήθηκαν σε οπτικοακουστικά συστήματα αναγνώρισης για να μειώσουν τα σφάλματα [26, 27]. Η χρήση των active appearance models γίνεται ώστε να αναγκάσει μια μάσκα με ορισμένα σημεία η απόσταση των οποίων αλλάζει για να χωρέσει το σχήμα του κεφαλιού. Η θέση των σημείων που προκύπτει από αυτή τη διαδικασία μας δίνει τα οπτικά χαρακτηριστικά. Στην αναγνώριση χρησιμοποιούνται δυο active appearance models. Το πρώτο χρησιμοποιεί σαν είσοδο τα χαρακτηριστικά όλων των σημείων του προσώπου ενώ το άλλο μόνο τα σημεία γύρο από το στόμα. το στόμα υποδεικνύεται χρησιμοποιώντας μια έκλειψη της οποίας ο πρώτος άξονας υποδεικνύει το πλάτος του στόματος ενώ ο δεύτερος το πόσο ανοιχτό είναι.

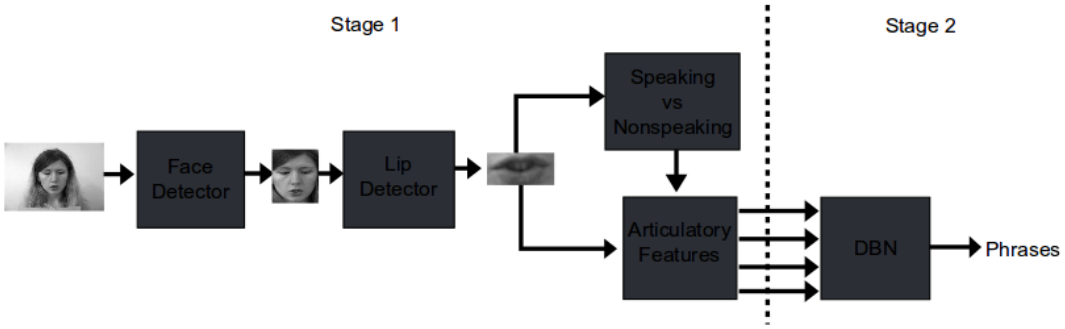
Στο [5], δίνεται περισσότερη έμφαση στην εξαγωγή articulatory χαρακτηριστικών. Εικόνες από την περιοχή του στόματος χρησιμοποιούνται για τα πειράματα. Γισ κάθε φώνημα καταγράφεται στο αντίστοιχο viseme. Ορίζονται τρεις κατηγορίες: κατηγορία βάσει εμφάνισης, κατηγορία βάσει σχήματος και ένας συνδυασμός των δυο. Η πρώτη κατηγορία περιλαμβάνει πληροφορίες της περιοχής ενώ η δεύτερη του σχήματος της όπως την απόσταση των χειλιών. Ξεχωριστοί ταξινομητές Support Vector Machine (SVM) εκπαιδεύονται για κάθε κατηγορία και ύστερα συγκρίνουν τις αποφάσεις τους στην test περίοδο. Η δομή του συστήματος παρατηρείται παρακάτω.



Σχήμα 2.1: Articulatory χαρακτηριστικά με χρήση SVM ταξινομητών. [5]

Μια άλλη ιδέα παρουσιάζεται στο [6]. Εικόνες του προσώπου του ομιλητή χρησιμοποιούνται σαν είσοδος. Το σύστημα βρίσκει το πρόσωπο και αναγνωρίζει τα χείλη του ομιλητή. Βασισμένο στα χείλη κάθε frame ταξινομείται είτε σαν speaking είτε σαν non-speaking. Για κάθε frame εξάγονται articulatory χαρακτηριστικά. Αυτά βασίζονται μόνο στο σχήμα των χειλιών και είναι διακριτά (σε μερικά frame ίσως υπάρχουν πληροφορίες σχετικά με την γλώσσα οι οποίες όμως δεν χρησιμοποιούνται) Έως αυτό το σημείο έχουν χρησιμοποιηθεί μόνο ταξινομητές SVM. Μέτα το βήμα της εξαγωγής χαρακτηριστικών ένα Dynamic Bayesian Network χρησιμοποιείται για περαιτέρω αναγνώριση. Το σχηματικό διάγραμμα του συστήματος παρατηρείται παρακάτω.

Μια ακόμη έρευνα στην οποία αξίζει να αναφερθούμε είναι [28] παρόλο που δεν συνδέεται άμεσα με την δική μας. Όπως τονίζεται στο [29], υπάρχει μεγάλος συσχετισμός μεταξύ του τι βλέπει και τι λέει κάποιος. Στο [28], χρησιμοποιούν οπτικές πληροφορίες του ομιλητή για να αυξήσουμε την επίδοση της αναγνώρισης ομιλίας. Σε αντίθεση με άλλες έρευνες, σε αυτήν την έρευνα οι οπτικές πληροφορίες είναι το αντικείμενο το οποίο βλέπει ο ομιλητής. Χρησιμοποιώντας ένα kinet βρήκαν το που βρίσκεται ο ομιλητής και που κοιτάζει. Οι ερευνητές χρησιμοποιούν τις ιδιότητες του αντικείμενου (όπως σχήμα, χρώμα κτλ.) το οποίο κοιτάζει ο ομιλητής για να



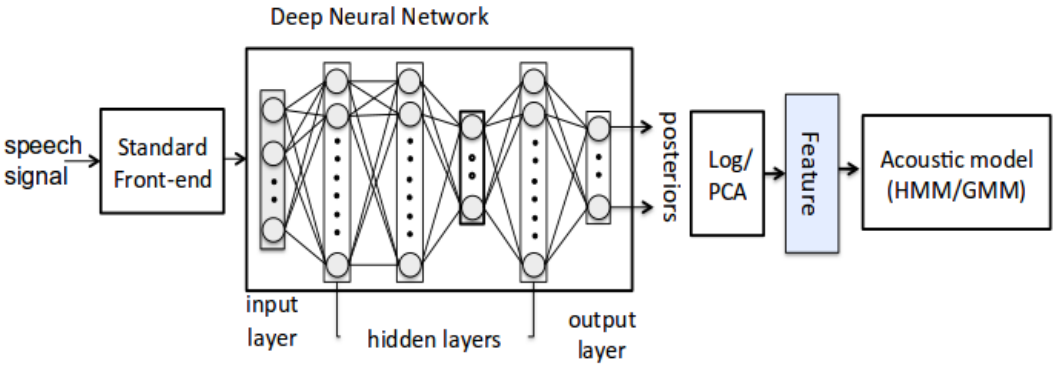
Σχήμα 2.2: Οπτικά χαρακτηριστικά βασισμένα στα χείλη του ομιλητή με χρήση SVM ταξινομητών. [6]

προκαταλάβουν το γλωσσικό μοντέλο του συστήματος αναγνώρισης τους.

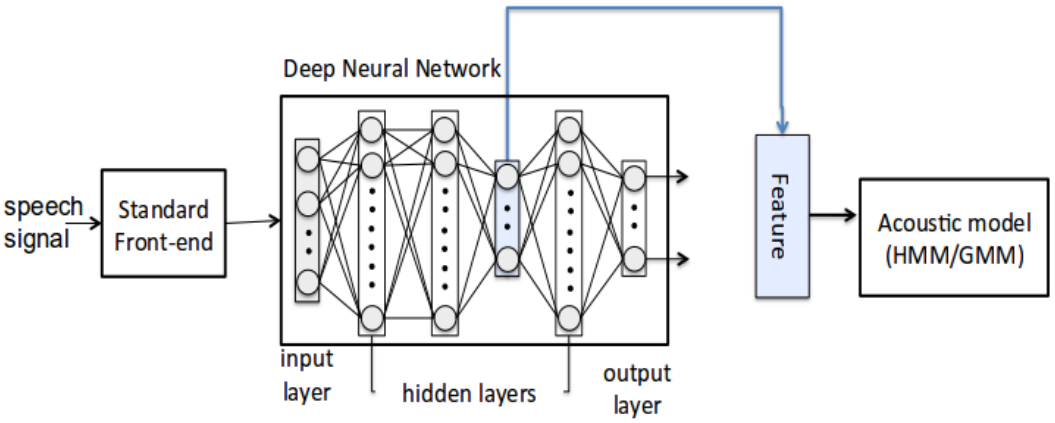
2.2 Αρχιτεκτονικές συστημάτων

Σε αυτό το σημείο θα αναφέρουμε μερικές εργασίες οι οποίες χρησιμοποιούν ενδιαφέρουσες προσεγγίσεις σε ότι αφορά την εκπαίδευση των συστημάτων. Ο λόγος που το κάνουμε αυτό είναι γιατί, συνήθως, μια πιο περίπλοκη αρχιτεκτονική (που τις περισσότερες περιπτώσεις αποτελεί ένα συνδυασμό απλούστερων ιδεών) έχει βελτιωμένα αποτελέσματα σε σχέση με μια πιο απλή, καθώς επιτυγχάνει καλύτερη μοντελοποίηση του προβλήματος.

Μια ενδιαφέρουσα εργασία [30] χρησιμοποιεί ένα Υβριδικό σύστημα Artificial Neural Network (ANN)/ Hidden Markcv Model (HMM) το οποίο συνδυάζει ακουστικές και articulatory πληροφορίες για την αναγνώριση ομιλίας σε θορυβώδες περιβάλλον. Η κύρια ιδέα είναι ότι διαφορετικές αναπαραστάσεις των ίδιων πληροφοριών παρά το ότι τα σφάλματα αναγνώρισης κάθε αναπαράστασης ξεχωριστά είναι περίπου τα ίδια, έχουν διαφορετικούς τύπους σφαλμάτων. Έτσι όταν συνδυάζονται το τελικό ποσοστό σφαλμάτων μειώνεται. Μια ακόμη ενδιαφέρουσα εργασία είναι [31], [32], οπού χρησιμοποιείται ένας συνδυασμός από νευρωνικά δίκτυα και Gaussian mixture models (tandem setup) για αυξημένα αποτελέσματα αναγνώρισης ομιλίας. Στο tandem setup ένα νευρωνικό δίκτυο εκπαιδεύεται για να αναγνωρίσει το σετ φωνημάτων. Η έξοδος του νευρωνικού δικτύου (NN) είναι ένα διάνυσμα με διαστάσεις ίσες των φωνημάτων το οποίο περιλαμβάνει τις πιθανότητες κάθε φωνήματος. Αυτό το διάνυσμα εισάγεται σε ένα HMM για να εκπαιδευτεί το Gaussian mixture model.



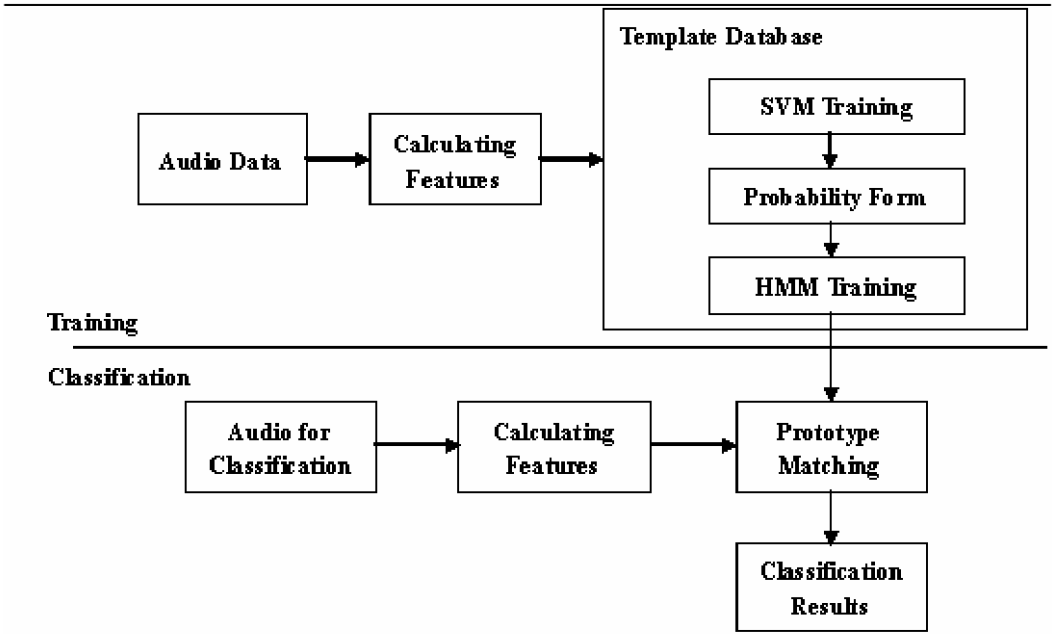
Σχήμα 2.3: Σύστημα Tandem με χρήση εκ των υστέρων πιθανότητας. [7]



Σχήμα 2.4: Σύστημα Tandem με χρήση bottleneck χαρακτηριστικών. [7]

Μία άλλη ευρέως διαδεδομένη αρχιτεκτονική, είναι τα βαθιά νευρωνικά δίκτυα (Deep Neural Networks (DNNs)). Παρόλο που τα DNNs απαιτούν γενικά μεγάλο όγκο δεδομένων για να καταγράψουν επιτυχώς τους συσχετισμούς μεταξύ χαρακτηριστικών [33], υπάρχουν έρευνες που μας προτείνουν τρόπους να εκπαιδεύσουμε DNNs με περιορισμένες πήγες χρησιμοποιώντας πολύγλωσσα δεδομένα. [34].

Άλλες ιδέες όπως υβριδικά HMM-SVM μοντέλα για ταξινόμηση ήχου έχουν επίσης χρησιμοποιηθεί με επιτυχία [35], [36]. Η δομή του HMM-SVM είναι η ακόλουθη: Εξάγουμε χαρακτηριστικά από τα ηχητικά δεδομένα, Μετά ένας ταξινομητης multi-label SVM εκπαιδεύεται για να διακρίνει τις κλάσεις. Κάθε ταξινομημένο αποτέλεσμα μετατρέπεται σε μια πιθανότητα ανάλογα με την σιγουριά ότι το αποτέλεσμα που έχει προβλεφθεί είναι όντως σωστό. Αυτές οι πιθανότητες εισάγονται σε ένα HMM. Το HMM εκπαιδεύεται χρησιμοποιώντας τους κλασσικούς αλγορίθμους.



Σχήμα 2.5: Υβριδικό SVM/HMM μοντέλο. [7]

Ορισμένα συστήματα που παρουσιάζουν ενδιαφέρον είναι τα συστήματα που είναι βασισμένα στην προσέγγιση multi- view [37]. Η κύρια ιδέα είναι να χρησιμοποιήσουμε διαφορετικά είδη μετρήσεων συγκεντρωμένα την ίδια ώρα για τον ίδιο σκοπό, με στόχο να χρησιμοποιήσουμε μια από αυτές για την αποτελεσματική μετατροπή των μετρήσεων της άλλης. Συνήθως χρησιμοποιούνται δυο μετρήσεις αλλά αυτό δεν είναι απόλυτο. Συνηθισμένες όψεις για την αναγνώριση ομιλίας είναι ήχος με οπτικά η articulatory χαρακτηριστικά. Μια άλλη επιλογή είναι να χρησιμοποιήσουμε τις ίδιες τις ετικέτες

αλλά αυτό δεν είναι σύνηθες στην πράξη. Σε αντίθεση με τα multi- modal setups, που απαιτούν δύο όψεις τόσο κατά τη διάρκεια της εκπαίδευσης, όσο και κατά τη διάρκεια των δοκιμών, τα multi-view setups μπορούν να διαχειριστούν δεδομένα με δυο όψεις, μια εκ των οποίων είναι διαθέσιμη μόνο κατά τη διάρκεια των δοκιμών (ενώ κατά τη διάρκεια της εκπαίδευσης είναι και οι δύο διαθέσιμες). Συνήθως χρησιμοποιείτε Canonical Correlation Analysis (CCA) για να υπολογίσουμε τους μετασχηματισμούς.

Παρόμοια setup έχουν χρησιμοποιηθεί με επιτυχία στην αναγνώριση ομιλίας όπως στο [38] στο οποίο χρησιμοποιείται η βάση δεδομένων XRMB. Σε αυτή την εργασία χρησιμοποιήθηκαν πολλές μέθοδοι για να συνδυαστούν οπτικοακουστικά χαρακτηριστικά όπως MFCCA (Mel Frequency Cepstral Coefficients & CCA), Non-Consolidating Components Analysis (NCCA) με πολλαπλές μετατροπές όπως PCA. Όλα τα χαρακτηριστικά δοκιμαστήκαν κάτω από δυο συστήματα, το πρώτο είναι kNN-based ενώ το άλλο multi- label SVM- based.

Στο κεφάλαιο αυτό παρουσιάσαμε συνοπτικά βασικές μεθόδους απόκτησης οπτικής πληροφορίας (DWT) και πληροφορία άρθρωσης ((XRMB), (EMA)) καθώς και ορισμένες αρχιτεκτονικές συστημάτων αναγνώρισης ομιλίας ((ANN/HMM), (DNNs), (SVM/HMM)) που χρησιμοποιούνται για την βελτίωση των αποτελεσμάτων αναγνώρισης.

Κεφάλαιο 3

Υλοποίηση των συστημάτων αυτόματης αναγνώρισης ομιλίας

Σε αυτό το κεφάλαιο θα περιγράψουμε τους βασικούς αλγορίθμους καθώς και το θεωρητικό υπόβαθρο στο οποίο είναι βασισμένο ένα ASR σύστημα. Θα παρουσιάσουμε επίσης δημοφιλείς ιδέες για την υλοποίηση ενός ASR συστήματος και θα εξηγήσουμε πως όλα αυτά συνδυάζονται για να δημιουργήσουν το τελικό σύστημα στην πράξη. Ένα κλασσικό σύστημα βασίζεται στα λεγόμενα κρυφά Μαρκοβιανά μοντέλα (Hidden Markov Model (HMM)) τα οποία είναι στατιστικά μοντέλα και χρησιμοποιούνται για τη μοντελοποίηση σημάτων. Παρακάτω θα αναφερθούμε εκτενώς στο πως λειτουργούν αυτά τα μοντέλα, πως εκπαιδεύονται αλλά και εναλλακτικές προσεγγίσεις για το πρόβλημα της αναγνώρισης ομιλίας. Επίσης θα περιγράψουμε διαφόρων ειδών χαρακτηριστικά (π. χ. MFCCs), μεθόδους (π. χ. LDA) και προσεγγίσεις (π. χ. SGMM) που χρησιμοποιούνται ευρέως καθώς και τις βελτιώσεις που επιφέρουν αυτές στα συστήματα αναγνώρισης ομιλίας.

3.1 Κρυφά Μαρκοβιανά Μοντέλα (HMM)

Το HMM είναι ένα στατιστικό μοντέλο το οποίο μας βοηθά να περιγράψουμε γεγονότα που αλλάζουν δυναμικά στο πεδίο του χρόνου όπως τα σήματα ομιλίας [39]. Το HMM χρησιμοποιείται κυρίως για τη μοντελοποίηση γεγονότων στα οποία οι ένέργειες΄ είναι κρυφές αλλά τα αποτελέσματα από αυτές τις πράξεις είναι γνωστά σε εμάς. Το κύριο πρόβλημα που λύνει το HMM είναι η εκτίμηση των ένεργειών΄ που έγιναν,όταν είναι γνωστά τα επερχόμενα αποτελέσματα (παρατηρήσεις). Ένα κλασσικό παράδειγμα είναι το πρόβλημα των δοχείων με αντικατάσταση. Ας υποθέσουμε ότι υπάρχει ένας μεγάλος αριθμός από μπάλες διαφορετικών χρωμάτων M, χωρισμένων (όχι απαραίτητα εξίσου) σε N δοχεία. Ένα άτομο επιλέγει μια μπάλα από ένα δοχείο, χωρίς εμείς να μπορούμε να δούμε το δοχείο από το οποίο επιλέχθηκε η μπάλα, μας

ενημερώνει σχετικά με το χρώμα της μπάλας και την επανατοποθετεί στο ίδιο δοχείο. Ξανά επιλέγει ένα δοχείο,παίρνει μια μπάλα, μας λέει το χρωμα κ.ο.κ. Εμείς τώρα θα πρέπει να εκτιμήσουμε τη σειρά των δοχείων που ήταν κρυμμένα, γνωρίζοντας μόνο τη σειρά των χρωμάτων που μας ανακοινώθηκε. Χρησιμοποιώντας το HMM για την επίλυση του παραπάνω προβλήματος,είναι προφανές ότι πρέπει να προσδιοριστούν οι τιμές πέντε παραμέτρων:

1. Ο αριθμός των καταστάσεων N, που αντιστοιχεί στον αριθμό των δοχείων. Για πρακτικές εφαρμογές, συνήθως υπάρχουν ενδείξεις για τον αριθμό των καταστάσεων, ανάλογα τη φύση του προβλήματος έτσι ώστε κάθε κατάσταση να έχει λόγο ύπαρξης, παρά το γεγονός ότι οι καταστάσεις είναι κρυμμένες. Οι καταστάσεις είναι labeled με τιμές από το 1 μέχρι το N και η κατάσταση σε χρόνο t συμβολίζεται ως q_t.

2. Ο αριθμός όλων των πιθανών παρατηρήσεων M ανά κατάσταση. Στο παράδειγμά μας, το M ισούται με τον αριθμό των διαφορετικών χρωμάτων μπάλας. Συμβολίζουμε τις διαφορετικές τιμές των παρατηρήσεων ανά κατάσταση ως $V = \{v_1, v_2, ..., v_M\}$. Κάθε στοιχείο του V αντιστοιχεί στη φυσική έξοδο του συστήματος που μοντελοποιείται.

3. Η κατανομή πιθανότητας της μετάβασης καταστάσεων

$$a_{ij} = P[q_{t+1} = j | q_t = i], \qquad 1 \leq i, j \leq N$$

που δίνει την πιθανότητα μετάβασης στην κατάσταση j δεδομένου ότι η παρούσα κατάσταση είναι η i.

Στην περίπτωσή μας, το A δίνει την πιθανότητα το άτομο να επιλέξει το j-στο δοχείο δεδομένου ότι προηγουμένως είχε επιλέξει το i-στο δοχείο. Το A είναι ένας NxN πίνακας. Το άθροισμα των στοιχείων κάθε σειράς ισούται με 1και οι τιμές όλων των στοιχείων a_{ij} είναι μεταξύ 0 και 1, συμπεριλαμβανομένων και των δύο.

$$\sum_{j=1}^{N} a_{ij} = 1$$

$$0 \leq a_{ij} \leq 1$$

Το μοντέλο αυτό λέγεται εργοδικό επειδή από κάθε κατάσταση, η μετάβαση σε κάθε άλλη κατάσταση είναι επιτρεπτή.

4. Η κατανομή πιθανότητας των παρατηρήσεων $B = \{b_{ij}\}$ όπου

$$b_{ij} = P[o_t = v_j | q_t = i], \qquad 1 \leq j \leq M, \qquad 1 \leq i \leq N$$

Στο υπό μελέτη πρόβλημα, το B αντιστοιχεί στην πιθανότητα το άτομο, δοθέντος του δοχείου i που επέλεξε, να επιλέξει μια μπάλα j χρώματος. Το B είναι ένας $N \times M$ πίνακας,με ιδιότητες ίδιες με αυτές που περιγράφησαν για τον πίνακα A.

$$\sum_{j=1}^{M} b_{ij} = 1$$

$$0 \leq b_{ij} \leq 1$$

5. Ο πίνακας κατανομής αρχικής κατάστασης π_i που δίνει την πιθανότητα για κάθε κατάσταση να επιλεγεί ως αρχική κατάσταση για τη μοντελοποιημένη διαδικασία. Στο παράδειγμά μας, ο π_i δίνει την πιθανότητα το άτομο να επιλέξει το $i-th$ δοχείο πρώτο. Όπως είναι προφανές,ο π είναι ένας $N \times 1$ πίνακας όπου

$$\sum_{i=1}^{N} \pi_{ij} = 1$$

Ο συμβολισμός $\lambda = (A, B, \pi)$ γενικά χρησιμοποιείται σαν ολοκληρωμένη περιγραφή των παραμέτρων του μοντέλου.

3.1.1 Χρήση των HMMs στην αναγνώριση φωνής

Για αναγνώριση φωνής, χρησιμοποιούνται Gaussian Mixtures Models για τη μοντελοποίηση των πιθανοτήτων B κάθε διαφορετική παρατήρηση j να προέρχεται από κάθε κατάσταση i, αντί για συγκεκριμένους αριθμούς. Για κάθε κατάσταση έχουμε

$$b_i(o_t) = \sum_{k=1}^{M} c_{ik} \mathcal{N}(o_t, \mu_{ik}, U_{ik}), \qquad 1 \leq i \leq N$$

όπου

$$\sum_{k=1}^{M} c_{ik} = 1$$

και c_{ik}, μ_{ik}, U_{ik} αντιστοιχούν στο συντελεστή,τη μέση τιμή και τον πίνακα συνδιακύμανσης της κάθε Gaussian αντίστοιχα.

Επιπλέον, τα HMMs που χρησιμοποιούνται στις περισσότερες περιπτώσεις είναι left-right (Bakis) μοντέλα,που σημαίνει ότι από κάθε κατάσταση i, μόνο μεταβάσεις σε κατάσταση j με $j \geq i$ είναι επιτρεπτές. Με άλλα λόγια,τα left-right HMMs είναι αυτά για τα οποία ο πίνακας μεταβάσεων A είναι άνω τριγωνικός και ο πίνακας κατανομής αρχικής κατάστασης $\pi = [1, 0, 0, ..., 0]'$. Τα μοντέλα left-right ταιριάζουν ιδιαίτερα για αναγνώριση φωνής γιατί μπορούν να περιγράψουν γεγονότα που είναι σε χρονική σειρά,ακόμα και αν υπάρχουν μερικές τυχαίες παραλλαγές,όπως ακριβώς είναι και οι λέξεις.

Στη συνέχεια παρουσιάζεται μία απλή, εισαγωγική αντιμετώπιση του προβλήματος της αναγνώρισης ομιλίας προκειμένου ο αναγνώστης να κατανοήσει την αρχή λειτουργίας βασικών αλγορίθμων. Αναλυτική περιγραφή της δομής και λειτουργίας των σύγχρονων συστημάτων αναγνώρισης ομιλίας παρουσιάζεται στο τέλος της ενότητας.

Για μια απλή λέξη (ή φώνημα) το σύστημα αναγνώρισης φωνής δουλεύει όπως εξηγείται στα παρακάτω βήματα:

- Κάθε φωνητικό σήμα από μια δοσμένη λέξη W αναπαριστάται ως μια σειρά από διανύσματα. Ένα codebook με M μοναδικά διανύσματα χρησιμοποιείται για τον καθορισμό του φωνητικού σήματος στο δείκτη του διανύσματος του κοντινότερου codebook στο διάνυσμα αναπαράστασης του χρόνου. Για κάθε λέξη W στο λεξιλόγιό μας εκπαιδεύουμε ένα ξεχωριστό HMM N καταστάσεων χρησιμοποιώντας ως παρατηρήσεις την αναπαράσταση του σήματος.

- Διαφορετικές τιμές των παραμέτρων εξετάζονται, όπως οι καταστάσεις των HMMs και το μέγεθος του codebook, για τη βελτίωση της ικανότητας των HMMs να μοντελοποιούν την ακολουθία λέξεων.

- Όταν μια άγνωστη λέξη πρέπει να αναγνωριστεί,η πιθανότητα υπολογίζεται για κάθε μοντέλο, δοθέντων των παρατηρήσεων της εξεταζόμενης λέξης. Η εξεταζόμενη λέξη κατηγοριοποιείται ως η λέξη W της οποίας το μοντέλο έχει την υψηλότερη πιθανότητα.

Ωστόσο,για κάθε βήμα πρέπει να λυθεί και ένα πρόβλημα (3 προβλήματα στο σύνολο) έτσι ώστε να εφαρμοστούν τα HMMs σε πρακτικές εφαρμογές όπως περιγράφησαν παραπάνω:

- Πώς προσαρμόζουμε στο μοντέλο τις παραμέτρους μοντελοποίησης $\lambda = (A, B, \pi)$ για να μεγιστοποιήσουμε την $P(O|\lambda)$;

- Πώς θα βρούμε τη 'βέλτιστη' ακολουθία καταστάσεων $Q = q_1, q_2, ..., q_T$ που σχετίζεται με μια δοθείσα ακολουθία παρατηρήσεων $O = \{o_1, o_2, ..., o_T\}$ και ένα μοντέλο $\lambda = (A, B, \pi)$;

- Δοθείσας της ακολουθίας παρατηρήσεων $O = \{o_1, o_2, ..., o_T\}$ και του μοντέλου $\lambda = (A, B, \pi)$, πώς υπολογίζουμε αποτελεσματικά την $P(O|\lambda)$, την πιθανότητα της ακολουθίας παρατηρήσεων,δοθέντος του μοντέλου·

Διάφοροι αλγόριθμοι έχουν φτιαχτεί για την επίλυση των προβλημάτων που περιγράφησαν. Παρουσιάζουμε συνοπτικά τους πιο δημοφιλείς στις παρακάτω ενότητες [40].

3.1.2 Αλγόριθμος Forward and backward

Οι αλγόριθμοι Forward and Backward μας βοηθούν στο να υπολογίσουμε αποτελεσματικά την πιθανότητα $P(O|\lambda)$. Η κύρια ιδέα του αλγόριθμου Forward είναι ότι μόνο το πρώτο μέρος της ακολουθίας παρατηρήσεων χρησιμοποιείται για τον υπολογισμό της πιθανότητας και, μόλις αυτή υπολογιστεί, με χρήση των πινάκων μετάβασης και εκπομπής υπολογίζουμε την πιθανότητα της προηγούμενης ακολουθίας παρατηρήσεων συν την επόμενη παρατήρηση. Η διαδικασία συνεχίζεται έως ότου χρησιμοποιηθεί όλο το μήκος της ακολουθίας παρατηρήσεων. Ο αλγόριθμος λειτουργεί ως εξής:

Έστω ότι

$$\alpha_t(i) = P(o_1 o_2 ... o_t, q_t = i|\lambda)$$

είναι η πιθανότητα των πρώτων t παρατηρήσεων σε κατάσταση i τη χρονική στιγμή t, δοθέντων των παραμέτρων του μοντέλου λ.

Μπορούμε να υπολογίσουμε την $P(O|\lambda)$ εφαρμόζοντας τα παρακάτω βήματα:

1.
$$\alpha_1(i) = \pi_i b_i(o_1), \qquad 1 \leq i \leq N$$

2.
$$\alpha_{t+1} = \sum_{i=1}^{N} \alpha_t(i) a_{ij} b_j(o_{t+1}), \qquad 1 \leq t \leq T-1, \qquad 1 \leq j \leq N$$

3.
$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

Η κύρια ιδέα του αλγορίθμου Backward είναι περίπου ίδια με αυτή του αλγορίθμου Forward με μια βασική διαφορά. Στον Backward αλγόριθμο χρησιμοποιούμε το τελευταίο μέρος της ακολουθίας παρατηρήσεων και μετά από κάθε απαιτούμενο υπολογισμό προσθέτουμε το προηγούμενο μέρος της ακολουθίας στην ακολουθία που χρησιμοποιήθηκε για την εκτίμηση της πιθανότητας. Συνεπώς ορίζουμε

$$\beta_t(i) = P(o_{t+1} o_{t+2} ... o_T|q_t = i, \lambda)$$

όπου β είναι η πιθανότητα των τελευταίων $(T-t)$ παρατηρήσεων σε κατάσταση i τη χρονική στιγμή t, δοθέντων των παραμέτρων μοντέλου λ. Ο αλγόριθμος λειτουργεί ως εξής:

1.
$$\beta_T(i) = 1, \qquad 1 \leq i \leq N$$

2.

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), \qquad 1 \leq i \leq N, \qquad t = T-1, T-2, ..., 1$$

Και για τους δύο αλγορίθμους,το κόστος υπολογισμού είναι της τάξης N^2T

3.1.3 Αλγόριθμος Viterbi

Ο αλγόριθμος Viterbi χρησιμοποιείται για την επιλογή της 'βέλτιστης' ακολουθίας καταστάσεων όταν μας δίνονται οι παρατηρήσεις και το μοντέλο. Το κριτήριο βελτιστοποίησης του αλγορίθμου είναι η μεγιστοποίηση του αναμενόμενου αριθμού σωστών μεμονωμένων καταστάσεων. Υπάρχουν ορισμένα προβλήματα που μπορεί να προκύψουν με χρήση αυτού του κριτηρίου, για παράδειγμα η ακολουθία καταστάσεων που προκύπτει από τον αλγόριθμο Viterbi ίσως να μην είναι έγκυρη επειδή η μετάβαση από την κατάσταση q_t στην q_{t+1} μπορεί να έχει μηδενική πιθανότητα. Παρόλο που υπάρχουν λύσεις για τέτοιου είδους προβλήματα, ο αλγόριθμος Viterbi χρησιμοποιείται ακόμα στην πλειοψηφία των εφαρμογών.

Ορίζουμε $\delta_t(i)$ το καλύτερο score κατά μήκος μιας διαδρομής και $\psi_t(j)$ την κατάσταση που μεγιστοποιεί το $\delta_t(i)$. Ο αλγόριθμος αποτελείται από τα ακόλουθα βήματα:

1.

$$\delta_1(i) = \pi_i b_i(o_1), \qquad 1 \leq i \leq N$$

$$\psi_1(i) = 0$$

2.

$$\delta_t(j) = \max_{1 \leq i \leq N} [\delta_{t-1} a_{ij}] b_j(o_t)$$

$$\psi_t(j) = \arg \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}]$$

όπου $1 \leq j \leq N$, $2 \leq t \leq T$ για τις δύο τελευταίες εξισώσεις.

3.

$$P^* = \max_{1 \leq i \leq N} [\delta_T(i)]$$

$$q_T^* = \arg \max_{1 \leq i \leq N} [\delta_T(i)]$$

4.

$$q_T^* = \psi_{t+1}(q_{t+1}^*), \qquad t = T-1, T-2, ..., 1$$

3.1.4 Αλγόριθμος Baum-Welch

Ο αλγόριθμος Baum-Welch χρησιμοποιείται για εκπαίδευση HMM. Σκοπός του είναι να βρεί εκτιμήσεις για τους πίνακες A, B, π τέτοιες ώστε η πιθανότητα των ακολουθιών των παρατηρήσεων $P(O|\lambda)$ να μεγιστοποιείται. Βασίζεται στον αλγόριθμο expectation-maximization (EM) για να υπολογίσει τις παραμέτρους ενός HMM. Για το πρόβλημα αυτό δεν υπάρχει αναλυτική λύση, συνεπώς ο αλγόριθμος μπορεί ενδεχομένως να συγκλίνει σε τοπικό μέγιστο.

Περιγράφουμε συνοπτικά τη βασική ιδέα του αλγορίθμου. Αρχικά, οι πίνακες A, B, π αρχικοποιούνται με τυχαίες τιμές. Για την εκτίμηση του A, οι πιθανότητες των παρατηρήσεων υπολογίζονται και αθροίζονται για να δώσουν το τελικό άθροισμα. Στη συνέχεια υπολογίζουμε,προσθέτουμε και διαιρούμε με το τελικό άθροισμα τις πιθανότητες των ακολουθιών παρατηρήσεων που αρχικά ήταν στην κατάσταση S_i και μετα πήγαν στην κατάσταση S_j. Αυτός ο αριθμός τοποθετείται στη θέση (i, j) ενός πίνακα A_{NEW}. Η διαδικασία επαναλαμβάνεται για όλα τα ζευγάρια $1 \leq i, j \leq N$ και μόλις ο πίνακας A_{NEW} συμπληρωθεί, τον κανονικοποιούμε με τέτοιο τρόπο ώστε το άθροισμα κάθε σειράς να ισούται με 1. Για τον πίνακα B θεωρούμε ότι η παρατήρηση O_j προήλθε από την κατάσταση S_i και ξανά υπολογίζουμε, προσθέτουμε και διαιρούμε την πιθανότητα της ακολουθίας παρατηρήσεων για όλα τα ζευγάρια $1 \leq i \leq N, 1 \leq j \leq M$. Ο πίνακας B_{NEW} είναι επίσης κανονικοποιημένος. Για τον υπολογισμό των αρχικών πιθανοτήτων υποθέτουμε ότι όλες οι καταστάσεις ξεκινούν με την κρυφή κατάσταση S_i και υπολογίζουμε την υψηλότερη πιθανότητα. Η διαδικασία πάλι επαναλαμβάνεται για όλα τα $1 \leq i \leq N$. Έπειτα ο πίνακας π_{NEW} κανονικοποιείται και αυτός. Τέλος, επαναλαμβάνουμε όλα τα βήματα που περιγράφησαν παραπάνω μέχρι οι προκύπτουσες πιθανότητες να συγκλίνουν ικανοποιητικά.

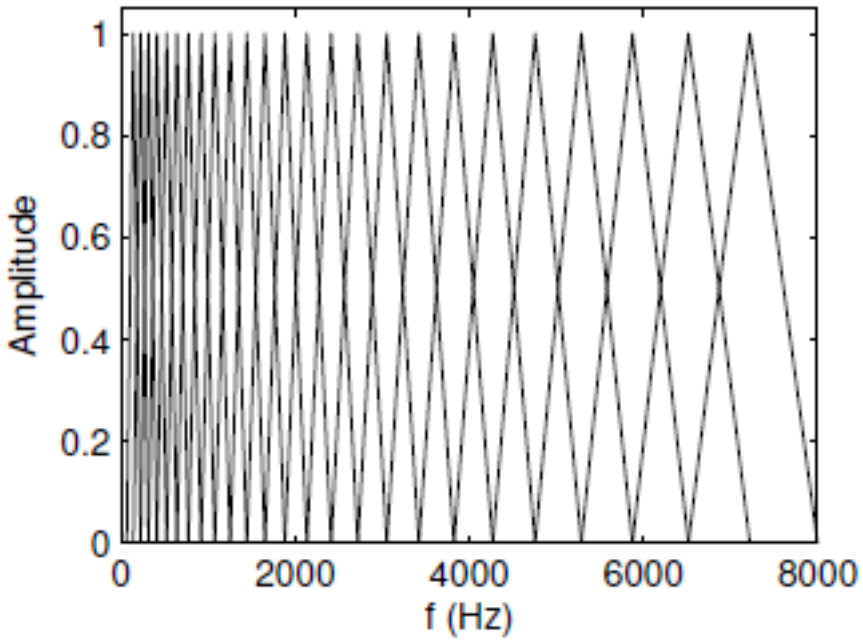
3.2 Mel Frequency Cepstral Coefficient (MFCC)

Για τη μοντελοποίηση της φωνής με HMMs απαιτείται μια κατάλληλη αναπαράσταση του σήματος του ήχου. Αυτή επιτυγχάνεται μέσω της εξαγωγής κατάλληλων χαρακτηριστικών. Τα MFCCs είναι μια δημοφιλής επιλογή για ηχητικά σήματα. Η κύρια ιδέα των MFCCs είναι ότι προσπαθούν να εξάγουν από το σήμα την πληροφορία που χρησιμοποιεί και το ανθρώπινο αυτί για την αναγνώριση των ήχων. Ο αλγόριθμος (Σχήμα 3.2)αποτελείται από τα ακόλουθα βήματα:

1. Χωρίζουμε το σήμα σε μικρά frames.
2. Για κάθε frame υπολογίζουμε το periodogram estimate του φάσματος ισχύος.
3. Εφαρμόζουμε το φίλτρο mel στο φάσμα ισχύος (Σχήμα 3.1).
4. Αθροίζουμε την ενέργεια σε κάθε φίλτρο.
5. Υπολογίζουμε το λογάριθμο όλων των ενεργειών των φίλτρων.

6. Υπολογίζουμε το DCT του λογαρίθμου των ενεργειών των φίλτρων.

7. Κρατάμε τους συντελεστές του DCT (μερικούς από τους πρώτους, συνήθως $10-13$) και απορρίπτουμε τους υπόλοιπους. Ο λόγος που το κάνουμε αυτό, είναι γιατί οι υπόλοιποι συντελεστές περιγράφουν πολύ γρήγορες εναλλαγές στο φάσμα ισχύος οι οποίες δε συνεισφέρουν ιδιαίτερα στην αναγνώριση ομιλίας.

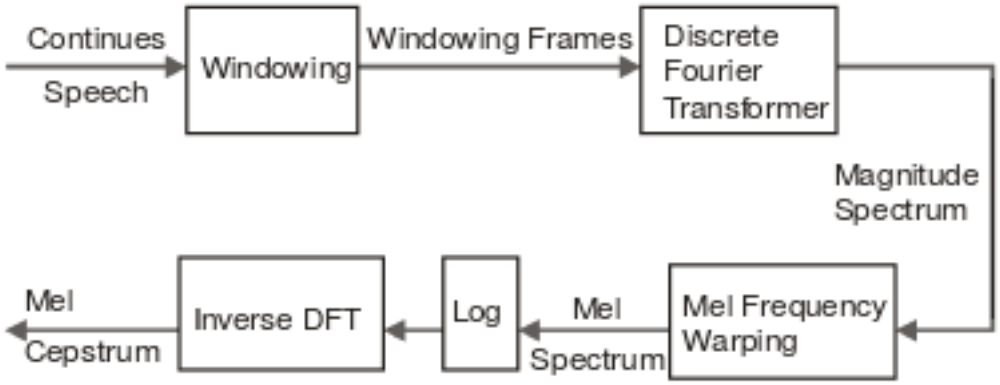


Σχήμα 3.1: Φίλτρο Mel. [8]

Παρόλο που το φωνητικό σήμα συνεχώς αλλάζει, θεωρούμε ότι παραμένει ίδιο και έχει την ίδια πληροφορία για περίπου $20ms-40ms$. Επομένως επιλέγουμε τη διάρκεια του frame να είναι $25ms$ με $10ms$ μετατόπιση του frame. Συνήθως $10-13$ DCT συντελεστές κρατούνται. Στην περίπτωσή μας επιλέγουμε να κρατήσουμε 13. Επίσης,είναι πολύ συνηθισμένο να προσαρτούμε στο βασικό σετ συντελεστών και τις πρώτες και δεύτερες παραγώγους τους.

3.3 Βασική αρχιτεκτονική του συστήματος

Η βασική αρχιτεκτονική ενός συστήματος αναγνώρισης ομιλίας συστήματος απεικονίζεται στο Σχήμα 3.3. Από τον ήχο εισόδου παίρνουμε τα διανύσματα χαρακτηριστικών $Y = y_1, y_2, ..., y_N$ που στην περίπτωσή μας είναι τα MFCCs. Ο αποκωδικοποιητής χρησιμοποιεί τα διανύσματα αυτά ως εισόδους και προσπαθεί να βρεί μια



Σχήμα 3.2: Σχηματική αναπαράσταση της διαδικασίας εξαγωγής χαρακτηριστικών MFCCs. [9]

ακολουθία λέξεων $W = w_1, w_2, ..., w_M$ που είναι το πιθανότερο να δημιούργησε το U. Με άλλα λόγια, ο αποκωδικοποιητής προσπαθεί να υπολογίσει το

$$w = argmax\{P(w|Y)\}$$

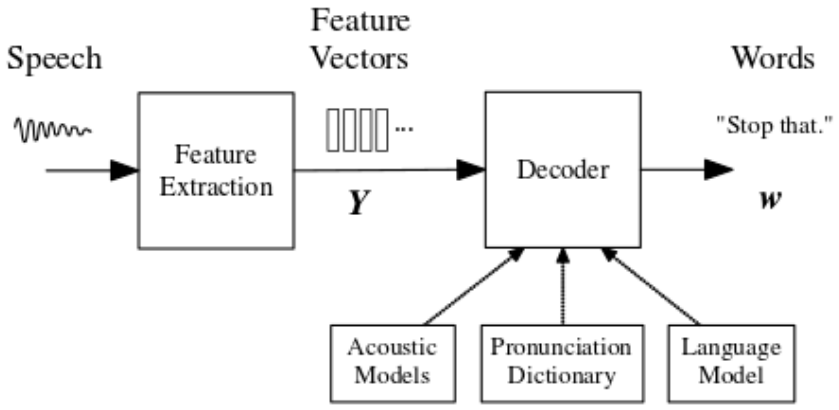
Ο υπολογισμός του $P(w|Y)$ όμως είναι δύσκολο εγχείρημα. Γι΄ αυτόν το λόγο, μετασχηματίζουμε το πρόβλημά μας σε ένα ισοδύναμο με χρηση του θεωρήματος Bayes. Το νέο μας πρόβλημα είναι ο υπολογισμός του

$$w = argmax\{P(Y|w)P(w)\}$$

Τώρα είναι ευκολότερο να ορίσουμε και τις δύο ζητούμενες πιθανότητες. Η $P(w)$ προσδιορίζεται από το γλωσσικό μοντέλο και η $P(Y|w)$ από το ακουστικό μοντέλο. Η βασική μονάδα ήχου που αντιπροσωπεύεται από το ακουστικό μοντέλο είναι το φώνημα. Για παράδειγμα η λέξη 'bat' συντίθεται απο τρία φωνήματα /b/ /ae/ /t/. Περίπου 40 τέτοια φωνήματα απαιτούνται για την Αγγλική γλώσσα. Μια πλήρης λίστα με τα φωνήματα που χρησιμοποιούνται υπάρχει στο παράρτημα Β.

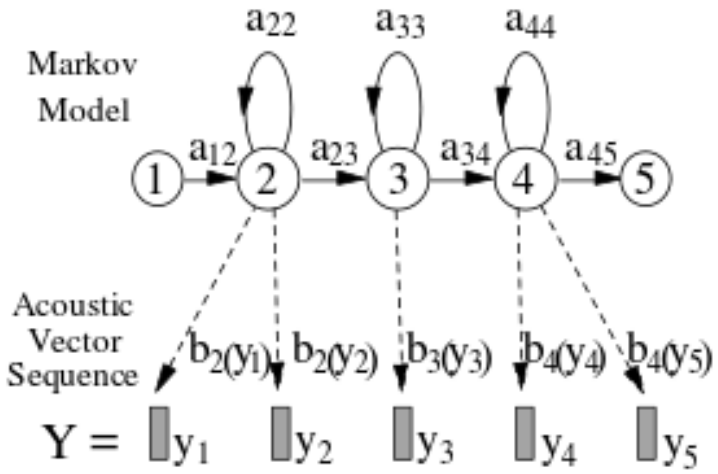
Για να δημιουργήσουμε ένα ακουστικό μοντέλο μιας ακολουθίας από λέξεις ενώνουμε τα φωνήματα της κάθε λέξης όπως τα βρίσκουμε στο λεξικό. Για να κατασκευάσουμε το γλωσσικό μοντέλο, το οποίο είναι συνήθως ένα N-gram μοντέλο,υπολογίζουμε την πιθανότητα της κάθε λέξης όταν δίνονται μόνο οι $N-1$ προηγούμενες. Αυτές οι πιθανότητες υπολογίζονται κατευθείαν από το κείμενο εκπαίδευσης. Έπειτα ο αποκωδικοποιητής χρησιμοποιεί και το γλωσσικό και το ακουστικό μοντέλο για να βρεί την πιθανότερη ακολουθία λέξεων.

Κάθε φώνημα q αναπαριστάται από ένα HMM με πιθανότητες μετάβασης κατάστασης a_{ij} και πιθανότητες εξόδου b_j. Δοθείσης της ακολουθίας των χαρακτηριστικών διανυσμάτων,εκτιμούμε την πιθανότερη ακολουθία κατάστασης κατά μήκος όλων των



Σχήμα 3.3: Βασική αρχιτεκτονική του συστήματος.

μοντέλων φωνημάτων. Αυτό με την υψηλότερη πιθανότητα είναι και αυτό που επιλέγουμε 3.4. Επαναλαμβάνουμε τη διαδικασία και για τα υπόλοιπα φωνήματα. Στο τέλος, επιλέγουμε σαν ακολουθία φωνημάτων τα φωνήματα με την υψηλότερη πιθανότητα.



Σχήμα 3.4: Μοντέλο φωνήματος [10]

Όσον αφορά τις πιθανότητες εξόδου, θα χρησιμοποιήσουμε κατανομή Gauss αντί για ένα μόνο αριθμό, όπως περιγράφηκε στην προηγούμενη ενότητα. Οι παράμετροι του μοντέλου μπορούν να υπολογιστούν με χρήση του αλγορίθμου Forward-Backward.

Καθόλη τη διάρκεια της διαδικασίας κάναμε τις ακόλουθες υποθέσεις:

- Κάθε κατάσταση είναι ανεξάρτητη των υπολοίπων, αν γνωρίζουμε την αμέσως

προηγούμενή της.

- Κάθε παρατήρηση είναι ανεξάρτητη των υπολοίπων, αν γνωρίζουμε την κατάσταση από την οποία παράχθηκε.

Το μεγαλύτερο πρόβλημα είναι ότι η αποσύνθεση κάθε λέξης σε μια σειρά από context-independent βασικά φωνήματα αποτυγχάνει στο να καλύψει μεγάλο εύρος των διάφορων context-dependent που υπάρχουν στον πραγματικό κόσμο. Για παράδειγμα, η βασική φόρμα προφοράς για τις λέξεις "mood" και "cool" θα χρησιμοποιούσε το ίδιο φωνήεν "oo", όμως στην πράξη η προφορά του "oo" στις δύο λέξεις είναι πολύ διαφορετική λόγω του επηρεασμού από το προηγούμενο και το επόμενο σύμφωνο. Τα Context Independent μοντέλα φωνημάτων αναφέρονται ως monophones. Ένας απλός τρόπος εξομάλυνσης του προβλήματος είναι η χρήση μοναδικού μοντέλου φωνημάτων για κάθε πιθανό ζευγάρι δεξιών και αριστερών γειτονικών φωνημάτων. Τα μοντέλα που προκύπτουν ονομάζονται triphones και αν υπάρχουν N βασικά φωνήματα τότε υπάρχουν N^3 πιθανά triphones. Για να αποφύγουμε data sparsity προβλήματα, το πλήρες σετ των πιθανών triphones L μπορεί να αντιστοιχηθεί σε ένα μειωμένο σετ φυσικών μοντέλων P με το να ομαδοποιήσουμε μαζί τις παραμέτρους σε κάθεcluster.

Η βασική διαδικασία για τη δημιουργία ακουστικού μοντέλου είναι η ακόλουθη:

- Monophone HMM μοντέλα χρησιμοποιούνται με κατανομή Gauss ως πιθανότητες εξόδου.
- Οι αρχικές παράμετροι της κατανομής Gauss είναι ίδιες με αυτές των δεδομένων εκπαίδευσης.
- Γίνεται χρήση EM για τον υπολογισμό των πραγματικών παραμέτρων.
- Επεκτείνουμε το monophone μοντέλο σε triphone χρησιμοποιώντας τις αντίστοιχες παραμέτρους Gauss από κάθε διακριτό triphone, για όλα τα τρία στάδια φωνημάτων (προηγούμενο-τρέχον-επόμενο) ως αρχικές παραμέτρους.
- Γίνεται ξανά χρήση EM για την εκτίμηση των triphone παραμέτρων Gauss.

Για να αυξήσουμε περαιτέρω τα αποτελέσματα αναγνώρισης μπορούμε να αλλάξουμε τις πιθανότητες κατάστασης εξόδου. Χρησιμοποιώντας την κατανομή Gauss που περιγράψαμε προηγουμένως, κάποιος εμμέσως υποθέτει ότι η παρατηρούμενη ακολουθία χαρακτηριστικών είναι συμμετρική. Ωστόσο αυτή δεν είναι η συνηθισμένη περίπτωση για διάφορους λόγους όπως λόγω παραλλαγών στην προφορά των ομιλητών. Για να αντιμετωπιστεί αυτό το ζήτημα αντικαθιστούμε τη μονή κατανομή Gauss με ένα μείγμα κατανομών Gauss. Κάνοντας αυτό μπορούμε να μοντελοποιήσουμε πολύ πιο πολύπλοκα δεδομένα. Για να βρούμε την εκτιμώμενη πιθανότητα μιας δοθείσας

ακολουθίας προσθέτουμε τις πιθανότητες κάθε διαφορετικής κατανομής πολλαπλασιαζόμενη με ένα συντελεστή βαρύτητας. Ο αλγόριθμος EM χρησιμοποιείται ξανά για τον υπολογισμό των παραμέτρων Gauss.

3.4 Περαιτέρω βελτιώσεις του βασικού συστήματος αναγνώρισης

Προκειμένου να μειώσουμε τη συσχέτιση μεταξύ των χαρακτηριστικών και να μειώσουμε τη διάστασή τους, μπορούμε να χρησιμοποιήσουμε πολλές προβολές ή συνδυασμούς αυτών πριν τα χαρακτηριστικά δοθούν ως είσοδος στο σύστημά μας. Γνωστές επιλογές είναι οι PCA και LDA οι οποίες περιγράφονται στη συνέχεια. Διάφορες μελέτες δείχνουν βελτιωμένα αποτελέσματα με χρήση αυτών των μεθόδων, όπως για παράδειγμα, στο [41] όπου παρατηρείται βελτίωση από 18.6 σε 13.2 με χρήση της μεθόδου LDA

3.4.1 Principal Components Analysis (PCA)

Principal Components Analysis (PCA) είναι μια τεχνική που χρησιμοποιείται στην πολυμεταβλητή ανάλυση δεδομένων. Χρησιμοποιείται κυρίως για τη μείωση της διάστασης των χαρακτηριστικών, αναλύοντας το σύνολο των δεδομένων στα κύρια συστατικά του, τα οποία είναι τα ιδιοδιανύσματα της μήτρας συνδιακύμανσης [42].

Ας υποθέσουμε ότι έχουμε N διανύσματα $X_n, n = 1.2...., N$ διαστάσεων $1 \times M$ το καθ΄ ένα. Σε κάθε ένα X_n αντιστοιχίζεται ένα διάνυσμα Y_n το οποίο είναι γραμμικός συνδυασμός των X_n.

$$Y_1 = e_{11}X_1 + e_{12}X_2 + ... + e_{1N}X_N$$
$$Y_2 = e_{21}X_1 + e_{22}X_2 + ... + e_{2N}X_N$$

κ.ο.κ.

Αφαιρούμε το μέσο όρο του δείγματος έτσι ώστε

$$\mu = \frac{1}{N-1}\sum_{n=1}^{N} X_n = 0$$

Υπολογίζουμε τη μήτρα συνδιακύμανσης

$$cov(X) = \Sigma = \frac{1}{N}\sum_{n=1}^{N} X(n)X^T(n)$$

$$cov(X) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \dots & \sigma_N^2 \end{pmatrix}$$

Επιπλέον, Y_i, Y_j έχουν συνδιακύμανση

$$cov(Y_i, Y_j) = \sum_{k=1}^{N} \sum_{l=1}^{N} e_{ik} e_{jl} \sigma_{kl} = e_i^T \Sigma e_j$$

Ο στόχος της PCA είναι η εύρεση των $e_{i1}, e_{i2}, ...e_{iN}$ έτσι ώστε η ποσότητα

$$cov(Y_i, Y_i) = e_i^T \Sigma e_i$$

να μεγιστοποιείται.

Ωστόσο, πρέπει να πληρούνται ορισμένοι περιορισμοί ώστε το πρόβλημα έχει μοναδική λύση.

$$e_i^T e_i = 1$$

Για κάθε $v = 1, 2, ..., i-1$

$$cov(Y_v, Y_i) = e_v^T \Sigma e_i = 0$$

Έχουμε ότι

$$e_i = \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{iN} \end{pmatrix}$$

είναι το ιδιοδιάνυσμα του πίνακα Σ που αντιστοιχεί στην ιδιοτιμή λ_i με

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N$$

καθώς και ότι η διακύμανση της $i-$οστής κύριας συνιστώσας ισούται με την $i-$οστή ιδιοτιμή.

$$var(Y_i) = var(e_{i1} X_1 + e_{i2} X_2 + \dots e_{iN} X_N) = \lambda_i$$

Η συνολική διακύμανση του X μπορεί να οριστεί ως το ίχνος του πίνακα Σ

$$trace(\Sigma) = \sigma_1^2 + \sigma_2^2 + ... + \sigma_N^2 = \lambda_1 + \lambda_2 + ...\lambda_N$$

Ως εκ τούτου, το ποσοστό της διακύμανσης που εκφράζεται με την ith κύρια συνιστώσα είναι

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + ...\lambda_N}$$

Αν κρατούσαμε όλες τις κύριες συνιστώσες, ο μετασχηματισμός θα ήταν ισοδύναμος με μία περιστροφή των αξόνων. Μπορούμε να μειώσουμε το πλήθος των διαστάσεων των δεδομένων μας με το να διατηρήσουμε τις πρώτες i κύριες συνιστώσες. Για να επιτευχθεί αυτό, θα πρέπει η διακύμανση που περιγράφεται από τις πρώτες i συνιστώσες να είναι περίπου ίση με τη συνολική διακύμανση έτσι ώστε η απώλεια πληροφορίας να είναι αμελητέα.

$$\frac{\lambda_1 + \lambda_2 + ...\lambda_i}{\lambda_1 + \lambda_2 + ...\lambda_N} \cong 1$$

Έχοντας επιλέξει το κατάλληλο πλήθος συνιστωσών i, δημιουργούμε έναν πίνακα με τα πρώτα i ιδιοδιανυσματα με τον οποίο πολλαπλασιάζουμε τα αρχικά δεδομένα μας X ώστε να αποκτήσουμε τα μετασχηματισμένα σε λιγότερες διαστάσεις δεδομένα.

3.4.2 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) είναι μια άλλη μέθοδος που χρησιμοποιούμε στην έρευνά μας για τη μείωση της διάστασης των χαρακτηριστικών εισόδου. Η κύρια ιδέα της μεθόδου είναι η εύρεση μιας προβολής των δεδομένων πάνω σε μία γραμμή (υπερεπίπεδο στη γενική περίπτωση) έτσι ώστε οι δύο κλάσεις του προβλήματος να είναι εύκολα διαχωρίσιμες (με μικρές τροποποιήσεις, αυτή η μέθοδος μπορεί να χρησιμοποιηθεί για τα προβλήματα με περισσότερες από δύο κλάσεις) [43]. Η μέθοδος LDA χρησιμοποιεί δυο κριτήρια που λαμβάνονται υπ΄ όψην ταυτόχρονα για την κατασκευή των νέων αξόνων μετασχηματισμού:

1. Μεγιστοποίηση της απόστασης μεταξύ των μέσων τιμών των κλάσεων.

2. Ελαχιστοποίηση της διακύμανσης (scatter) εντός κάθε κλάσης.

Ας υποθέσουμε ότι έχουμε n δείγματα x από κάθε μία από τις δύο κλάσεις(C_1, C_2) όπου n_1 είναι το πλήθος των δειγμάτων που ανήκουν στην πρώτη ομάδα και n_2 το πλήθος των δειγμάτων που ανήκουν στη δεύτερη ομάδα (με $n = n_1 + n_2$).

Η μέση τιμή των προβολών των δειγμάτων πάνω σε μία γραμμή με μοναδιαίο διάνυσμα β, δίνεται από τις ακόλουθες εξισώσεις για κάθε ομάδα:

$$\mu_1 = \frac{1}{n_1} \sum_{x_i \in C_1}^{n_1} \beta^T x_i = \beta^T (\frac{1}{n_1} \sum_{x_i \in C_1}^{n_1} x_i) = \beta^T m_1$$

$$\mu_2 = \frac{1}{n_2} \sum_{x_i \in C_2}^{n_2} \beta^T x_i = \beta^T (\frac{1}{n_2} \sum_{x_i \in C_2}^{n_2} x_i) = \beta^T m_2$$

όπου m_1 είναι η μέση τιμή της κλάσης C_1 και m_2 είναι η μέση τιμή της κλάσης C_2.

Η μέση τιμή του δείγματος είναι

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i$$

και η διακύμανση

$$var = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$$

Στην LDA, αντί για τη διακύμανση, χρησιμοποιούμε την ποσότητα scatter s για να μετρήσουμε την εξάπλωση των δεδομένων γύρο από τη μέση τιμή. Η ποσότητα scatter ορίζεται ως:

$$s = \sum_{i=1}^{n}(x_i - m)^2$$

Παρατηρούμε ότι η ποσότητα scatter είναι η διακύμανση του δείγματος πολλαπλασιασμένη με n.

Το scatter για τις προβολές των δεδομένων κάθε κλάσης είναι:

$$\sigma_1^2 = \sum_{x_i \in C_1} (\beta^T x_i - \mu_1)^2 = \sum_{y_i \in C_1} (y_i - \mu_1)^2$$

$$\sigma_2^2 = \sum_{x_i \in C_2} (\beta^T x_i - \mu_2)^2 = \sum_{y_i \in C_2} (y_i - \mu_2)^2$$

όπου $y_i = \beta^T x_i$

Για να μεταφράσουμε την έννοια της διαχωρισιμότητας σε μαθηματικά, ορίζουμε μια συνάρτηση κόστους ως εξής:

$$J(\beta) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

Στόχος μας είναι η εύρεση της τιμής του β η οποία μεγιστοποιεί την $J(\beta)$. Όσο μεγαλύτερη είναι η τιμή της $J(\beta)$ τόσο μεγαλύτερη είναι η διαχωρισιμότητα των κλάσεων.

Ορίζουμε τον *within* scatter πίνακα ως

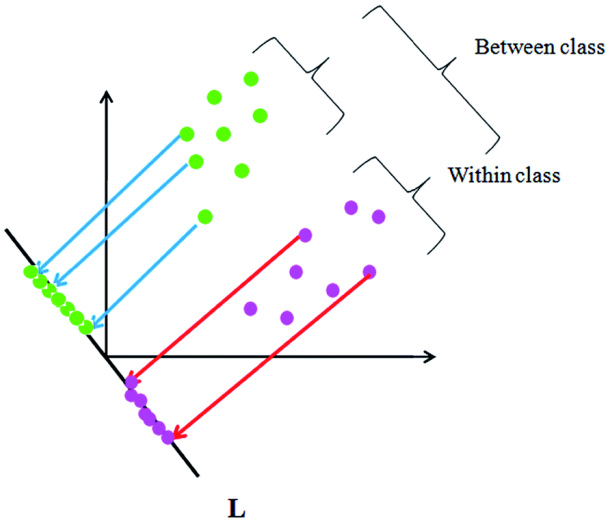
$$S_w = s_1 + s_2$$

και τον *between* scatter πίνακα ως

$$S_b = (m_1 - m_2)(m_1 - m_2)^T$$

Ο *within* πίνακας περιγράφει τη διακύμανση εσωτερικά των κλάσεων ενώ ο *between* πίνακας περιγράφει τη διακύμανση μεταξύ των κλάσεων.

Παρακάτω φαίνεται μία οπτική εξήγηση αυτών των πινάκων (Σχήμα 3.5).



Σχήμα 3.5: Οπτική εξήγηση του πίνακα Within - Between.

Μπορούμε τώρα να ξαναγράψουμε τη συνάρτηση κόστους ώστε να αποτελείται μόνο από τον όρο β ώστε να τη μεγιστοποιήσουμε.

$$J(\beta) = \frac{\beta^T S_b \beta}{\beta^T S_w \beta}$$

Υπολογίζουμε την πρώτη παράγωγο του $J(\beta)$ ως προς β και τη θέτουμε ίση με μηδέν. Μετά από κάποιους υπολογισμούς καταλήγουμε στο:

$$S_b\beta = \frac{\beta^T S_b \beta}{\beta^T S_w \beta} S_w\beta = \alpha S_w\beta$$

Η λύση στο πρόβλημα ιδιοτιμών είναι:

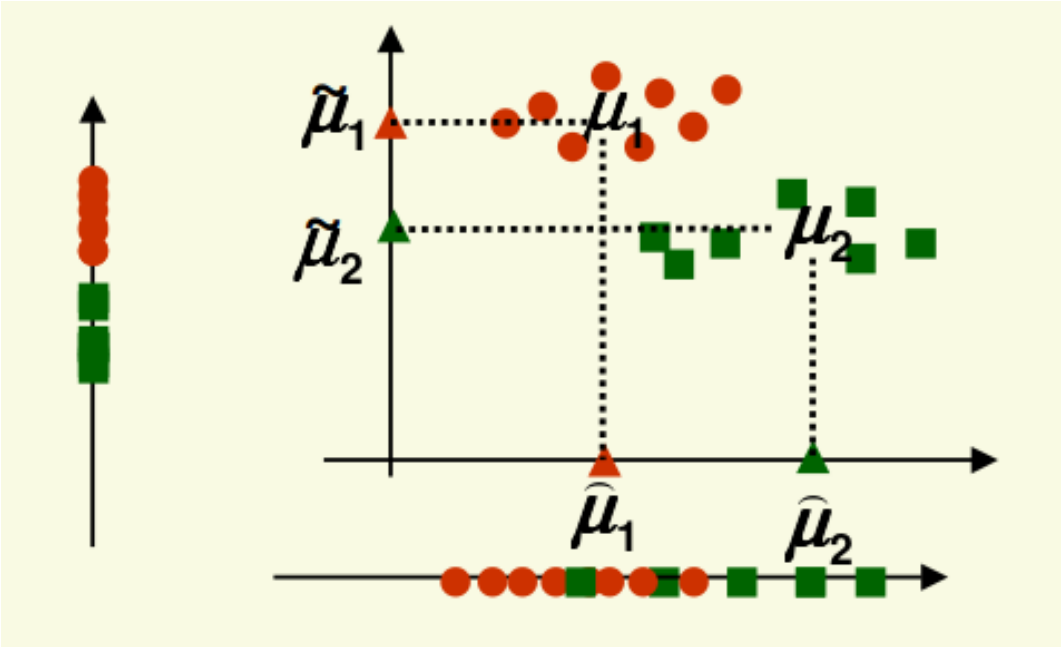
$$\beta = S_w^{-1}(m_1 - m_2)$$

Κρατάμε τα πρώτα k ιδιοδιανύσματα τα οποία αντιστοιχούν στις k μεγαλύτερες ιδιοτιμές, όπου k είναι το πλήθος των διαστάσεων που επιλέγουμε να κρατήσουμε. Η διαδικασία στη συνέχεια είναι η ίδια με αυτή που περιγράψαμε στο αντίστοιχο τμήμα της μεθόδου PCA.

Σε αυτό το σημείο κάποιος μπορεί να αναρωτηθεί 'Τιατί χρειάζονται δύο κριτήρια; Δεν μπορούμε απλά να χρησιμοποιήσουμε μόνο το πρώτο;' (Το να χρησιμοποιήσει κανείς μόνο το δεύτερο δεν βγάζει νόημα).

Η απάντηση στο ερώτημα αυτό είναι ότι είναι πολύ πιθανό η προβολή των δεδομένων μας σε μια γραμμή που μεγιστοποιεί την απόσταση των μέσων των δύο τάξεων, μπορεί επίσης να κρατήσει περισσότερες πληροφορίες από ότι απαιτείται για τη διακύμανση των τάξεων, οι οποίες είναι πιθανόν να οδηγήσουν σε μεγάλη αλληλεπικάλυψη μεταξύ των τάξεων. Προκειμένου να διορθώσουμε όσο δυνατόν περισσότερες προβληματικές περιπτώσεις όπως αυτές που περιγράφονται παραπάνω, χρησιμοποιούμε το δεύτερο κριτήριο ώστε να κρατήσουμε τις προβολές των κλάσεων όσο γίνεται πιο συμπαγείς.

Μια οπτική εξήγηση φαίνεται παρακάτω.

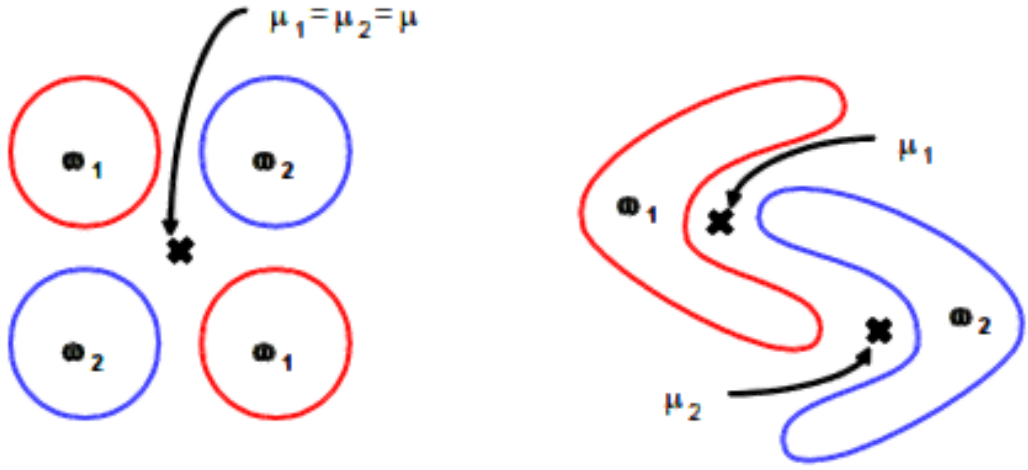


Σχήμα 3.6: Χρήση μόνο του πρώτου κριτηρίου.

Είναι προφανές ότι η προβολή των δεδομένων στον X-άξονα έχει σαν αποτέλεσμα η απόσταση μεταξύ των μέσων τιμών των προβαλόμενων δειγμάτων να είναι μεγαλύτερη. Ωστόσο, οι δύο κλάσεις δεν είναι επαρκώς διαχωρίσιμες καθώς διατηρείται πολύ μεγάλο μέρος της διακύμανσης. Από την άλλη πλευρά, η προβολή των δεδομένων στον Y-άξονα έχει σαν αποτέλεσμα αυξημένη τα δεδομένα να είναι ευκόλως διαχωρίσιμα, χωρίς να υπάρχει επικάλυψη μεταξύ των δύο κλάσεων, πάρα το γεγονός ότι η απόσταση των μέσων τιμών των προβλημένων δεδομένων είναι μικρότερη.

Αν και η LDA έχει παρόμοια επίδοση με πιο περίπλοκες τεχνικές, υπάρχουν περιπτώσεις που αποτυγχάνει να βρει μια καλή προβολή για τα δεδομένα. Στις περισσότερες από αυτές τις περιπτώσεις, η προβολή των δεδομένων σε οποιαδήποτε γραμμή θα είχε σαν αποτέλεσμα τα δεδομένα να είναι μη διαχωρίσιμα. Στις υπόλοιπες περιπτώσεις υπάρχει κάποια κατάλληλη γραμμή για την προβολή των δεδομένων όμως η LDA δεν μπορεί να την εντοπίσει. Άλλες, απλούστερες μέθοδοι μπορεί να βρουν μία καλή γραμμή σε τέτοιες περιπτώσεις. Για παράδειγμα, η PCA ενδέχεται να βρει μια καλή γραμμή για τα δεδομένα όταν η LDA αποτυγχάνει, παρόλο που η LDA είναι καλύτερη μέθοδος γενικά. Επιπροσθέτως, η LDA μπορεί να αποτύχει στο να δώσει καλά αποτελέσματα όταν εφαρμόζεται σε περίπλοκα σύνολα δεδομένων τα οποία δεν έχουν κανονική κατανομή ή όταν οι μέσες τιμές των κλάσεων ισούται μεταξύ τους.

Μερικά παραδείγματα στα οποία η LDA δεν έχει καλά αποτελέσματα φαίνονται στη ακόλουθη εικόνα.



Σχήμα 3.7: Σύνολα δεδομένων που αποτυγχάνει η μέθοδος LDA.

3.4.3 Maximum Likelihood Linear Transform (MLLT)

Σε πολλές περιπτώσεις εφαρμόζεται επίσης και Maximum Likelihood Linear Transform [44]. Αυτός ο μετασχηματισμός μας επιτρέπει να χρησιμοποιήσουμε μη διαγώνιους πίνακες συνδιακύμανσης για τις Gaussians. Γράφουμε τον πίνακα συνδιακύμανσης της j κατανομής στη μορφή

$$\Sigma_j = W D_j W^T$$

όπου D_j είναι ο διαγώνιος πίνακας που είναι μοναδικός ανά κατανομή και W είναι ο πίνακας μετασχηματισμού ο οποίος είναι ίδιος για όλες τις κατανομές (ή για ένα σετ κατανομών). Πειράματα έδειξαν μείωση του WER κατά 3% με χρήση του μετασχηματισμού MLLT [44].

3.4.4 Προσαρμοστική σε ομιλητές εκπαίδευση

Ένα τελευταίο πράγμα που θέλουμε να αναφέρουμε είναι η προσαρμοστική σε ομιλητές τεχνική εκπαίδευσης για τη δημιουργία μοντέλων ανεξαρτήτου ομιλητή. Η τεχνική αυτή έρχεται να λύσει το πρόβλημα στο οποίο μεγάλος όγκος δεδομένων χρησιμοποιείται για τη μοντελοποίηση της μεταβλητότητας μεταξύ των διαφορετικών ομιλητών παρά για τη μεταβλητότητα μεταξύ των διαφορετικών ήχων. Πολλοί μετασχηματισμοί μπορούν να χρησιμοποιηθούν για προσαρμοστική σε ομιλητές εκπαίδευση. Ένας πολύ γνωστός μετασχηματισμός είναι ο MLLR ο οποίος εφαρμόζεται επίσης στα διανύσματα χαρακτηριστικών πριν δοθούν ως είσοδοι στον αποκωδικοποιητή. Πειράματα έχουν δείξει βελτίωση των αποτελεσμάτων κατά περίπου 1.5% [45]. Συνήθως συνδυάζεται με άλλες τεχνικές μετασχηματισμού όπως αυτές που αναφέραμε στην προηγούμενη παράγραφο.

3.4.5 Maximum Mutual Information (MMI)

Κατά τη διάρκεια της φάσης εκπαίδευσης του HMM,ο στόχος είναι να εκτιμήσουμε τις παραμέτρους του HMM που μεγιστοποιούν την πιθανότητα του μοντέλου να παράξει την ακολουθία εκπαίδευσης. Οι παράμετροι αυτοί βρίσκονται με χρήση του αλγορίθμου EM. Στις περισσότερες περιπτώσεις ωστόσο δεν υπάρχουν αρκετά δεδομένα εκπαίδευσης διαθέσιμα για τη μεγιστοποίηση πιθανοφάνειας Maximum Likelihood (ML) ώστε να λειτουργεί ικανοποιητικά. Επίσης, σε ένα σύστημα αναγνώρισης φωνής αυτό που τελικά ενδιαφέρει είναι να μεγιστοποιηθεί η διακριτική ικανότητα των μοντέλων, την ικανότητά τους δηλαδή να ξεχωρίζουν μεταξύ ήχων και όχι τόσο να μοντελοποιούν πλήρως τους ήχους αυτούς. Γι΄ αυτό έχουν αναπτυχθεί μια σειρά εναλλακτικών κριτηριών προς αυτή την κατεύθυνση. Ο κύριος στόχος των διακριτικών (discriminative) αυτών κριτηρίων είναι η βελτιστοποίηση της εκ των υστέρων πιθανότητας των λέξεων δοθείσης του ακουστικού μοντέλου.

Μια δημοφιλής επιλογή τέτοιου κριτηρίου είναι το Maximum Mutual Information (MMI) που στοχεύει στη μεγιστοποίηση της posterior πιθανότητας και επιτυγχάνει

βελτίωση της τάξεως του 2% [46]. Η συνάρτηση εκπαίδευσης που έχουμε στόχο να μεγιστοποιηθεί είναι η

$$F(\lambda) = \sum_{i=1}^{N} \log(\frac{p(Y_r|w_r;\lambda)P(w_r)}{\sum_w p(Y_r|w;\lambda)P(w)})$$

όπου λ είναι οι παράμετροι του μοντέλου, U η ακολουθία παρατηρήσεων και w η ακολουθία παρατηρήσεων [47].

3.4.6 Subspace Gaussian Mixture Model (SGMM)

Οι τεχνικές που περιγράφησαν προηγουμένως σχετικά με την προσαρμοστική σε ομιλητές εκπαίδευση και τη μείωση των εκτιμώμενων παραμέτρων μπορούν να παράξουν μοντέλα με βελτιωμένη απόδοση. Ωστόσο,μια νεότερη μέθοδος που συνδυάζει τα πλεονεκτήματα και των δύο τεχνικών και παράγει συστήματα με ακόμα περισσότερο βελτιωμένη απόδοση έχει βρεθεί.

Η μέθοδος αυτή ονομάζεται Subspace Gaussian Mixture Model (SGMM) και βασίζεται στην κλασσική GMM προσέγγιση [48]. Η ιδέα των SGMMs είναι ότι το ακουστικό μοντέλο μπορεί να εκπαιδευτεί χρησιμοποιώντας έναν υπόχωρο με χαμηλότερη διάσταση από το κανονικό GMM μοντέλο. Οι εξισώσεις του μοντέλου φαίνονται παρακάτω.

$$p(x|j) = \sum_{i=1}^{I} w_{ij} N(x; \mu_{ji}, \Sigma_i)$$

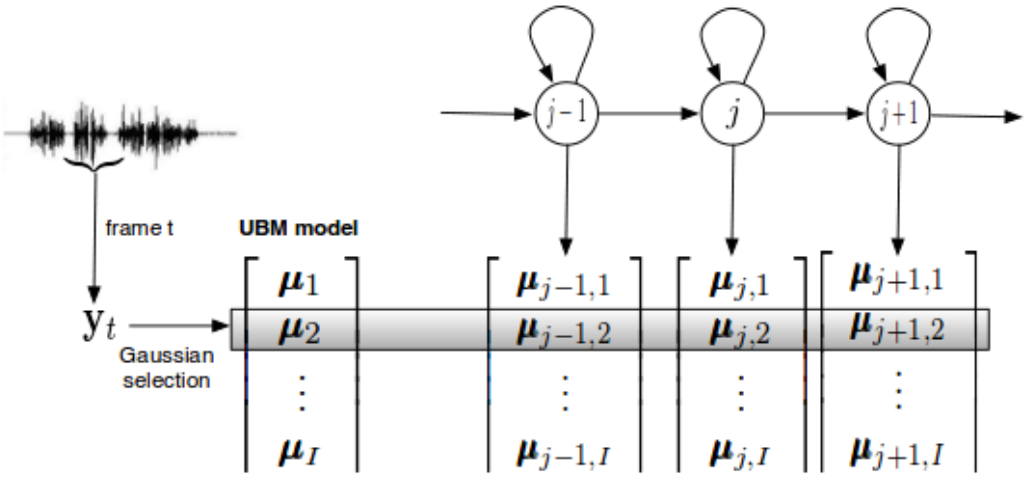
$$\mu_{ji} = M_i v_j$$

$$w_{ji} = \frac{exp(w_i^T v_j)}{\sum_{i=1}^{I} exp(w_i^T v_j)}$$

Οι παράμετροι μπορούν να χωριστούν σε δύο sets. Το πρώτο περιλαμβάνει τις global παραμέτρους $(M_i, N_i, w_i, \Sigma_i)$, οι οποίες πρέπει να υπολογιστούν μια φορά και αποθηκεύονται για να χρησιμοποιηθούν όταν χρειαστεί. Το άλλο περιλαμβάνει τις state-dependent (v_{jm}, c_{jm}) που υπολογίζονται κατά τα συνηθισμένα. Ο συνδυασμός όλων των παραμέτρων δίνει τις συνιστώσες Gauss των σταδίων του HMM [11]. Δεδομένου ότι χρησιμοποιούμε χώρο χαμηλής διάστασης, ο αριθμός των παραμέτρων που πρέπει να υπολογιστεί είναι μικρός. Ωστόσο οι προκύπτουσες Gaussians σε κάθε υποχώρο είναι πολύ μεγάλες (π.χ. $I = 400$), γεγονός που έχει σαν αποτέλεσμα αυξημένο χρόνο εκπαίδευσης και αποκωδικοποίησης.

Μια ενδιαφέρουσα λύση είναι κάθε φορά να διαλέγουμε τις Γκαουσιανές κατανομές που έχουν καλές πιθανότητες να παράγουν ένα καλό μοντέλο και να χρησιμοποιήσουμε

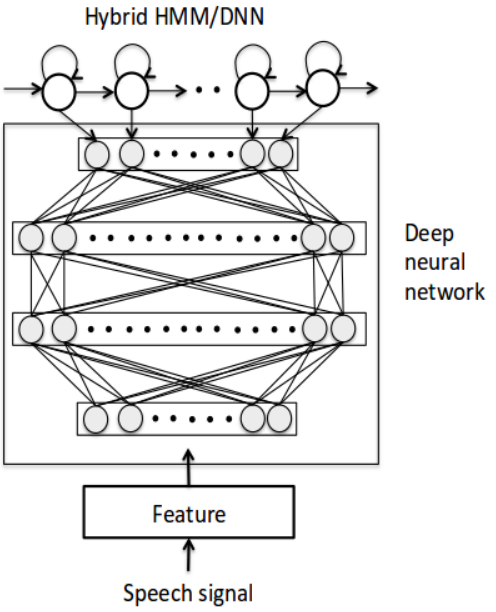
μόνο αυτές. Στην πράξη, ένα Καθολικό Μοντέλο Υποβάθρου Universal Background Model (UBM) χρησιμοποιείται για τη διαδικασία επιλογής αυτών των Gaussians [49]. Το UBM είναι ένα μείγμα από I Gaussians. Για κάθε ακουστικό frame υπολογίζεται η πιθανότητα κάθε Gaussian του UBM. Οι Gaussians (π.χ. 15 Gaussians) του UBM με το υψηλότερο σκορ χρησιμοποιούνται για την επιλογή των αντίστοιχων Gaussians από κάθε κατάσταση του HMM που θα χρησιμοποιηθούν για το συγκεκριμένο frame. Το Σχήμα 3.8 δείχνει την όλη διαδικασία. Εφόσον τα SGMMs είναι mixture μοντέλα όλες οι προηγούμενες τεχνικές που έχουν περιγραφεί μπορούν να εφαρμοστούν.



Σχήμα 3.8: Χρήση του UBM για την επιλογή των Γκαουσσιανών. [11]

3.4.7 Συνδιάζοντας DNNs και HMMs

Μια άλλη ενδιαφέρουσα ιδέα είναι η χρήση DNNs σε συνδυασμό με HMMs. Τα DNNs χρησιμοποιούνται για να μοντελοποιήσουν τις πιθανότητες παρατήρησης και να αντικαταστήσουν τα GMMs. Το κύριο πλεονέκτημα αυτής της προσέγγισης είναι ότι τα DNNs απαιτούν λιγότερες παραδοχές για τα δεδομένα και είναι περισσότερο διακριτικά (discriminant) [7]. Ένα υβριδικό HMM/DNN setup φαίνεται στο Σχήμα 3.9. Περισσότερες πληροφορίες σχετικά με τα βαθιά νευρωνικά δίκτυα παρουσιάζονται στο κεφάλαιο 2.



Σχήμα 3.9: Υβριδικό HMM/DNN σύστημα.

3.5 Πως δουλεύουν τα συστήματα του kaldi

Αφού περιγράψαμε τους βασικούς αλγόριθμους και ιδέες θα εξηγήσουμε πως μπορούμε να τα χρησιμοποιήσουμε όλα αυτά για να κατασκευάσουμε ένα σύστημα αναγνώρισης ομιλίας χρησιμοποιώντας το kaldi. Το kaldi είναι ένα πρόγραμμα που προσφέρει ένα μεγάλο εύρος δυνατοτήτων όσον αφορά την εκπαίδευση τελευταίας γενιάς συστημάτων αναγνώρισης ομιλίας. Το kaldi [50] δουλεύει ως εξής:

Από τα δεδομένα εκπαίδευσης εξάγουμε τα MFCCs χαρακτηριστικά για να εκπαιδεύσουμε το ακουστικό μοντέλο. Τα MFCCs ενώνονται με τις πρώτες και δεύτερες παραγώγους τους για να δημιουργήσουν το τελικό χαρακτηριστικό. Πρώτα εκπαιδεύουμε ένα κλασικό σύστημα monophone (*mono*) χρησιμοποιώντας τον forward-backward αλγόριθμο για την εκτίμηση των παραμέτρων και τον αλγόριθμο viterbi για αποκωδικοποίηση. Τα GMMs χρησιμοποιούνται για να αναπαραστήσουν τις πιθανότητες παρατήρησης.

Βασισμένο σε αυτό το μοντέλο, το πρώτο (tri_1) (από τα τρία) triphone μοντέλα εκπαιδεύεται χρησιμοποιώντας αρχικά Gaussians mixture models. Στο δεύτερο triphone model (tri_2) εφαρμόζεται LDA μετασχηματισμό για την μείωση της διάστασης των χαρακτηριστικών εισόδου σε συνδυασμό με MLLT που μας επιτρέπει να χρησιμοποιήσουμε covariance matrices που δεν είναι διαγώνια για Gaussians και ως εκ τούτου να βελτιώσει την ικανότητα των Gaussians να μοντελοποιήσουν τις ακολουθίες παρατηρήσεων [51]. Στο τρίτο triphone model (tri_3), εφαρμοζουμε SAT training (μαζί τους άλλους μετασχηματισμούς που χρησιμοποιούμε στο δεύτερο triphone model)

έτσι ώστε να καταγράψουμε την μεταβλητότητα των φωνημάτων των ομιλητών και όχι την μεταβλητότητα των ίδιων.

Το επόμενο μοντέλο ($SGMM_1$) βασίζεται σε SGMMs . Τα μοντέλα αυτά χρησιμοποιούν επίσης triphones άλλα αντί για πλήρη Gaussian mixtures χρησιμοποιούν μονό ένα μικρό μέρος των Gaussians (διαφορετικό για κάθε ακουστικό frame). Η επιλογή γίνεται σύμφωνα με τα σκορ των Gaussians ενός UBM. Το επόμενο μοντέλο ($SGMM_2$) που δημιουργείται είναι βασισμένο πάλι σε SGMMs. Ομως δεν στοχεύει στην μεγιστοποίηση της πιθανότητας της παράγωγης της ακολουθίας εκπαίδευσης. Χρησιμοποιει το MMI criterion για να μεγιστοποιήσει την εκ των υστέρων πιθανότητα μιας λέξης δεδομένου του ακουστικού μοντέλου [52].

Το Kaldi μπορεί να εκπαιδεύσει μοντέλα (DNN) συνδυάζοντας τη δύναμη των Deep Neural Networks με την κλασική triphone HMM προσέγγιση. Εξ΄ ορισμού χρησιμοποιεί έξι hidden layers με 2048 διαστάσεις. Εφαρμόζεται LDA στα χαρακτηριστικά εισόδου. Το DNN χρησιμοποιείται για να υπολογίσει τις πιθανότητες παρατήρησης του HMM. Το μεγαλύτερο πλεονέκτημα της μεθόδου hybrid HMM/DNN είναι ότι τα DNNs μπορούν να υπολογίσουν με μεγάλη επιτυχία τις πιθανότητες παρατήρησης καθώς έχουν μεγάλη διακριτική ικανότητα.

3.6 Σύνοψη

Στο κεφάλαιο αυτό παρουσιάσαμε της βασικές αρχές πάνω στις οποίες βασίζονται τα συστήματα αναγνώρισης ομιλίας καθώς και διάφορες σύγχρονες μεθόδους και προσεγγίσεις οι οποίες χρησιμοποιούνται για την περαιτέρω βελτίωση των αποτελεσμάτων.

Συγκεκριμένα, εξηγήσαμε πως δουλεύουν τα κρυφά Μαρκοβιανά μοντέλα και οι κύριοι αλγόριθμοι που χρησιμοποιούνται για την εκπαίδευσή τους. Στη συνέχεια, αναλύσαμε πως γίνεται η εξαγωγή των χαρακτηριστικών ήχου. Έπειτα, περιγράψαμε τη βασική αρχιτεκτονική ενός συστήματος αναγνώρισης ομιλίας και επεκτείναμε αυτές τις ιδέες, εξηγώντας τις διάφορες προσεγγίσεις που εφαρμόζονται σήμερα. Τέλος, έγινε μία περιγραφή της λειτουργίας του κύριου εργαλείου που χρησιμοποιήσαμε για την εκπόνηση της παρούσας εργασίας.

Κεφάλαιο 4

Μέθοδοι και αλγόριθμοι

Σε αυτό το κεφάλαιο περιγράφουμε τις βασικές μεθόδους και αλγόριθμους που χρησιμοποιούνται για τα πειράματα. Πρώτα εξηγούμε πώς λειτουργεί ο αλγόριθμος εξαγωγής χαρακτηριστικών SIFT. Στη συνέχεια, περιγράφουμε τις διάφορες τεχνικές μείωσης των διαστάσεων των χαρακτηριστικών, όπως PCA, LDA και CCA. Τέλος, εξηγούμε την τεχνική Bow σε συνδυασμό με SVMs, καθώς και τη μέθοδο SMOTE που χρησιμοποιείται για την αντιμετώπιση των προβλημάτων που αφορούν την ανισορροπία στο πλήθος των δεδομένων.

Συνοπτικά, χρησιμοποιήσαμε τα SIFT (Scale-Invariant Feature Transform) χαρακτηριστικά για την περιγραφή κάθε frame στο βίντεο. Εφαρμόζοντας τη μέθοδο Bag of Visual Words μετατρέπουμε αυτούς τους descriptors σε ένα ιστόγραμμα ανά εικόνα. Εξάγουμε MFCCs τα οποία μαζί με τα articulatory- ακουστικά ιστογράμματα είναι οι δυο όψεις του πειράματος μας. Τέλος εφαρμόζουμε ένα multi- view setup χρησιμοποιώντας CCA. Τα πειραματικά αποτελέσματα παρουσιάζουν βελτιώσεις στην αναγνώριση φωνής σε σύγκριση με την κλασική audio-based προσέγγιση.

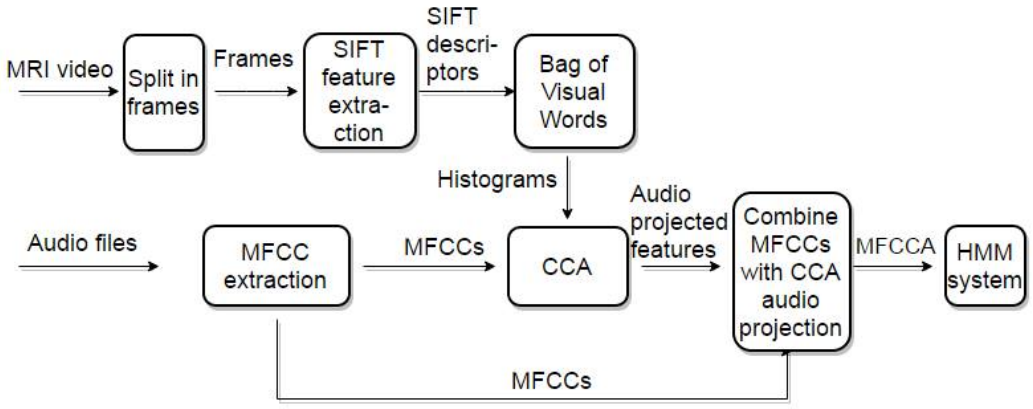
Το σχηματικό διάγραμμα του συστήματος μας φαίνεται παρακάτω: (Σχήμα 4.1).

4.1 Πως δουλεύει το σύστημά μας

Τώρα θα περιγράψουμε πώς οι θεωρητικές μέθοδοι εφαρμόζονται στην πράξη και πώς υλοποιείται το σύστημά μας. Το σύστημά μας μπορεί να διαιρεθεί σε τρία μέρη: Το ακουστικό, αυτό που έχει σχέση με το σύστημα παραγωγής φωνής (articulatory) και το κομμάτι που αφορά στον συνδυασμό (acoustic-articulatory).

4.2 Εξαγωγή χαρακτηριστικών

Το ακουστικό μέρος αποτελείται κυρίως από το βήμα της εξαγωγής των χαρακτηριστικών του ήχου. Τα χαρακτηριστικά που χρησιμοποιήθηκαν είναι τα MFCCs. Επιλέγουμε τη διάσταση των MFCCs να είναι 13. Ως εκ τούτου, μετά την προεπε-



Σχήμα 4.1: Multi-view σύστημα για τη βάση δεδομένων rt-MRI-TIMIT.

ξεργασία των αρχείων ήχου, εξάγονται για κάθε παράθυρο τα MFCCs καθώς και οι πρώτες και δεύτερες παράγωγοι τους που σχηματίζουν τελικά το διάνυσμα χαρακτηριστικών του παραθύρου αυτού.

Το articulatory μέρος αποτελείται από τα εξής: πρώτα θα χωρίσει το βίντεο μας σε πλαίσια. Όλα τα ξεχωριστά πλαίσια κατηγοριοποιούνται με τη χρήση των phone alignment (αρχεία που περιγράφουν κάθε χρονική στιγμή ποιο φώνημα εκφωνείται στο αρχείο αντίστοιχο αρχείο ήχου.) αρχείων που δημιουργήθηκαν από τα αρχεία ήχου. Τα πλαίσια χρησιμοποιούνται για την εξαγωγή των articulatory χαρακτηριστικών για το σύστημά μας. Χρησιμοποιούμε SIFT χαρακτηριστικά στο setup μας. Για κάθε frame, ο αλγόριθμος εύρεσης των SIFT παράγει πολλά σημεία ενδιαφέροντος. Το πλήθος αυτών των σημείων διαφέρει από εικόνα σε εικόνα. Ο αλγόριθμος εύρεσης των SIFT, υπολογίζει δύο διανύσματα για κάθε σημείο ενδιαφέροντος, το descriptor και το detector. Κρατάμε μόνο τους descriptor στη συνέχεια των πειραμάτων μας. Οι descriptors που βρέθηκαν μέσω αυτής της διαδικασίας δίνονται ως είσοδος στο σύστημα BoW. Τα κέντρα του αλγορίθμου k-means χρησιμοποιούνται για να δημιουργήσουν ένα ιστόγραμμα για κάθε frame, που είναι τα τελικά articulatory μας χαρακτηριστικά. Για να ελέγξουμε το πόσο καλά διάφορες τεχνικές κανονικοποίησης ιστογραμμάτων λειτουργούν, έχουμε δημιουργήσει ένα ξεχωριστό σύστημα που χρησιμοποιεί ως είσοδο τα τελικά ιστογράμματα, εφαρμόζει τεχνικές κανονικοποίησης και χρησιμοποιεί multi-label SVMs να κατατάξει τα ιστογράμματα. Η μέθοδος SMOTE χρησιμοποιείται για να χειριστεί την ανισορροπία που εμφανίζεται των δεδομένων σε αυτή τη διαδικασία. Έχοντας καθορίσει τη μέθοδο κανονικοποίησης με τα καλύτερα αποτελέσματα, την εφαρμόζονται στα τελικά χαρακτηριστικά μας.

Το μέρος που συνδυάζεται η πληροφορία από το σύστημα παραγωγής φωνής με την ακουστική είναι το μέρος που ακουστική πληροφορία συνδυάζεται με την articulatory πληροφορία για να δώσει το τελικό audio-articulatory χαρακτηριστικό. Η

μέθοδος CCA εφαρμόζεται στα audio και articulatory χαρακτηριστικά. Στα πειράματά μας, κρατάμε τη προβολή των audio χαρακτηριστικών. Τα τελικά audio-articulatory χαρακτηριστικά (MFCCA) είναι οι προβολές των ηχητικών χαρακτηριστικών που επισυνάπτονται με τα αρχικά χαρακτηριστικά ήχου.

Τέλος χρησιμοποιήσαμε τις διάφορες συνταγές του kaldi για να δημιουργήσουμε το τελικό μας σύστημα. Εξετάζουμε τέσσερις από τις διαθέσιμες επιλογές. Στην αρχή, έχουμε δημιουργήσει ένα σύστημα monophone και στη συνέχεια, δοκιμάζουμε τρεις παραλλαγές των συστημάτων triphone. Η διαφορά μεταξύ των triphone συστημάτων είναι οι μετασχηματισμοί, για παράδειγμα LDA, που εφαρμόζονται στα χαρακτηριστικά εισόδου. Οι ακριβής διαφορές μεταξύ των triphone συστημάτων, καθώς και το πώς λειτουργούν, εξηγούνται αναλυτικά στο κεφάλαιο 3.

4.3 Εξαγωγή χαρακτηριστικών από τα δεδομένα MRI

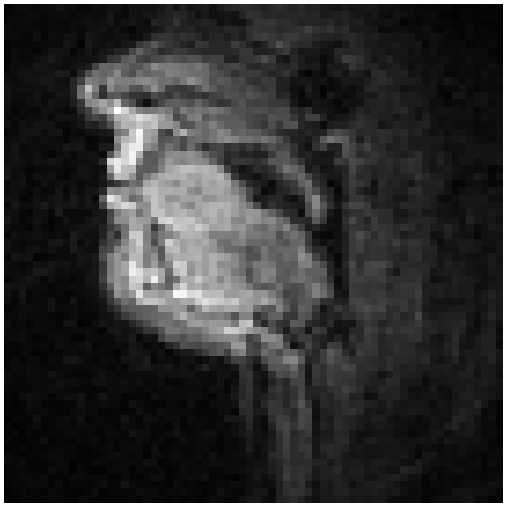
Σε αυτή την ενότητα θα περιγράψουμε τις διάφορες τεχνικές που εφαρμόσαμε προκειμένου να αντιμετωπίσουμε τα διάφορα προβλήματα που συναντήσαμε κατά τη διαδικασία εξαγωγής χαρακτηριστικών.

Ο στόχος μας είναι από κάθε πλαίσιο του βίντεο να εξάγουμε ένα διάνυσμα που θα περιγράφει όσο το δυνατόν καλύτερα τη διάταξη των διαφόρων μερών του συστήματος παραγωγής ομιλίας, χωρίς όμως το διάνυσμα αυτό να επηρεάζεται από τα γεωμετρικά χαρακτηριστικά του εκάστοτε ομιλητή. Επίσης, θέλουμε η εξαγωγή των χαρακτηριστικών να γίνεται με αυτόματο τρόπο και να μην επηρεάζεται από μικρο-αποκλίσεις στη θέση του ομιλητή που μπορεί να υπάρξουν από βίντεο σε βίντεο.

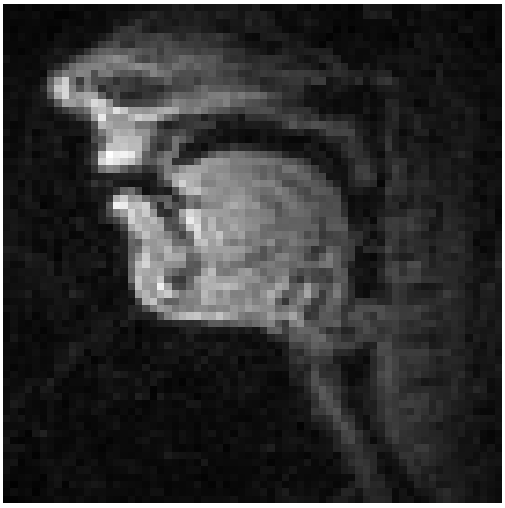
Κατά την όλη διαδικασία συναντήσαμε αρκετές δυσκολίες όπως το είδος του χαρακτηριστικού που θα επιλέξουμε. Άλλα προβλήματα που προέκυψαν δεδομένης της επιλογής του χαρακτηριστικού ήταν η αντιστοίχηση πολλαπλών διανυσμάτων σε κάθε πλαίσιο του βίντεο, η ανισορροπία δεδομένων καθώς και ο επιτυχής συνδυασμός της ακουστικής με την articulatory πληροφορία. Για την επίλυση όλων αυτών εφαρμόσαμε διάφορες μεθόδους τις οποίες περιγράφουμε αναλυτικά παρακάτω.

4.3.1 Scale Invariant Feature Transform (SIFT)

Στην προσέγγισή μας, χρησιμοποιούμε χαρακτηριστικά SIFT [12] , τα οποία είναι robust, scale invariant και θεωρείται ότι είναι τα καλύτερα για ταυτοποίηση κάτω από ποικίλες συνθήκες [53]. Επομένως, αναμένουμε ότι οι κινήσεις του κεφαλιού ή το θόλωμα των εικόνων που οφείλονται στην κακή ποιότητα των εικόνων δεν θα επηρεάσει σημαντικά το αποτέλεσμα της αναγνώρισης. Έχει επίσης παρατηρηθεί ότι τα SIFT-like χαρακτηριστικά δίνουν τα καλύτερα αποτελέσματα κατά την ταξινόμηση εικόνων [54]. Εξάλλου, SIFT θα ανιχνεύσουν την κίνηση των διαφόρων τμημάτων της



Σχήμα 4.2: Παράδειγμα βίντεο πλαισίου του ομιλητή f3



Σχήμα 4.3: Παράδειγμα βίντεο πλαισίου του ομιλητή m3

φωνητικής οδού, όπως τη γλώσσα, η οποία είναι ένας ακόμη λόγος για την επιλογή τους. Παρά το γεγονός ότι τα SIFT είναι υπολογιστικά ιδιαίτερα δαπανηρά, δεδομένου ότι η βάση δεδομένων rtMRI-TIMIT έχει σχετικά χαμηλή ανάλυση, ο χρόνος που απαιτείται για τον υπολογισμό τους δεν είναι πραγματικά πρόβλημα.

Ο αλγόριθμος SIFT αποτελείται κυρίως από τέσσερα στάδια.

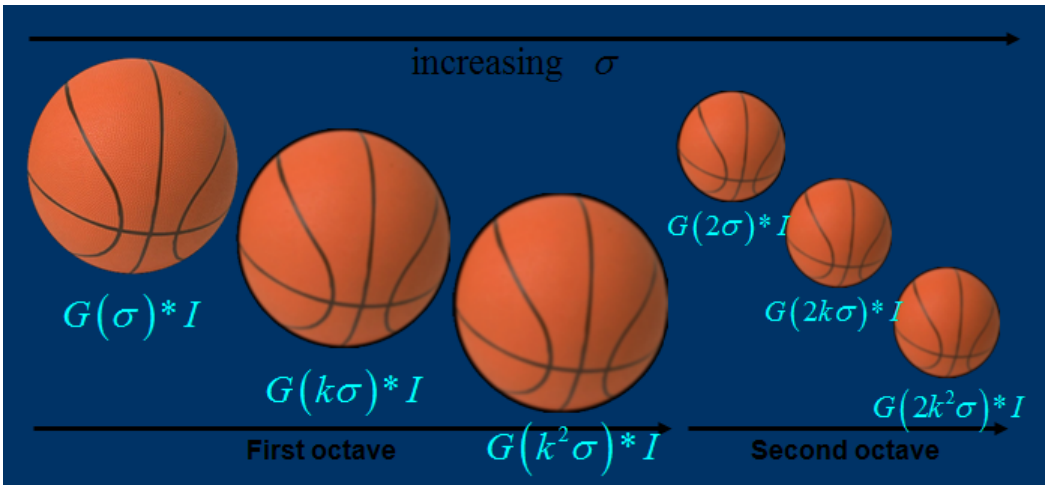
1. Ανίχνευση ακρότατων σε διαφορετικές κλίμακες
2. Ακριβής εντοπισμός σημείων ενδιαφέροντος
3. Υπολογισμός προσανατολισμού και κλίμακας των σημείων
4. Δημιουργία περιγραφικών χαρακτηριστικών

Το πρώτο βήμα είναι η εύρεση των ακρότατων σε διαφορετικές κλίμακες, δηλαδή η εύρεση των σημείων που δεν διαφέρουν ιδιαίτερα από διάφορες οπτικές γωνίες καθώς και της χαρακτηριστικής του κλίμακας σ. Αυτό επιτυγχάνεται με την εφαρμογή μίας Scale-Space συνάρτησης που βασίζεται στη Gaussian συνάρτηση.

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y)$$

Ένα εύρος τιμών της παραμέτρου θολώματος k και κλίμακας (octaves) χρησιμοποιούνται. Οι παράμετροι αυτοί, πολλαπλασιάζονται με την παράμετρο σ με αποτέλεσμα να προκύπτουν εικόνες διαφορετικής κλίμακας και με διαφορετικό επίπεδο

θόλωσης από την αρχική. Ο λόγος που το γίνεται αυτό είναι για να εντοπιστούν τα σημεία που μένουν αναλλοίωτα στο χώρο κλιμάκων. Βέλτιστη τιμή των octaves είναι τέσσερα και η βέλτιστη τιμή της παραμέτρου θόλωσης είναι πέντε. Σαν αρχικές τιμές χρησιμοποιούνται $k = \sqrt{2}, \sigma = 2$.



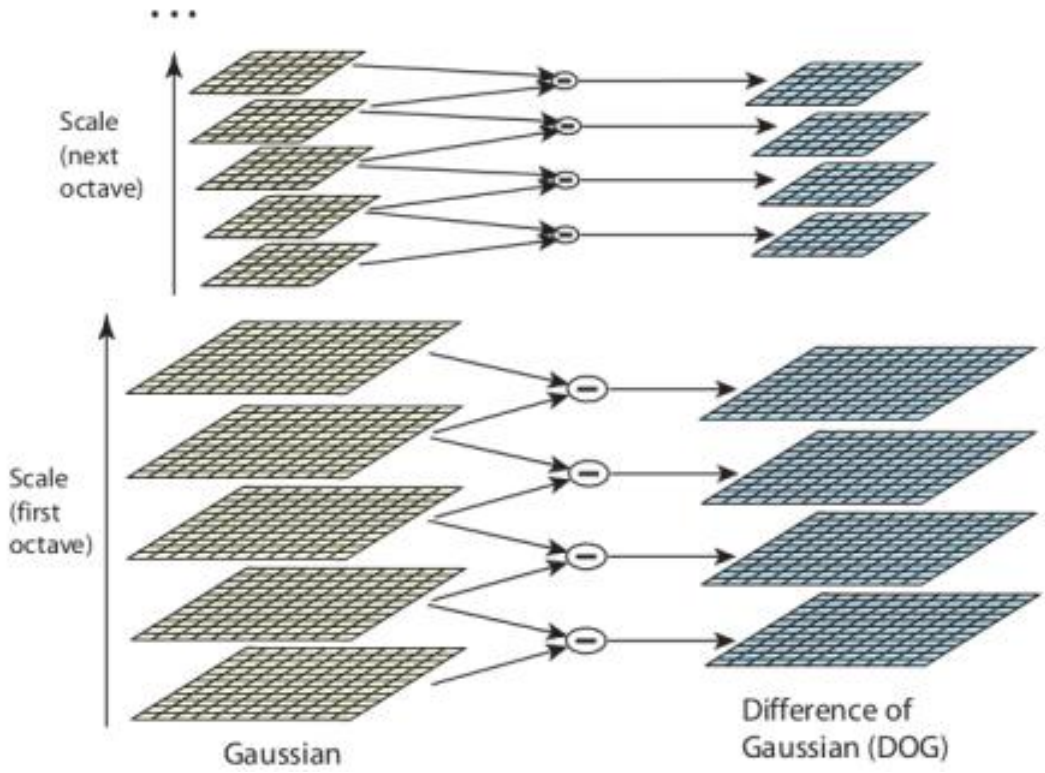
Σχήμα 4.4: Παραδείγματα εικόνων για διάφορες τιμές οκτάβων και k.

Για την επιλογή της κατάλληλης κλίμακας, η εφαρμογή στις διάφορες εικόνες του τελεστής Laplacian of Gaussian δίνει τα καλύτερα αποτελέσματα. Ωστόσο, είναι υπολογιστικά ακριβός, επομένως χρησιμοποιείται στην πράξη ο τελεστής Difference of Gaussians καθώς είναι μία πολύ καλή προσέγγιση.

$$D(x, y, \sigma) = L(x, y, m * k^n\sigma) - L(x, y, m * k^{n-1}\sigma)$$

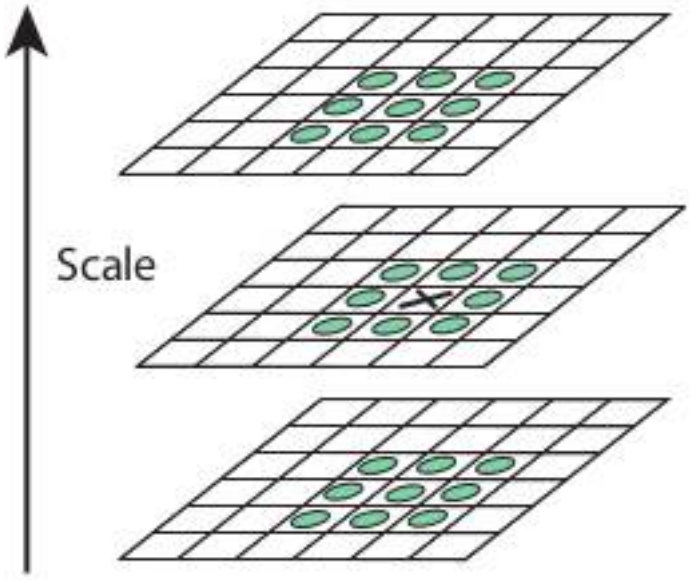
όπου m είναι η τρέχουσα octave και n το τρέχων επίπεδο θολώματος.

Η οπτική αναπαράσταση της διαδικασίας φαίνεται στην επόμενη εικόνα:



Σχήμα 4.5: Οπτική αναπαράσταση της υλοποίησης Difference of Gaussians (DoG). [12]

Για την εύρεση των τοπικών ακρότατων, συγκρίνουμε την τιμή κάθε σημείου με αυτή των γειτόνων του: οχτώ σημεία από την τρέχουσα κλίμακα και άλλα δεκαοχτώ σημεία από την επόμενη και προηγούμενη κλίμακα. Επιλέγουμε όλα τα ακρότατα που βρίσκονται σε αυτές τις $3 \times 3 \times 3$ περιοχές.



Σχήμα 4.6: Σύγκριση εντός μιας $3 \times 3 \times 3$ περιοχής για εύρεση της θέσης των ακρότατων. [12]

Το δεύτερο βήμα είναι η εύρεση στο χώρο των σημείων ενδιαφέροντος. Σε αυτό το στάδιο αφαιρούμε τα σημεία που εντοπίσαμε στο προηγούμενο στάδιο τα οποία έχουν χαμηλή αντίθεση. Αυτό επιτυγχάνεται με χρήση του αναπτύγματος Taylor για την D.

$$D(x) = D + \frac{\partial D^T}{\partial x}x + \frac{1}{2}x^T\frac{\partial^2 D^T}{\partial x^2}x$$

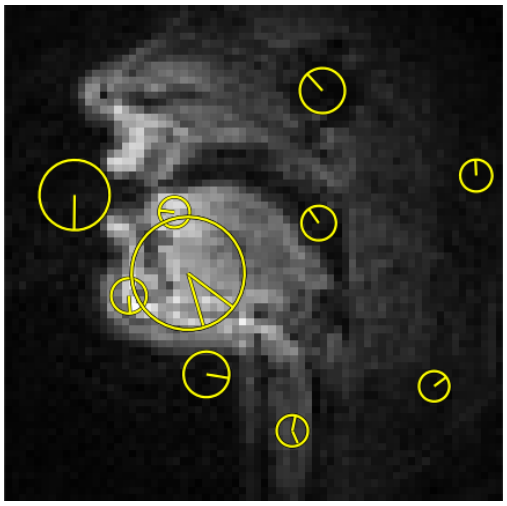
Για την εύρεση των ακροτάτων, ελαχιστοποιούμε την παραπάνω συνάρτηση.

$$z = -\frac{\partial^2 D^{-1}}{\partial x^2}\frac{\partial D}{\partial x}$$

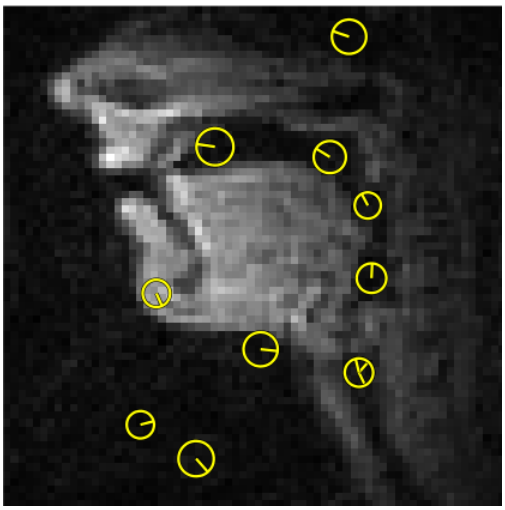
Όλα τα σημεία για τα οποία ισχύει $D(z) < 0.03$ (τιμές εικόνας στο [0,1]) αφαιρούνται καθώς έχουν χαμηλή αντίθεση. Για την εξάλειψη των ασήμαντων γωνιών υπολογίζουμε τον πίνακα Hessian της D

$$H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{yx} & D_{yy} \end{bmatrix}$$

και υπολογίζουμε τις ιδιοτιμές του. Αν ο λόγος της μεγαλύτερης ιδιοτιμής προς τη μικρότερη ιδιοτιμή στη θέση και στην κλίμακα του σημείου ενδιαφέροντος είναι μεγαλύτερος από δέκα τότε το σημείο απορρίπτεται. Τα εναπομείναντα σημεία είναι οι τελικοί SIFT Detectors.



Σχήμα 4.7: Παράδειγμα SIFT Detectors του ομιλητή f3



Σχήμα 4.8: Παράδειγμα SIFT Detectors του ομιλητή m3

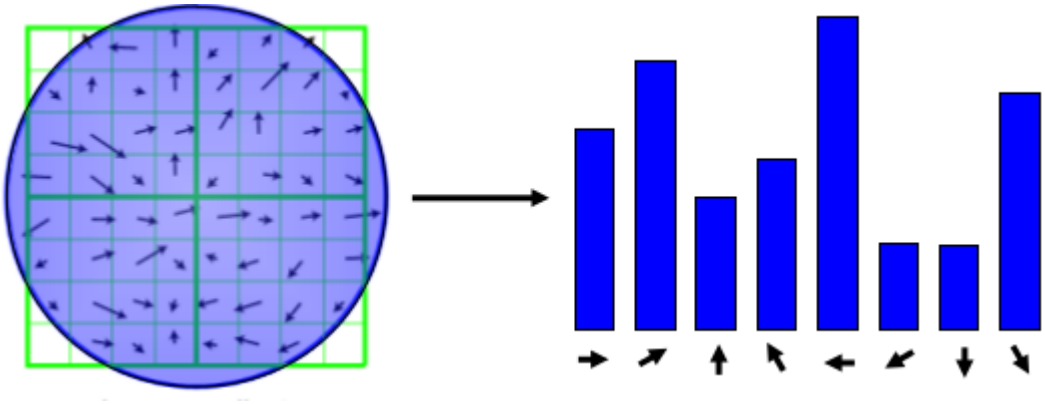
Στο τρίτο βήμα, την Ανάθεση Προσανατολισμού, θέλουμε να αφαιρέσουμε την επίδραση της περιστροφής και της διαφοράς κλίμακας που οφείλονται σε τοπικές ιδιότητες των εικόνων. Αρχικά, βρίσκουμε την σωστή εικόνα L δεδομένης της κλίμακας του σημείου. Έπειτα υπολογίζουμε το gradient magnitude m και το orientation θ.

$$m(x,y) = \sqrt{(L(x+1,y) - L(xx-1,y))^2 + (L(x,y+1) - L(x,y-1))^2}$$

$$\theta(x,y) = \tan^{-1}(\frac{L(x,y+1) - L(x,y-1)}{L(x+1,y) - L(x-1,y)})$$

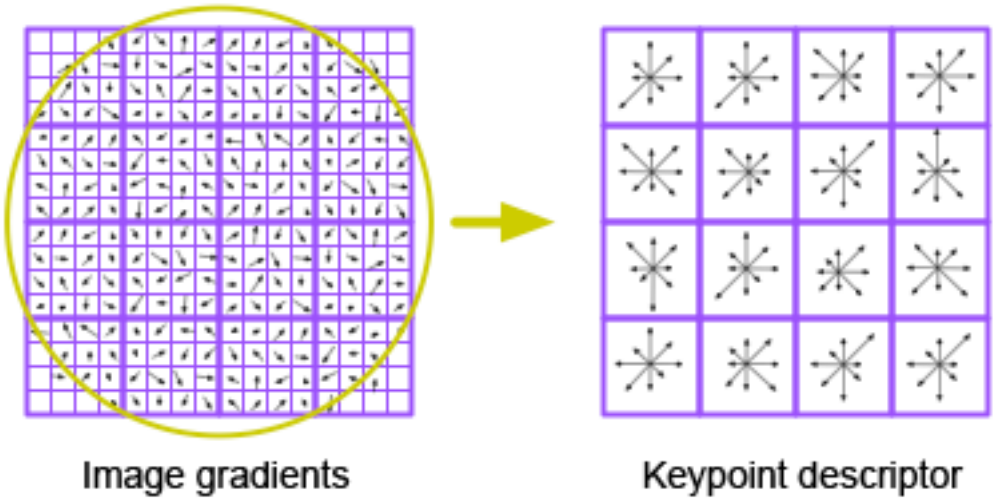
Κατασκευάζουμε το orientation ιστόγραμμα από το gradient orientation κάθε σημείου. Κάθε κορυφή με τιμή τουλάχιστον 0.8 του μέγιστου peak χρησιμοποιείται για τη δημιουργία ενός σημείου ενδιαφέροντος με αυτόν τον προσανατολισμό. Σε κάποια σημεία μπορεί να αντιστοιχιστούν περισσότεροι του ενός προσανατολισμού αλλά αυτό βοηθά ιδιαίτερα στη ευστάθεια. Τέλος, μία παραβολή προσαρμόζεται στα τρία ιστογράμματα με τις κοντινότερες τιμές σε κάθε peak με στόχο την επίτευξη μεγαλύτερης ακρίβειας όσον αφορά τη θέση των peaks.

Το τέταρτο και τελευταίο στάδιο είναι η δημιουργία των Descriptor. Η κλίση των δεδομένων που υπολογίστηκε στο προηγούμενο στάδιο, περιστρέφεται έως ότου συμπέσει με τον προσανατολισμό του σημείου ενδιαφέροντος. Στη συνέχεια σταθμίζεται



Σχήμα 4.9: Υπολογισμός προσανατολισμού. [12]

με μία Gaussian με διακύμανση 1.5 φορές τη διακύμανση της κλίμακας του σημείου. Έπειτα, αυτά τα δεδομένα χρησιμοποιούνται για να δημιουργήσουν ιστογράμματα με 4×4 δείγματα ανά παράθυρο (το οποίο επικεντρώνεται στο KeyPoint) γύρω από 8 κατευθύνσεις και η τελική διάσταση του SIFT descriptor είναι $4 \times 4 \times 8 = 128$.



Σχήμα 4.10: Περιγραφείς των σημείων ενδιαφέροντος. [12]

4.3.2 Bag of Words (BoW)

Η τεχνική αυτή πρωτοχρησιμοποιήθηκε για την ταξινόμηση εγγράφων και αργότερα προσαρμόστηκε και χρησιμοποιήθηκε στον τομέα της όρασης υπολογιστών για την ταξινόμηση εικόνων [55], [56]. Αρχικά θα περιγράψουμε πως η τεχνική BoW εφαρμόζεται στην κατηγοριοποίηση κειμένου ώστε να δώσουμε στον αναγνώστη τις βασικές ιδέες της μεθόδου, και στη συνέχεια θα εξηγήσουμε πως αυτές οι μέθοδοι εφαρμόζονται στην ταξινόμηση εικόνων. Η μέθοδος αυτή μας φάνηκε χρήσιμη στην έρευνά μας καθώς μας επέτρεψε να αντιστοιχίσουμε στο κάθε πλαίσιο του βίντεο ένα διάνυσμα που θα περιέγραφε συνολικά το πλαίσιο, βασιζόμενο στα πολλαπλά διανύσματα που είχαν αντιστοιχηθεί στο αντίστοιχο πλαίσιο από τον αλγόριθμο SIFT.

Η τεχνική BoW βασίζεται στην υπόθεση πως μπορούμε να αναπαραστήσουμε ένα δεδομένο αντικείμενο σαν ένα ιστόγραμμα κάποιων χαρακτηριστικών του [57]. Ας υποθέσουμε πως έχουμε σαν αντικείμενο ένα έγγραφο d. Κάθε διακεκριμένη λέξη του εγγράφου θεωρείται χαρακτηριστικό. Εκτός από μεμονωμένες λέξεις, κάποιος μπορεί να χρησιμοποιήσει ζεύγη δύο λέξεων (ή N λέξεων γενικά) σαν χαρακτηριστικά, ώστε να λάβει υπ΄όψην του τη χωρική κατανομή πληροφορίας μέσα στο έγγραφο [58]. Χάρην απλότητας, θα θεωρήσουμε στο παράδειγμά μας κάθε μία λέξη ξεχωριστά σαν χαρακτηριστικό. Επομένως, δημιουργούμε ένα ιστόγραμμα του κειμένου καταμετρώντας πόσες φορές εμφανίζεται κάθε λέξη στο έγγραφο. Πλέον το έγγραφο αντιπροσωπεύεται από το ιστόγραμμά του.

Συνεπώς

$$H_d = [n(w_{1,d}), n(w_{2,d}), ..., n(w_{N,d})]$$

όπου H_d είναι το ιστόγραμμα του εγγράφου d, $n(w_{i,d})$ είναι το πλήθος των φορών που η λέξη w_i εμφανίζεται στο κείμενο d και N το πλήθος των διακεκριμένων λέξεων που εμφανίζονται στο κείμενο d.

Για να συγκρίνουμε δύο έγγραφα, μπορούμε να χρησιμοποιήσουμε διαφόρων ειδών αποστάσεις για να μετρήσουμε την ομοιότητά τους (μέσω των ιστογραμμάτων τους). Στην πράξη όμως, η απόσταση του συνημιτόνου είναι αυτή που προτιμάται.

Γίνεται εύκολα αντιληπτό ότι δίνεται μεγάλη βαρύτητα στις λέξεις που εμφανίζονται πολλές φορές ενώ μικρή σε όσες εμφανίζονται λίγες. Παρ΄όλο που αυτό μοιάζει λογικό, υπάρχει ένα πρόβλημα. Στις περισσότερες περιπτώσεις, οι λέξεις που εμφανίζονται λίγες φορές είναι αυτές που έχουν μεγαλύτερη σημασία.

Για την επίλυση αυτού του προβλήματος, εφαρμόζουμε τη μέθοδο Term Frequency-Inverse Document Frequency (TF-IDF) για τη στάθμιση των όρων του ιστογράμματος έτσι ώστε οι λέξεις με χαμηλότερη συχνότητα να έχουν ισχυρότερη επίδραση στο τελικό αποτέλεσμα. Η κύρια ιδέα είναι πως κάθε λέξη που εμφανίζεται και στα δύο αρχεία δεν είναι καλή επιλογή όσον αφορά τη διάκριση των δύο αρχείων.

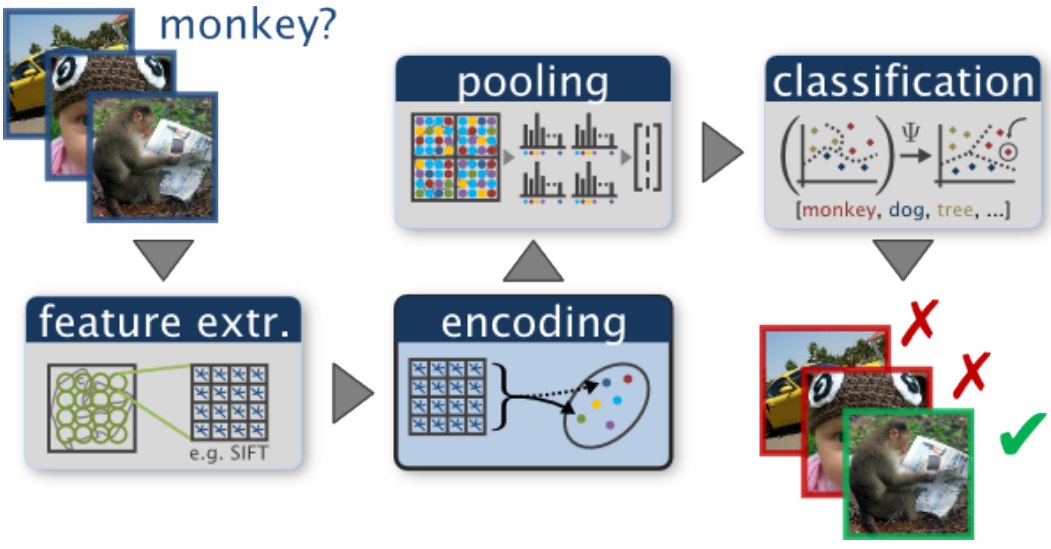
Κάθε όρος $n(w_{i,d})$ του ιστογράμματος H_d πολλαπλασιάζεται με ένα συντελεστή βάρους q_i ο οποίος υπολογίζεται από τον ακόλουθο τύπο:

$$q_i = \log \frac{D}{1 + \{d : w_{i,d} \in d\}}$$

όπου D είναι το πλήθος των εγγράφων που συγκρίνουμε και $\{d : w_{i,d} \in d\}$ είναι το πλήθος των αρχείων που εμφανίζεται η λέξη $w_{i,d}$.

Εφαρμογή της μεθόδου BoW στην κατηγοριοποίηση εικόνων

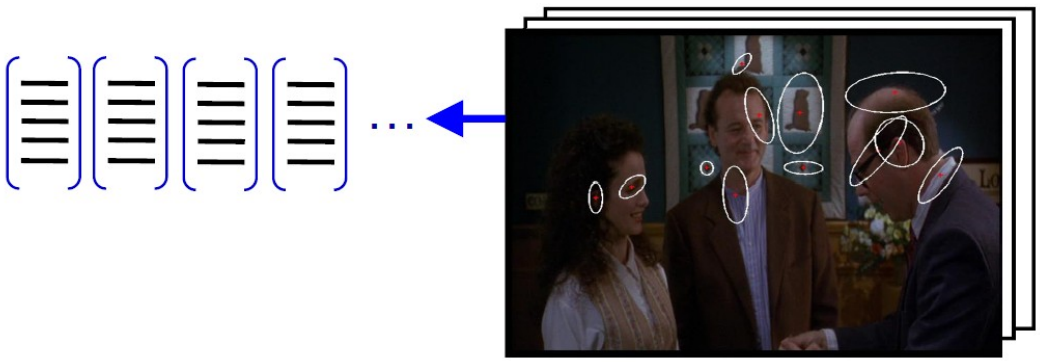
Τώρα θα περιγράψουμε πως εφαρμόζονται οι ιδέες που περιγράψαμε προηγουμένως σε προβλήματα κατηγοριοποίησης εικόνων [59]. Η διαδικασία αποτελείται κυρίως από τέσσερα στάδια. Μία συνοπτική περιγραφή των σταδίων αυτών είναι η ακόλουθη: Στο πρώτο βήμα γίνεται η εξαγωγή χαρακτηριστικών. Οι SIFT descriptors είναι μία συνηθισμένη επιλογή. Μια άλλη συνηθισμένη επιλογή είναι οι SURF descriptor [60]. Σαφώς υπάρχουν και άλλες επιλογές όπως περιγράφεται εδώ [61]. Το δεύτερο στάδιο περιλαμβάνει την εκμάθηση του οπτικού λεξιλογίου. Στον N-dimension χώρο, όπου N είναι η διάσταση του descriptor, εφαρμόζουμε k-means clustering σε όλους τους descriptors όλων των εικόνων. Με αυτόν τον τρόπο αποκτάμε k σημεία N διαστάσεων το καθένα τα οποία αποτελούν τα κέντρα των κλάσεων που δημιουργήθηκαν και ταυτόχρονα είναι και οι λέξεις του λεξικού μας. Το οπτικό λεξικό λέγεται και code- book. Το τρίτο στάδιο περιλαμβάνει την αντιστοίχιση του κάθε descriptor των εικόνων με την κοντινότερη λέξη του codebook. Το τελικό στάδιο είναι αντιστοίχιση κάθε εικόνας με ένα ιστόγραμμα k διαστάσεων που δείχνει πόσοι από τους N διαστάσεων descriptors της εικόνας αντιστοιχήθηκαν στην i- λέξη του λεξικού. Συνήθως μετά από το σύστημα BoW εκπαιδεύονται κάποια Support Vector Machines (SVMs) για κατηγοριοποίηση εικόνων(χρησιμοποιώντας τα ιστογράμματα σαν χαρακτηριστικά εισόδου).



Σχήμα 4.11: Οπτική αναπαράσταση της διαδικασίας BoW. [13]

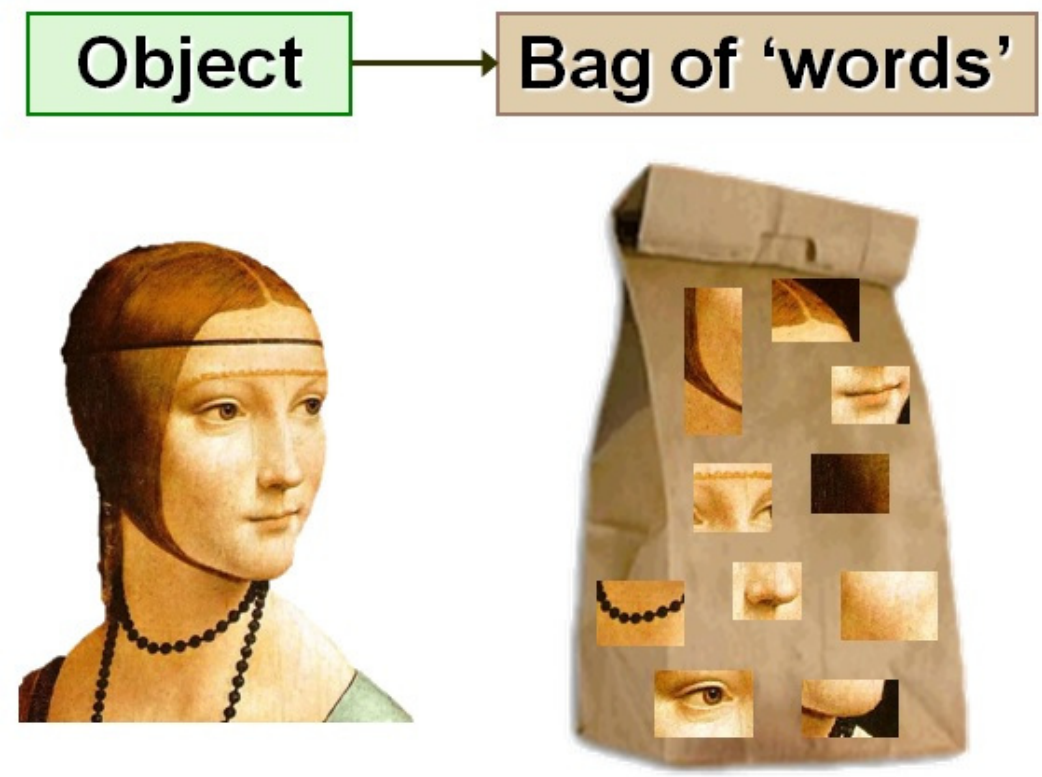
Το πρώτο στάδιο είναι η εύρεση των σημείων ενδιαφέροντος στις εικόνες και ο

μετασχηματισμός αυτής της πληροφορίας σε ένα διάνυσμα έτσι ώστε η ίδια εικόνα(ή παρόμοιες εικόνες) από διαφορετικές οπτικές γωνίες να έχουν περίπου ίδια διανύσματα, ενώ διαφορετικές εικόνες να έχουν εντελώς διαφορετικά διανύσματα. Αυτά τα σημεία μας βοηθάνε στο να εντοπίσουμε το είδος των αντικειμένων που δείχνει η εικόνα. Όπως αναφέρθηκε και νωρίτερα, οι SIFT descriptors είναι μία καλή επιλογή.



Σχήμα 4.12: Εξαγωγή χαρακτηριστικών. [13]

Στο Σχήμα 4.12 βλέπουμε ένα παράδειγμα εικόνας όπου απεικονίζονται τα σημεία ενδιαφέροντος και οι μετασχηματισμοί τους σε διανύσματα. Παρατηρούμε ότι ο αλγόριθμος που χρησιμοποιείται για την εύρεση των σημείων ενδιαφέροντος έχει εντοπίσει μία περιοχή πάνω από το φρύδι δύο ανθρώπων. Περιμένουμέ ότι αυτές οι δύο περιοχές θα έχουν παρόμοια διανύσματα, με μηδαμινές διαφορές, καθώς αντιστοιχούν στην ίδια περιοχή του ανθρώπινου προσώπου και επομένως αναμένουμε να αντιστοιχούν στην ίδια λέξη.



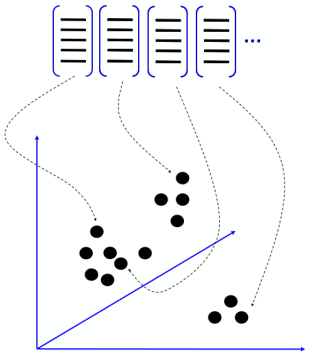
Σχήμα 4.13: Οπτική αναπαράσταση μιας οπτικής λέξης. [13]

Το στάδιο δύο αφορά τη δημιουργία του οπτικού λεξιλογίου. Με άλλα λόγια, μελετάμε τα διανύσματα που προέκυψαν από τα σημεία ενδιαφέροντος όλων των εικόνων και προσπαθούμε να εντοπίσουμε ποια από αυτά τα σημεία αντιστοιχούν σε παρόμοιες περιοχές(ή παρόμοια χαρακτηριστικά όπως το ανθρώπινο μάτι) και ποια όχι. Το οπτικό λεξιλόγιο αποτελείται από όλες τις διαφορετικές λέξεις(παρόμοιες περιοχές) που βρίσκονται με αυτόν τον τρόπο. Ένα παράδειγμα οπτικού λεξιλογίου φαίνεται στο Σχήμα 4.15

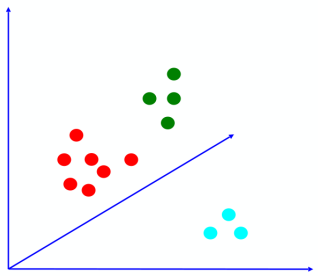
Σχήμα 4.14: Οπτικές λέξεις από διάφορες εικόνες. [13]

Σχήμα 4.15: Οπτικό λεξικό. [13]

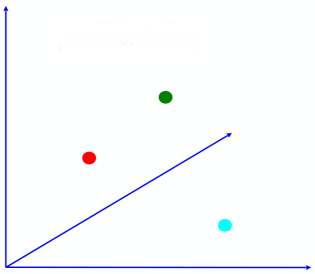
Για τη δημιουργία του οπτικού λεξιλογίου, σχεδιάζουμε τα N-διαστάσεων διανύσματα ($N = 128$ στην περίπτωση των SIFT descriptors) και εφαρμόζουμε k-means clustering. Το κέντρο κάθε κλάσης αντιστοιχεί και σε μία διαφορετική οπτική λέξη. Μια οπτική αναπαράσταση της όλης διαδικασίας φαίνεται στις επόμενες εικόνες.



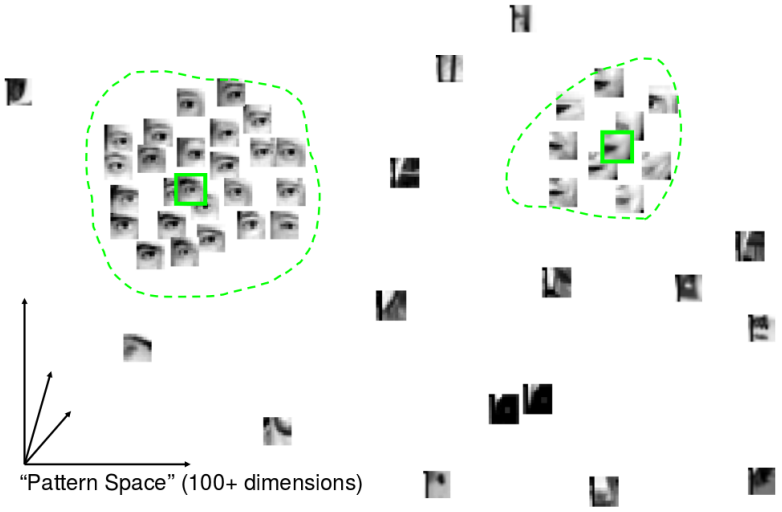
Σχήμα 4.16: Γραφική αναπαράσταση των διανυσμάτων. [13]



Σχήμα 4.17: Εφαρμογή k-means στα διανύσματα [13]



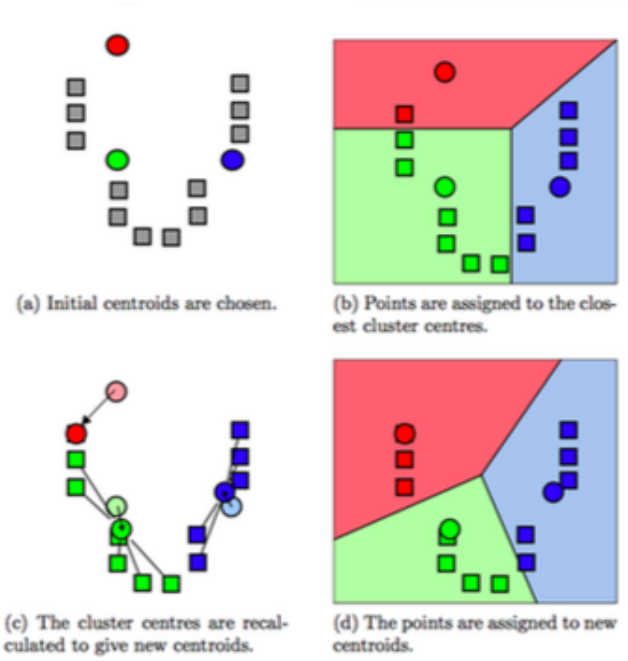
Σχήμα 4.18: Το κέντρο κάθε κλάσης απαρτίζει το οπτικό λεξιλόγια. [13]



Σχήμα 4.19: Συνοπτικό παράδειγμα παρουσίαση της διαδικασίας ομαδοποίησης. [13]

Ο αλγόριθμος k-means

Σε αυτό το σημείο θα περιγράψουμε πως δουλεύει ο αλγόριθμος k-means, ο οποίος είναι επαναληπτικός. Το πρώτο πράγμα που πρέπει να κάνουμε είναι να καθορίσουμε την τιμή της παραμέτρου k και να επιλέξουμε τα κατάλληλα σημεία σαν αρχικά κέντρα (k στο σύνολο). Κάθε διάνυσμα αντιστοιχείται στην κλάση με το κοντινότερο κέντρο. Συνήθως η Ευκλείδια απόσταση χρησιμοποιείται για τον υπολογισμό και τη σύγκριση των αποστάσεων. Το νέο κέντρο κάθε κλάσης υπολογίζεται σαν ο μέσος όρος των σημείων που αντιστοιχήθηκαν σε αυτή την κλάση. Η διαδικασία επαναλαμβάνεται μέχρις ότου να μην αλλάξει κανένα διάνυσμα κλάση.



Σχήμα 4.20: Πως δουλεύει ο αλγόριθμος k-means. [14]

Υπάρχουν δύο προβλήματα που πρέπει να επιλυθούν όταν εφαρμόζουμε k-means clustering. Το πρώτο είναι η επιλογή του πλήθους των κλάσεων που πρέπει να χρησιμοποιηθούν και το δεύτερο είναι με ποια σημεία θα επιλέξουμε σαν αρχικά κέντρα των κλάσεων.

Για το πρώτο πρόβλημα (εύρεση βέλτιστου k), μπορεί κανείς να σκεφτεί ότι μπορεί εύκολα να να λυθεί με την χρήση μιας συνάρτηση σφάλματος, για παράδειγμα, η μέση απόσταση των σημείων κάθε ομάδας από το κέντρο τους, και την ελαχιστοποίηση. Αυτό θα οδηγούσε στη επιλογή της παραμέτρου k ίση με τον αριθμό των δειγμάτων εκπαίδευσης, το οποία δεν έχει κανένα νόημα καθώς το σύστημα θα γινόταν υπέρ-εκπαιδευμένο. Καλύτεροι μέθοδοι έχουν αναπτυχθεί για την επίλυση αυτού του προβλήματος. Θα εξηγήσουμε συνοπτικά δύο ευριστικούς αλγορίθμους οι οποίοι βασίζονται στη μελέτη της μεταβολής της συνεκτικότητας των κλάσεων για τις διάφορες τιμές του k [62].

Η μέθοδος elbow: Ο πρώτος αλγόριθμος είναι η μέθοδος elbow. Βασίζεται στην απόσταση μεταξύ των σημείων μιας κλάσης και του κέντρου τους. Ο αλγόριθμος λειτουργεί ως εξής: Εκτελούμε τον αλγόριθμο k-means για διάφορες (συνεχόμενες στις περισσότερες περιπτώσεις) τιμές του k, για παράδειγμα, 1 ως 10. Για κάθε τιμή του k, υπολογίζουμε και σχεδιάζουμε το άθροισμα των τετραγώνων των αποστάσεων

(error) SSE ανάμεσα στα σημεία και το κέντρο στο οποίο αντιστοιχήθηκαν.

Ο μαθηματικός τύπος είναι

$$SSE = \sum_{i=1}^{n} (x_i - y_{x_i})^2$$

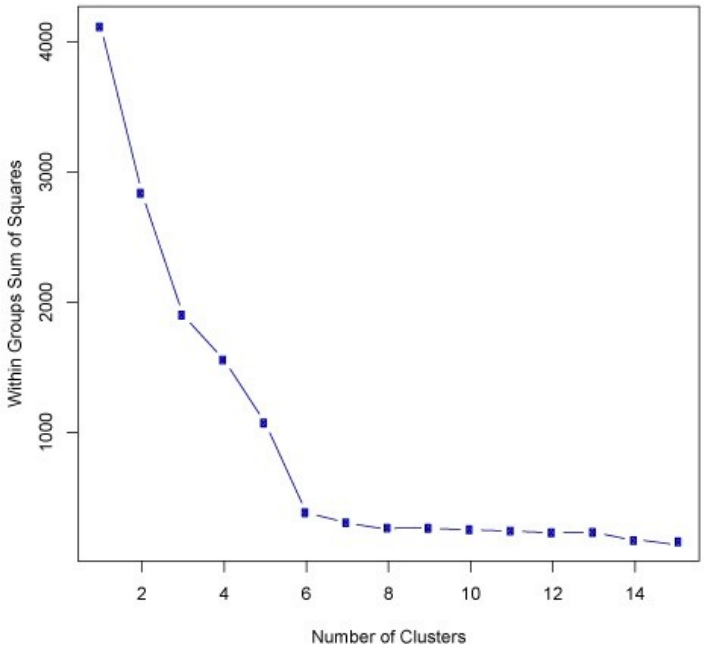
όπου

n είναι το πλήθος των δεδομένων εκπαίδευσης,

x_i είναι το τρέχων δείγμα,

y_{x_i} είναι το κέντρο της κλάσης στην οποία αντιστοιχήθηκε το διάνυσμα x_i.

Από το (k, SSE) γράφημα υπολογίζουμε την τιμή του k για την οποία μία αύξηση στο πλήθος των κλάσεων δεν έχει σαν αποτέλεσμα σημαντική μείωση στην τιμή του SSE. Ένα παράδειγμα φαίνεται στο Σχήμα 4.21. Καθώς το πλήθος των κλάσεων αυξάνεται, η τιμή SSE μειώνεται ραγδαία. Όμως, για τιμές του k μεγαλύτερες του έξι, δεν υπάρχει αξιόλογη μείωση του SSE επομένως η τιμή του $k = 6$ θεωρείται μία καλή επιλογή για το πλήθος των κλάσεων του συγκεκριμένου προβλήματος.



Σχήμα 4.21: Επιλογή μιας καλής τιμής για το πλήθος των κλάσεων εφαρμόζοντας τη μέθοδο ($k = 6$) elbow στο γράφημα (k, SSE). [15]

Η δεύτερη μέθοδος είναι η μέθοδος silhouette. Αυτή η μέθοδος βασίζεται στο πόσο όμοιο είναι ένα αντικείμενο με την κλάση του, σε σχέση με τις υπόλοιπες κλάσεις. Οι τιμές του Silhouette κυμαίνονται από μείον ένα έως ένα. Όσο κοντινότερες είναι οι τιμές του silhouette στη μονάδα, τόσο μεγαλύτερη είναι η ομοιότητα αυτού του σημείου με την κλάση του(καθώς και η πιθανότητα αυτό το σημείο να έχει αντιστοιχηθεί στη σωστή κλάση). Όσο μειώνεται η τιμή του silhouette, τόσο μειώνεται και η ομοιότητα του σημείου με την κλάση του. Αρνητικές τιμές silhouette πιθανότατα υποδηλώνουν λανθασμένο αριθμό επιλογής κλάσεων. Για την εκτέλεση του αλγορίθμου, εκτελούμε τον αλγόριθμο k-means για πολλές τιμές της παραμέτρου k. Η ομοιότητα (Silhouette value) κάθε σημείου i με την κλάση του υπολογίζεται χρησιμοποιώντας τον ακόλουθο τύπο:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

όπου

$a(i)$ είναι η μέση απόσταση ανάμεσα στο σημείο i και στα υπόλοιπα σημεία της

κλάσης του.

$b(i)$ είναι η μέση απόσταση μεταξύ του σημείου i και των σημείων της κοντινότερης κλάσης (εξαιρουμένης της κλάσης που ανήκει το σημείο i) στο σημείο i.

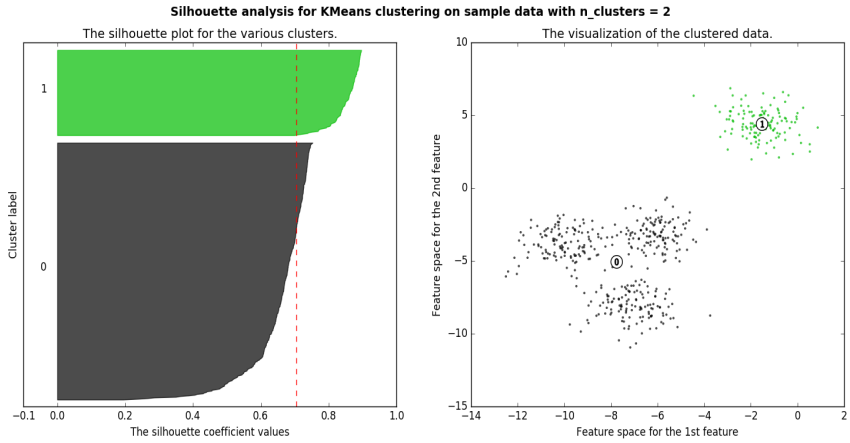
Για τη μέτρηση της απόστασης συνήθως χρησιμοποιούνται η Ευκλείδια ή η Manhattan απόσταση.

Έχοντας υπολογίσει τη μέση απόσταση κάθε σημείου για τις διάφορες τιμές του k, έχουμε δύο επιλογές:

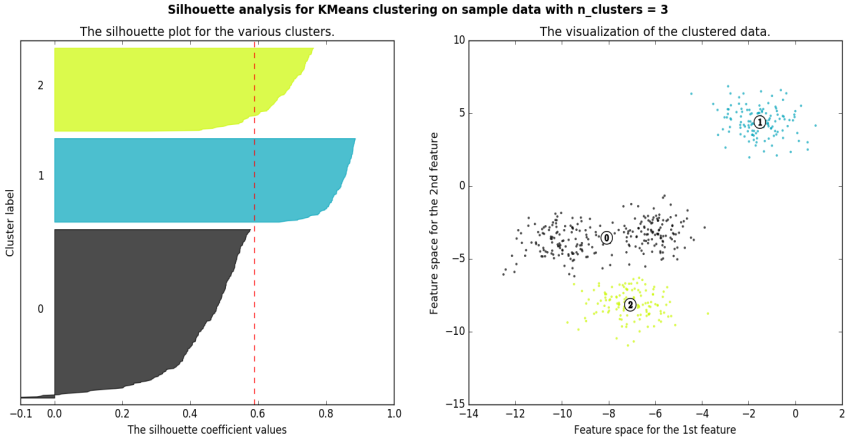
Η πρώτη είναι να σχεδιάσουμε για κάθε σημείο κάθε κλάσης της τιμές του silhouette, και να βρούμε για ποια τιμή του k έχουν τα περισσότερα σημεία τιμή του silhouette μεγαλύτερη από ένα όρια(το όριο αυτό συνήθως βρίσκεται ανάμεσα στο $0.6 - 0.8$).

Η δεύτερη επιλογή είναι να υπολογίσουμε τη μέση τιμή του silhouette για όλα τα σημεία και να φτιάξουμε ένα γράφημα με την τιμή του silhouette ανά πλήθος κλάσεων ((k, sil) graph). Αναζητούμε στο γράφημα την τιμή του k στη οποία υπάρχει κάποιο peak. Αν δεν είναι εμφανές κάποια κορυφή, τότε πιθανότατα απαιτείται να εξετάσουμε επιπλέον τιμές του k.

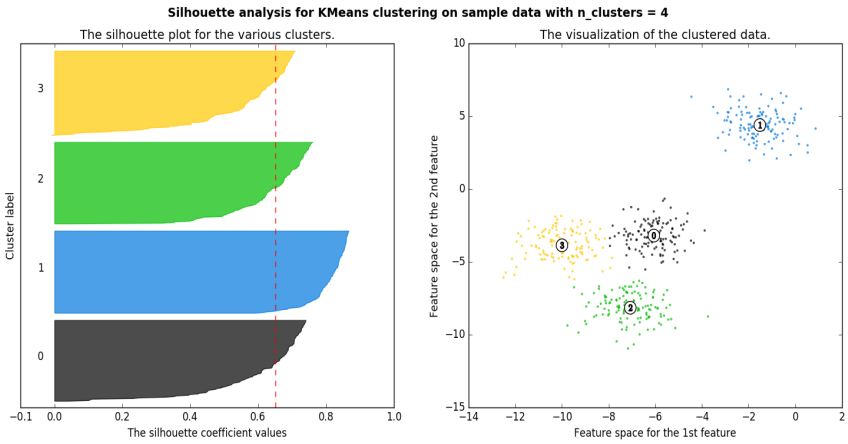
Παραδείγματα της επιλογής 1 παρουσιάζονται στις ακόλουθες εικόνες



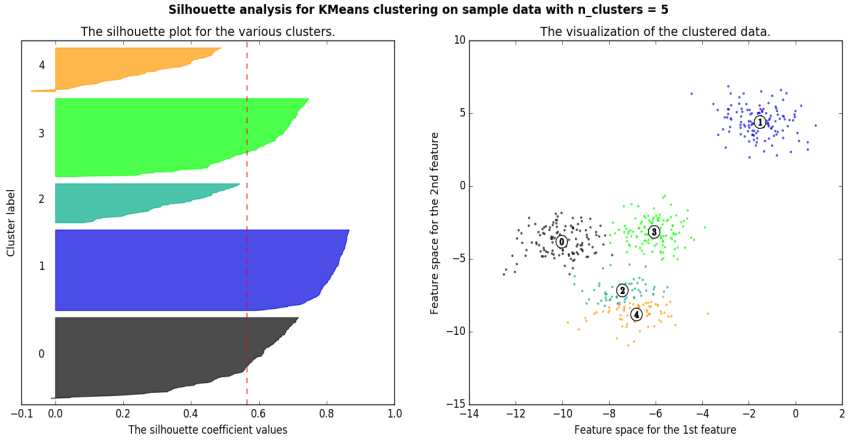
Σχήμα 4.22: Επιλογή 1 για τη μέθοδο silhouette με πλήθος κλάσεων $k = 2$. [16]



Σχήμα 4.23: Επιλογή 1 για τη μέθοδο silhouette με πλήθος κλάσεων $k = 3$. [16]



Σχήμα 4.24: Επιλογή 1 για τη μέθοδο silhouette με πλήθος κλάσεων $k = 4$. [16]



Σχήμα 4.25: Επιλογή 1 για τη μέθοδο silhouette με πλήθος κλάσεων $k = 5$. [16]

Όπως παρατηρούμε στο γράφημα των δεδομένων(δεξί μέρος των Σχημάτων 4.22 - 4.25), υπάρχουν δύο κύριες περιοχές δεδομένων στο παράδειγμά μας, η πάνω δεξιά περιοχή (περιοχή 1) και η κάτω αριστερά περιοχή (περιοχή 2). Η περιοχή 2 έχει αρκετά περισσότερα σημεία σε σύγκριση με την περιοχή 1, σχεδόν τα τριπλάσια. Η περιοχή 1 αποτελεί ξεκάθαρα μία κλάση από μόνη της. Από την άλλη μεριά, η περιοχή 2 μπορεί να διαιρεθεί σε τρεις υπό-περιοχές όπου κάθε υπό-περιοχή να έχει το ίδιο πλήθος δεδομένων με την περιοχή 1. Έτσι, οι πιθανότερες τιμές κλάσεων είναι είτε $k = 2$ είτε $k = 4$.

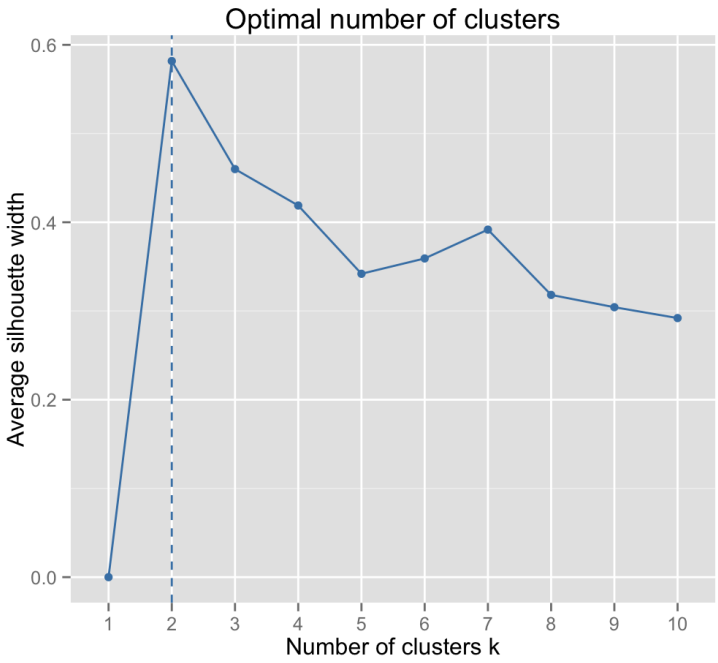
Ωστόσο, ένα σύνολο δεδομένων μπορεί να μην είναι 2-διαστάσεων και επομένως εύκολο να σχεδιαστεί ώστε να εξάγουμε τα παραπάνω συμπεράσματα αναλύοντας απλά το γράφημα. Ευτυχώς, όλες αυτές οι παρατηρήσεις μπορούν να εξαχθούν από το silhouette γράφημα.

Παρατηρούμε ότι η μέση τιμή του silhouette (η κόκκινη γραμμή στο γράφημα silhouette) είναι πάνω από 0.6 μόνο για $k = 2$ και $k = 4$. Επομένως τα δεδομένα μπορούν να διαιρεθούν σε 2 ή 4 κλάσεις. Η silhouette τιμή για $k = 2$ είναι μεγαλύτερη από την τιμή για $k = 4$. Υποπτευόμαστε ότι τα δεδομένα διαιρούνται κυρίως σε δύο κατηγορίες. Καθώς το πλάτος της γκρίζας περιοχής στον Ψ-άξονα είναι περίπου τρεις φορές μεγαλύτερο από το πλάτος της πράσινης περιοχής (για $k = 2$ silhouette γράφημα), η γκρίζα περιοχή έχει περίπου τρεις φορές περισσότερα σημεία. Κοιτάζοντας το γράφημα της άλλου πιθανού πλήθους κλάσεων($k = 4$) παρατηρούμε ότι η μπλε περιοχή έχει ακριβώς το ίδιο σχήμα με την πράσινη περιοχή στο προηγούμενο γράφημα για $k = 2$. Παρατηρώντας προσεχτικά τα γραφήματα και για τις υπόλοιπες τιμές του k βρίσκουμε ότι σε όλα υπάρχει αυτό το σχήμα. Συνεπώς για όλες τις τιμές του k αντιστοιχίζεται στην ίδια περιοχή σημείων (περιοχή 1) η οποία, όπως προκύπτει από τα γραφήματα, είναι εμφανώς διαχωρισμένη από τα υπόλοιπα σημεία. Επομένως τα σημεία που μπορούν να διαιρεθούν επιπλέον είναι μόνο αυτά που αντιστοιχούν στην γκρίζα περιοχή για $k = 2$ και μπορούν να διαιρεθούν σε τρεις υπό-περιοχές(κίτρινη, πράσινη, γκρίζα για $k = 4$).

Η επιλογή 2 προσφέρει πολύ λιγότερες πληροφορίες για το τι συμβαίνει εσωτερικά της κάθε κλάσης. Μπορούμε μόνο να πάρουμε μία γενική εικόνα για το πόσο καλά το πλήθος των κλάσεων ταιριάζει στο πρόβλημά μας. Πληροφορίες σχετικά με την ισορροπία του πλήθους δεδομένων ή τη διαχωρισιμότητα των κλάσεων δεν γίνεται να εξαχθούν με κάποιο τρόπο. Το πλεονέκτημα όμως αυτής της επιλογής είναι η βολικότητα και η ευκολία της. Στην πράξη, όταν κάποιος επιλέγει να χρησιμοποιήσει τη μέθοδο silhouette, συνήθως χρησιμοποιεί τη δεύτερη επιλογή πρώτα, και αν για κάποιο λόγο χρειάζεται περισσότερες πληροφορίες για την επιλογή της τιμής του k, τότε χρησιμοποιεί την πρώτη επιλογή.

Παράδειγμα της επιλογής 2 φαίνονται παρακάτω

Σε αυτό το σημείο μπορεί κάποιος να αναρωτηθεί: Άν έχω δύο καλές τιμές του k



Σχήμα 4.26: Επιλογή 2 για τη μέθοδο silhouette. [16]

ποια να επιλέξω΄ ;

Η απάντηση είναι ¨εξαρτάται¨. Η καλή κατανόηση του προς επίλυση προβλήματος είναι απαραίτητη ώστε να παρθεί η σωστή απόφαση. Όταν κάποιος αντιμετωπίζει ένα πρόβλημα σαν κι αυτό, συνήθως έχει μια εκτίμηση για το πλήθος των κλάσεων που απαιτούνται, επομένως μια λύση στο πρόβλημα είναι να επιλέξει την τιμή του k που είναι πιο κοντά στην εκτίμησή του. Παρόλο που η επιλογή της τιμής του k με τη μεγαλύτερη silhouette τιμή είναι συχνά μια καλή επιλογή ($k = 2$ στο παράδειγμά μας), η επιλογή μιας διαφορετικής τιμής του k με καλή τιμή silhouette, προσφέρει ενδεχομένως άλλα πλεονεκτήματα όπως ισορροπία πλήθους δεδομένων ανάμεσα στις κλάσεις (όπως φαίνεται για k=4 στο παράδειγμά μας, το πλάτος κάθε χρωματιστής περιοχής στον Y-άξονα του γραφήματος silhouette είναι περίπου ίδιο) και η επιλογή αυτής της τιμής του k μπορεί να είναι και καλύτερη για κάποια προβλήματα. Θυσιάζοντας ελαφρώς τη συνεκτικότητα των κλάσεων με στόχο την επίτευξη ισορροπημένου πλήθους δεδομένων μπορεί να μας εξοικονομήσει χρόνο και να μας γλυτώσει από διάφορα προβλήματα. Μερικές φορές ωστόσο, είναι αναπόφευκτη η ύπαρξη ανισόρροπου πλήθους δεδομένων. Το γεγονός αυτό μπορεί να προκαλέσει αρκετά προβλήματα και διάφορες μέθοδοι έχουν εφευρεθεί για την επίλυσή τους [63]. Θα αναφερθούμε εκτενώς σε αυτές τις μεθόδους αργότερα.

Ένα πρόβλημα που μπορεί να προκύψει κατά τη διαδικασία εύρεσης του πλήθους των κλάσεων, είναι το πλήθος των κλάσεων να είναι υπερβολικά μεγάλο, για παράδειγμα $k = 100$ ή ακόμα και $k = 1000$ ή και παραπάνω. Σε αυτές τις περιπτώσεις, δεν είναι ιδιαίτερα αποτελεσματικό να γίνουν όλοι οι υπολογισμοί με τον τρόπο που περιγράφηκαν παραπάνω(ισχύει και για τις δύο μεθόδους αυτό) για όλες τις πιθανές τιμές της παραμέτρου k μεταξύ $1 - 500$ για παράδειγμα. Αντί να υπολογίσουμε το SSE (μέθοδος 1) για συνεχόμενες τιμές του k, υπολογίζουμε τις τιμές του SSE για τιμές του k με βήμα 25 για παράδειγμα ($k = 2, k = 27, k = 52, ...$). Στη συνέχεια κάνουμε τα ίδια ακριβώς με αυτά που περιγράψαμε στην προηγούμενη περίπτωση με διαδοχικούς αριθμούς. Βρίσκουμε τη μικρότερη τιμή του k (μία εκτίμηση σε αυτήν την περίπτωση π.χ. $k = 75$) για την οποία μία αύξηση του πλήθους των κλάσεων δεν συνεπάγεται σημαντική μείωση στην τιμή του SSE. Έπειτα, ανάλογα βέβαια και με το πόσο μεγάλο είναι το αρχικό μας βήμα, συνεχίζουμε τη διαδικασία είτε με συνεχόμενες χτιμές (π.χ. $65, 66, 67, ..., 75, ...84, 85$) είτε με μικρότερο βήμα (π.γ. $60, 65, 70, 75, .., 90$), γύρω από την τιμή της εκτίμησής μας, μέχρις ώτου βρούμε την ακριβή τιμή του k που ψάχνουμε. Παρόλο που η μέθοδος silhouette δίνει περισσότερες πληροφορίες για το πως είναι οι κλάσεις, συνήθως προτιμάται η μέθοδος ελβοω ιδιαίτερα σε περιπτώσεις αυξημένου πλήθους k, εξ' αιτίας του χαμηλότερου υπολογιστικού κόστους.

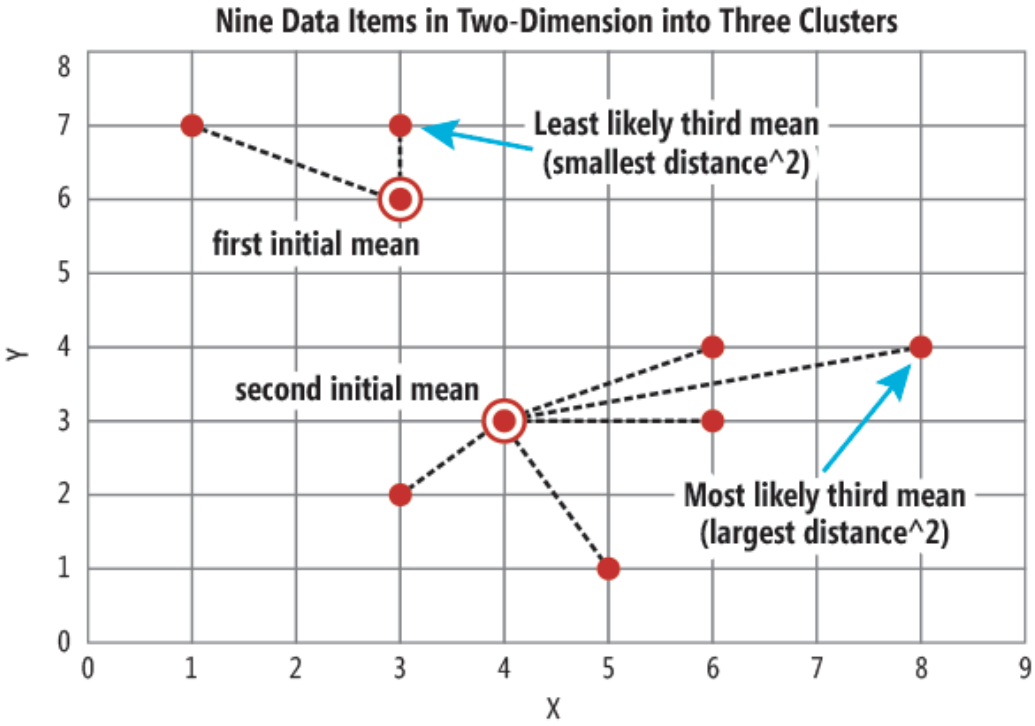
Αρχικά κέντρα κλάσεων

Ένα άλλο πρόβλημα είναι η επιλογή των αρχικών σημείων για κέντρα των κλάσεων. Η επιλογή των αρχικών κέντρων παίζει σημαντικό ρόλο στο πόσο καλή θα είναι η λύση στην οποία θα συγκλίνει ο αλγόριθμος k-means, καθώς και πόσο γρήγορα θα φτάσει σε αυτή τη λύση. Διάφορες μέθοδοι έχουν αναπτυχθεί, κάθε μία με τα πλεονεκτήματά της. Θα περιγράψουμε συνοπτικά τις κυριότερες από αυτές.

Μια δημοφιλής προσέγγιση είναι η επιλογή των αρχικών κέντρων στην τύχη. Αυτή η επιλογή δεν είναι τόσο κακή όσο ακούγεται, ωστόσο έχουν εφευρεθεί και βελτιωμένες τεχνικές. Μία άλλη επιλογή είναι να τρέξουμε τον αλγόριθμο k-means με τυχαία αρχικά σημεία, χρησιμοποιώντας μόνο το $20\% - 30\%$ των δεδομένων, και στη συνέχεια να επιλέξουμε τα τελικά κέντρα, σαν αρχικά κέντρα για την εκτέλεση του αλγορίθμου k-means σε όλα τα δεδομένα. Σαφώς μπορεί κάποιος να εφαρμόσει τις παραπάνω τεχνικές αρκετές φορές και να κρατήσει σαν αρχικά σημεία αυτά που έδωσαν τα καλύτερα αποτελέσματα. Επιπροσθέτως, πιο περίπλοκες τεχνικές έχουν αναπτυχθεί για την επίλυση αυτού του προβλήματος όπως k-means++ και canopy.

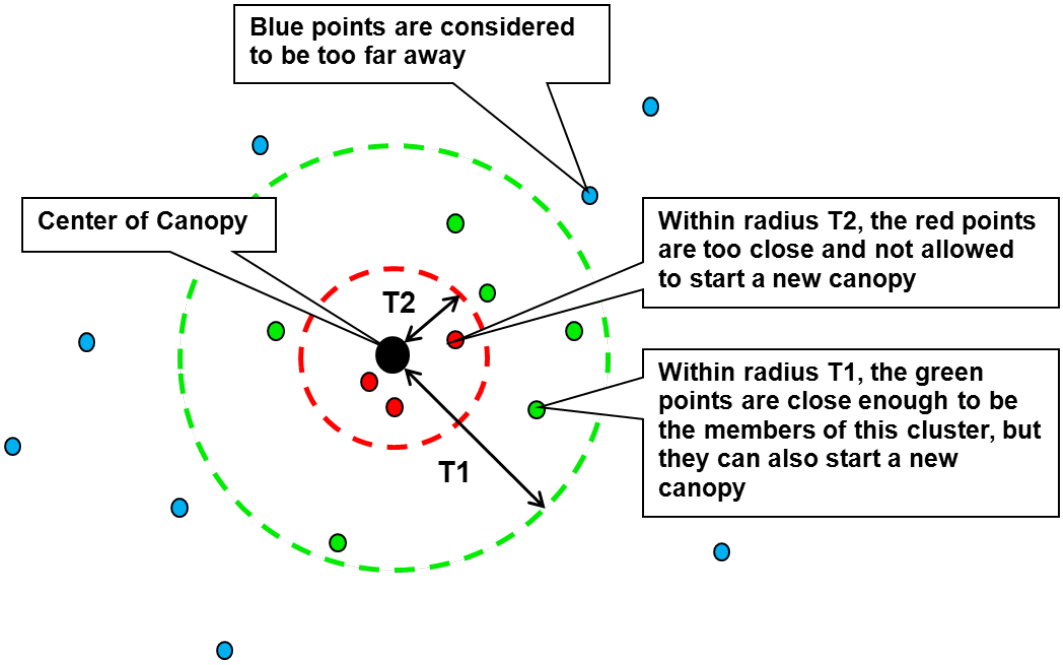
K-means++ βασίζεται στην ιδέα ότι τα αρχικά κέντρα θα πρέπει να είναι όσο το δυνατόν πιο εξαπλωμένα. Αρχικά επιλέγουμε στην τύχη σαν αρχικό κέντρο ένα σημείο από τα δεδομένα μας X. Για κάθε σημείο x, υπολογίζουμε την (Ευκλείδια) απόσταση $D(x)$ ανάμεσα στο x και στο κοντινότερο κέντρο από αυτά που έχουμε ήδη επιλέξει. Επιλέγουμε ένα νέο σημείο x σαν καινούριο κέντρο, χρησιμοποιώντας μία σταθμισμένη κατανομή πιθανότητας $\frac{D(x)^2}{\sum_{x \in X} D(x)^2}$. Επαναλαμβάνουμε τη διαδικασία μέχρι να επιλέξουμε k κέντρα. Τα επιλεγμένα κέντρα χρησιμοποιούνται σαν αρχικά

κέντρα για να τρέξουμε τον αλγόριθμο k-means.



Σχήμα 4.27: k-means++. [17]

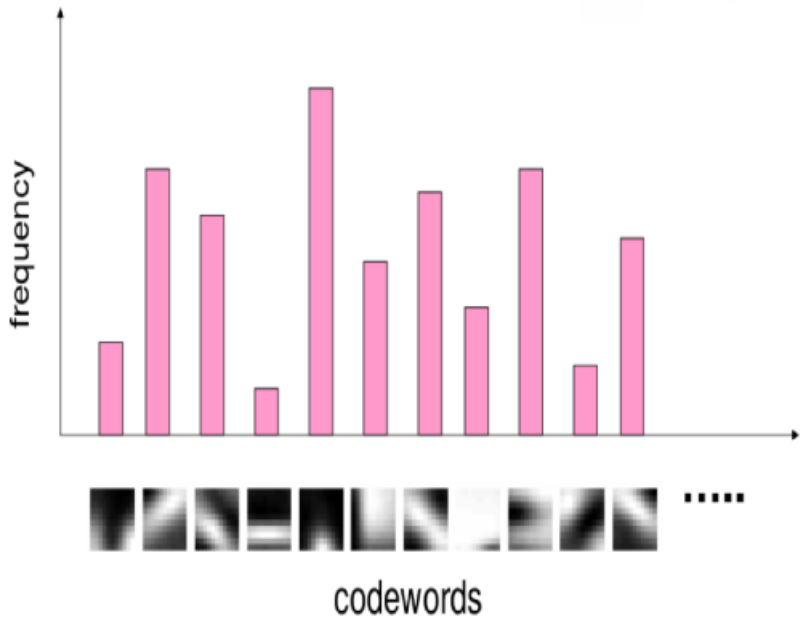
Canopy είναι μία τεχνική soft clustering. Ο αλγόριθμος δουλεύει ως εξής: Επιλέγουμε δύο thresholds, T_1 (loose απόσταση) και T_2 (tight απόσταση) με $T_1 > T_2$ (T_1 και T_2 μπορούν να επιλεχθούν χρησιμοποιώντας grid- search ή cross-validation). Επιλέγουμε ένα δείγμα x από τα δεδομένα μας X, το αφαιρούμε από το X και το θέτουμε σαν canopy. Έπειτα, για κάθε σημείο που έχει απομείνει στο X, αν η απόστασή του με το τελευταίο canopy είναι μικρότερη από T_2, αντιστοιχούμε αυτό το σημείο με αυτό το canopy και αφαιρούμε το σημείο από το X. Αν η απόσταση μεταξύ του σημείου και του τρέχον canopy είναι μεγαλύτερη από T_2 αλλά μικρότερη από T_1, αντιστοιχούμε αυτό το σημείο με αυτό το canopy αλλά δεν αφαιρούμε το σημείο από το X. Από τα σημεία που έχουν παραμείνει στο X, επιλέγουμε ένα άλλο σημείο στην τύχη σαν το τρέχων canopy και επαναλαμβάνουμε τη διαδικασία μέχρις ώτου να μην έχει μείνει κανένα σημείο στο X. Τα κέντρα από τα canopies χρησιμοποιούνται σαν αρχικά κέντρα για την εκτέλεση του αλγορίθμου k-means.



Σχήμα 4.28: Αλγόριθμος Canopy. [18]

Το τρίτο στάδιο της μεθόδου BoW είναι η αναπαράσταση κάθε descriptor από όλες τις εικόνες, χρησιμοποιώντας το οπτικό λεξικό. Κάθε descriptor αντιστοιχίζεται στο cluster (από τον k-means) που ανήκει. Με άλλα λόγια, το τρίτο στάδιο περιλαμβάνει την αντιστοίχιση μεταξύ των descriptors κάθε εικόνας και της κοντινότερης λέξης στο codebook.

Το τέταρτο στάδιο είναι η αναπαράσταση κάθε εικόνας σαν είναι k-dimension ιστόγραμμα το οποίο μας δείχνει πόσοι από τους N-dimension descriptors της εικόνας έχουν αντιστοιχιθεί στη $k-th$ λέξη του λεξικού. Επειδή οι κάθε εικόνα έχει διαφορετικό πλήθος descriptors, κάποια ιστογράμματα ενδέχεται να έχουν ιδιαίτερα μεγάλες τιμές σε σύγκριση με άλλα. Το γεγονός αυτό μπορεί να προκαλέσει διάφορα προβλήματα όταν εφαρμόζουμε επιπλέων τεχνικές machine learning. Για το λόγο αυτό συνήθως κανονικοποιούμε τα ιστογράμματα (το άθροισμα όλων των στοιχείων κάθε descriptor να ισούται με μονάδα).



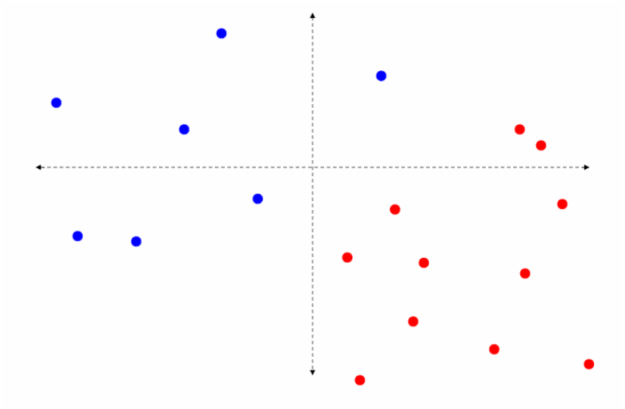
Σχήμα 4.29: Αναπαράσταση εικόνας χρησιμοποιώντας k-διάστατο ιστόγραμμα.

4.3.3 Support Vector Machine (SVM)

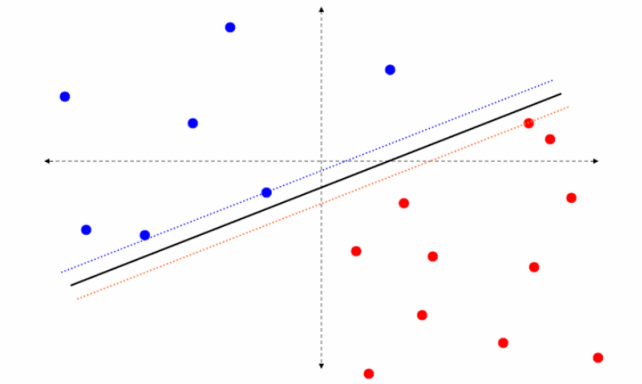
Μετά τη χρήση της τεχνικής BoW, είναι σύνηθες να χρησιμοποιούνται τα ιστογράμματα που προκύπτουν για την εκπαίδευση ενός SVM για κατηγοριοποίηση εικόνων. Παρόλο που τα SVMs χρησιμοποιούνται για να διαχωρίσουν τα δεδομένα σε δύο ομάδες, έχουν βρεθεί επεκτάσεις τους ώστε να μπορούν να χρησιμοποιηθούν για κατηγοριοποίηση multi-label δεδομένων. Θα αναφερθούμε εκτενέστερα σε αυτό αργότερα [64], [65]. Ο λόγος που χρησιμοποιήσαμε SVM στο σύστημά μας ήταν αφενός,για να ελέγξουμε διαφόρων ειδών κανονικοποιήσεις στα χαρακτηριστικά μας, και αφετέρου για να αποκτήσουμε μια εκτίμηση των αποτελεσμάτων που αναμέναμε από το τελικό μας σύστημα, διεξάγοντας ένα παράλληλο πείραμα.

Ας υποθέσουμε ότι έχουμε δύο κατηγορίες δισδιάστατων δεδομένων (μπλε και κόκκινα) και τα σχεδιάζουμε στο επίπεδο.

Στόχος μας είναι να βρούμε ένα υπέρ-επίπεδο (μία ευθεία στο παράδειγμά μας) η οποία διαχωρίζει τις δύο ομάδες. Γίνεται εύκολα αντιληπτό ότι στο παράδειγμά μας υπάρχουν αρκετές τέτοιες ευθείες. Μία από αυτές είναι η ακόλουθη:



Σχήμα 4.30: 2-διάστατα δεδομένα για ταξινόμηση με χρήση SVM.

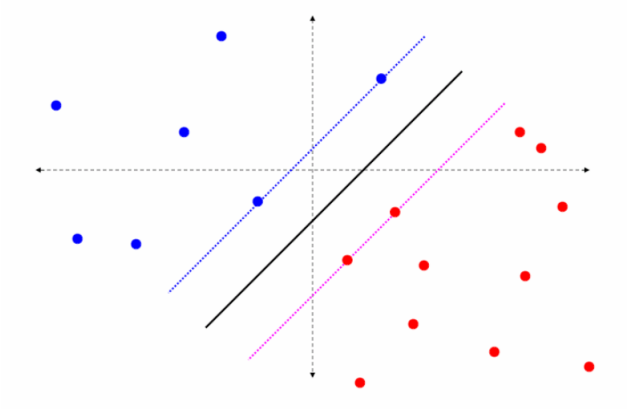


Σχήμα 4.31: Μία διαχωριστική γραμμή μεταξύ των κλάσεων.

Σε αυτό το σημείο, κάποιος θα μπορούσε να αναρωτηθεί: ΄Είναι αυτή η καλύτερη δυνατή ευθεία διαχωρισμού; Πως ορίζεται ο ΄όρος΄ καλύτερη;'

Θα ασχοληθούμε αρχικά με τη δεύτερη ερώτηση. Η καλύτερη διαχωριστική γραμμή είναι εκείνη που αφήνει το μεγαλύτερο κενό (margin) ανάμεσα στις δύο ομάδες. Με άλλα λόγια, θέλουμε να βρούμε την ευθεία, που η απόστασή της από το κοντινότερο σημείο κάθε ομάδας προς αυτήν, είναι η μέγιστη δυνατή. Επομένως η απάντηση στην πρώτη ερώτηση είναι ΄όχι΄ όπως φαίνεται και από το παρακάτω σχήμα όπου έχουμε σχεδιάσει την καλύτερη δυνατή διαχωριστική γραμμή.

Τώρα θα δώσουμε μια πιο μαθηματική περιγραφή για τη διαδικασία εύρεσης της καλύτερης διαχωριστικής γραμμής.



Σχήμα 4.32: Η καλύτερη διαχωριστική γραμμή μεταξύ των κλάσεων.

Αρχικά μετατρέπουμε τις ετικέτες ώστε οι τιμές τους να είναι νούμερα (1 και -1) αντί για κατηγορίες (κόκκινο και μπλε). Υποθέτουμε ότι έχουμε n πλήθος δεδομένων x_i (τα οποία είναι διανύσματα διαστάσεων k) με ετικέτες l_i, με $1 < l < n$. Το υπέρ-επίπεδο διαχωρισμού μπορεί να γραφτεί στη μορφή:

$$w^T * x - b = 0$$

όπου x είναι το διάνυσμα εισόδου, w είναι ο συντελεστής βάρους και b είναι ένα offset. Για κάθε σημείο δεδομένων έχουμε:

$$w^T * x_i - b \geq 1, \qquad l_i = 1$$

ή

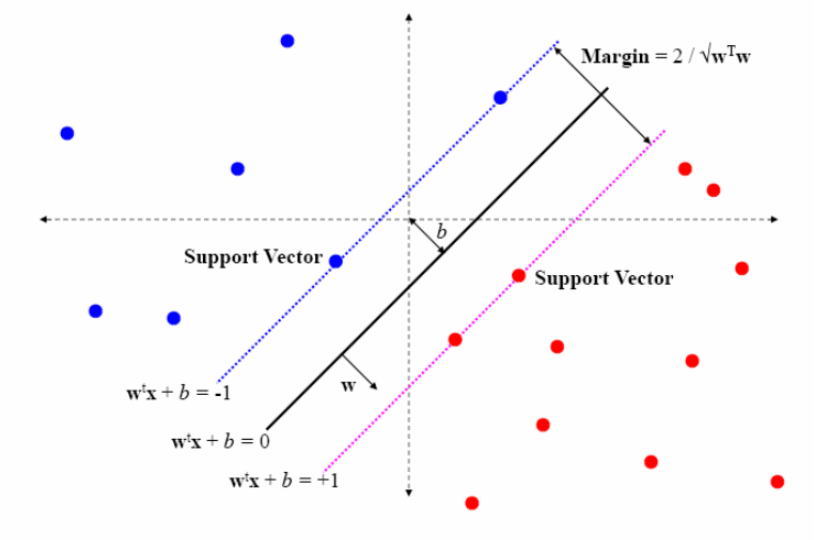
$$w^T * x_i - b \leq 1, \qquad l_i = -1$$

Οι πέριορισμοί αυτοί μπορούν να γραφούν στη μορφή

$$l_i * (w^T * x_i - b) \geq 1, \qquad 1 \leq i \leq n$$

Όπως φαίνεται, το margin ισούται με $\frac{2}{||w||}$. Για να μεγιστοποιήσουμε το margin, πρέπει να ελαχιστοποιήσουμε την ποσότητα $||w||$. Επομένως πρέπει να λύσουμε το ακόλουθο πρόβλημα βελτιστοποίησης: ελαχιστοποίηση του $||w||$, με χρήση της συνάρτησης $\Phi(w) = \frac{1}{2} w^T w$ σα συνάρτηση κόστους, με τον περιορισμό ότι $l_i * (w^T * x_i - b) \geq 1$ για όλα τα i με $1 \leq i \leq n$. Το SVM που προκύπτει μέσω αυτής της διαδικασίας λέγεται hard-margin SVM.

Όλα αυτά συνοψίζονται στην επόμενη εικόνα.



Σχήμα 4.33: Οπτική εξήγηση της ορολογίας που χρησιμοποιείται στα SVM. [19]

Καθ΄ όλη τη διάρκεια της ανάλυσής μας, υποθέσαμε ΄μυστικά΄ ότι τα σημεία δεδομένων μας μπορούν να διαχωριστούν σε δύο διακεκριμένες κλάσεις, υπό την έννοια ότι καμία από τις δύο ομάδες δεν περιλαμβάνει σημεία της άλλης ομάδας. Σε αρκετές περιπτώσεις όμως, αυτή η υπόθεση δεν ισχύει. Για την επίλυση αυτού του προβλήματος, τροποποιούμε τον αλγόριθμο εκπαίδευσης του SVM προκειμένου να είναι σε θέση να αντιμετωπίσει τέτοιου είδους καταστάσεις. Οι καινούριοι περιορισμοί σε αυτή την περίπτωση είναι:

$$l_i * (w^T * x_i - b) \geq 1 - \xi_i, \qquad 1 \leq i \leq n$$

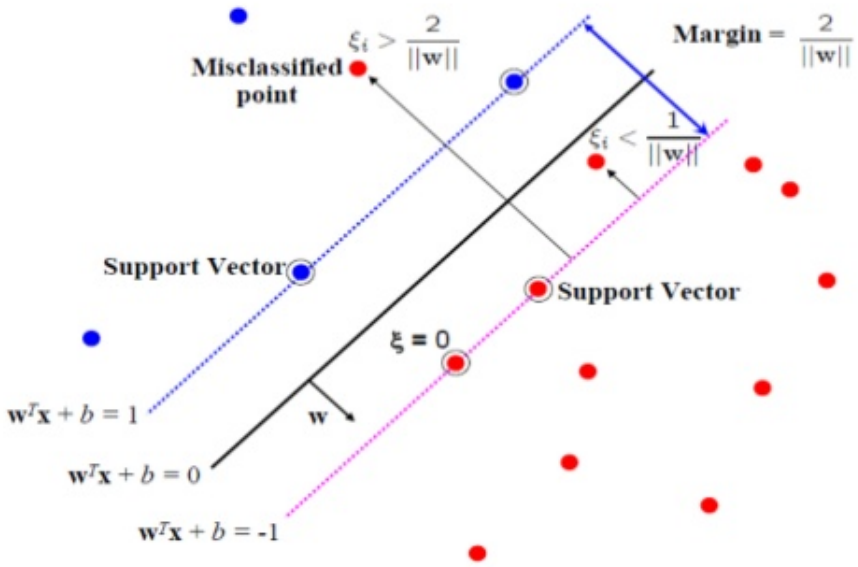
x_i είναι οι μεταβλητές χαλάρωσης (slack) και υπολογίζουν πόσο μέσα στη λάθος περιοχή είναι το σημείο i_{th} των δεδομένων. Οι xi_i μπορούν να πάρουν μόνο μη αρνητικές τιμές.

Υπάρχουν 3 περιπτώσεις

1. $\xi_i = 0$. Πρακτικά η προηγούμενη περίπτωση. Ένα σημείο με $\xi_i = 0$ είναι ταξινομημένο σωστά διατηρώντας το επιθυμητό περιθώριο.

2. $0 < \xi_i < 1$. Ένα σημείο με $0 < \xi_i < 1$ είναι ταξινομημένο σωστά χωρίς να έχει διατηρηθεί το επιθυμητό περιθώριο.

3. $1 < \xi_i$. Ένα σημείο με $1 < \xi_i$ έχει ταξινομηθεί λάθος.



Σχήμα 4.34: Παράδειγμα σημείων με διαφορετικές ξ_i τιμές.

Επίσης, χρειάζεται να προσαρμόσουμε και τη συνάρτηση κόστους στις αλλαγές που κάναμε. Η καινούρια συνάρτηση κόστους που πρέπει να ελαχιστοποιηθεί είναι:

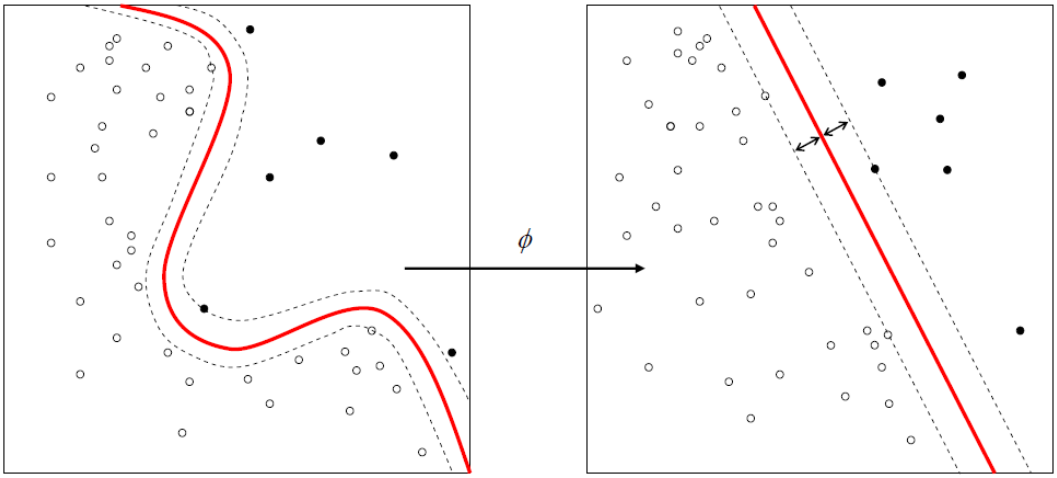
$$\Phi(w, \xi) = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i$$

Η παράμετρος C καθορίζεται από το χρήστη με χρήση cross-validation ή grid-search. Όσο η τιμή της παραμέτρου C αυξάνεται, τόσο το SVM τείνει να γίνεται hard margin και το αντίθετο.

Μία άλλη εξήγηση είναι ότι όσο αυξάνεται η τιμή του C, τόσο πιο σίγουροι είμαστε ότι τα δεδομένα μας είναι καλά. Όταν έχουμε υπόνοια ότι θα δεδομένα μας εμπεριέχουν αρκετό θόρυβο, συνήθως μειώνουμε τις τιμές της παραμέτρου C.

Μία ακόμα υπόθεση που έχουμε κάνει είναι ότι οι ομάδες των δεδομένων μας είναι γραμμικά διαχωρίσιμες. Ωστόσο, οι ομάδες μπορεί να είναι διαχωρίσιμες αλλά όχι με γραμμικό τρόπο. Για την επίλυση αυτού του θέματος, εφαρμόζουμε την τεχνική του πυρήνα (kernel). Μετασχηματίζουμε τα σημεία δεδομένων πολλαπλασιάζοντάς τα με μία συνάρτηση πυρήνα K, σε ένα χώρο μεγαλύτερης διάστασης όπου στον νέο χώρο οι κλάσεις είναι γραμμικά διαχωρίσιμες, και εφαρμόζουμε εκεί τις ίδιες τεχνικές που περιγράψαμε προηγουμένως.

Υπάρχουν αρκετές συναρτήσεις που μπορούν να χρησιμοποιηθούν σαν πυρήνες.



Σχήμα 4.35: Μετασχηματισμός σημείοων χρησιμοποιώντας την τεχνική του πυρήνα. [20]

Θα αναφέρουμε τις πιο συνηθισμένες απ΄ αυτές.

- Γραμμική συνάρτηση:
$$K(x, y) = x^T y$$
- Πολυωνυμική συνάρτηση:
$$K(x, y) = (x^T y + c)^n$$
- Gaussian radial basis συνάρτηση:
$$K(x, y) = e^{-\gamma||x-y||^2} \qquad \gamma > 0$$

Ένα τελευταίο πράγμα που θέλουμε να αναφέρουμε είναι πως μπορεί κανείς να χειριστεί multi-label προβλήματα ταξινόμησης [66]. Οι δημοφιλείς προσεγγίσεις είναι δύο: Η μία κλάση εναντίον όλων one versus all (OvA) και η μια κλάση εναντίον μιας άλλης one versus one (OvO) προσέγγιση.

Στην OvA προσέγγιση, αρχικά εκπαιδεύουμε k SVMs (k είναι το πλήθος των διαφορετικών ετικέτων του προβλήματος). Για κάθε i ($1 \leq i \leq k$) υποθέτουμε ότι όλα τα σημεία με ετικέτα i ανήκουν σε μία κλάση ενώ όλα τα υπόλοιπα σημεία ανήκουν στην άλλη. Με αυτόν τον τρόπο μετασχηματίσαμε multi- label το πρόβλημα σε k κλασσικά προβλήματα. Το επόμενο στάδιο, είναι να ορίσουμε μία αριθμητική ποσότητα που θα περιγράφει το πόσο σίγουροι είμαστε για το αποτέλεσμα της ταξινόμησης. Μία

επιλογή είναι να χρησιμοποιήσουμε την απόσταση του σημείου από το διαχωριστικό υπέρ-επίπεδο. Για να ταξινομήσουμε ένα καινούριο σημείο y, εφαρμόζουμε σε όλα τα SVMs το y και κατηγοριοποιούμε το y στην κλάση που αντιπροσωπεύει το SVM που είχε το μεγαλύτερο confidence score.

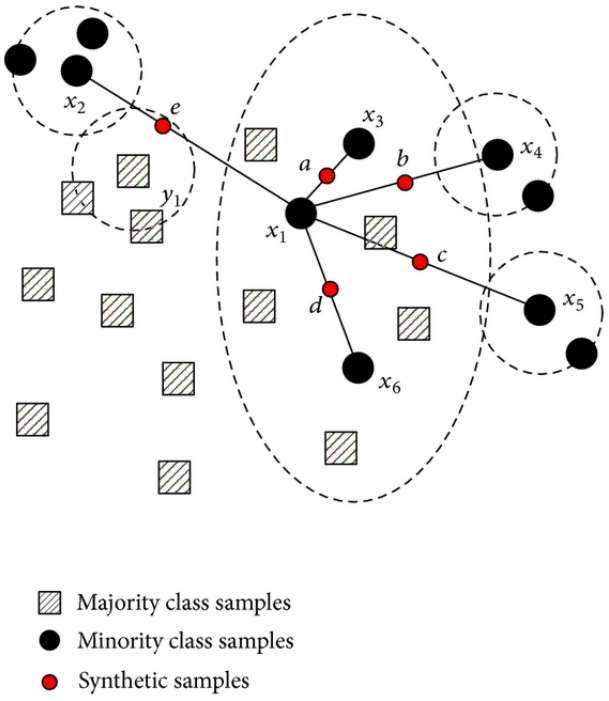
Στην OvO προσέγγιση, εκπαιδεύουμε ένα SVM για κάθε πιθανό ζεύγος από k λαβελς. Το συνολικό πλήθος των SVMs που θα χρειαστεί να εκπαιδεύσουμε είναι $k(k-1)/2$. Όταν έχουμε να ταξινομήσουμε ένα καινούριο σημείο y, εφαρμόζουμε όλα τα SVMs στο y. Κάθε SVM προτείνει (ψηφίζει) μία ετικέτα για το y. Το y κατηγοριοποιείται στην κλάση της ετικέτας με τις περισσότερες ψήφους.

Βέβαια, και οι δύο προσεγγίσεις έχουν κάποια μειονεκτήματα. Στη τεχνική OvA, ένα σοβαρό ζήτημα είναι ότι όλα τα δεδομένα εκπαιδεύονται σε imbalanced σετ δεδομένων, ακόμα και αν η κατανομή των δεδομένων σε κάθε κλάση είναι ισορροπημένη, καθώς τα δεδομένα που ανήκουν σε μία κλάση είναι γενικά πολύ λιγότερα από τα δεδομένα όλων των υπολοίπων κλάσεων αθροιστικά. Επίσης, η κλίμακα του confidence score μπορεί να διαφέρει ανάμεσα στα SVMs. Στην τεχνική OvO, ένα πρόβλημα που μπορεί να προκύψει, είναι ότι δύο (ή και περισσότερες) ετικέτες μπορεί να συγκεντρώσουν τον ίδιο αριθμό ψήφων.

Στην προσέγγισή μας, επιλέξαμε να ακολουθήσουμε την τεχνική OvA. Το μεγάλο μειονέκτημα αυτής της μεθόδου είναι όπως αναφέραμε η εκπαίδευση σε ανισόρροπο πλήθος δεδομένων κάθε κλάσης. Για την επίλυση αυτού του προβλήματος εφαρμόσαμε μια τεχνική δημιουργίας συνθετικών δειγμάτων την οποία περιγράφουμε παρακάτω.

4.3.4 Synthetic Minority Over-sampling Technique (SMOTE)

Η SMOTE είναι μία μέθοδος για τη διαχείριση imbalanced σετ δεδομένων [67]. Η τεχνική αυτή μας φάνηκε ιδιαίτερα χρήσιμη καθώς μας έλυσε το πρόβλημα που είχαμε κατά το στάδιο της εκπαίδευσης των SVMs. Αυτή η μέθοδος δημιουργεί καινούρια, συνθετικά δείγματα για την κλάση με τα λίγα δεδομένα. Ο αλγόριθμος δημιουργεί δείγματα μέχρις ότου προστεθεί ένα συγκεκριμένο πλήθος δειγμάτων. Για κάθε ένα από τα αρχικά σημεία του συνόλου δεδομένων της κλάσης μειονότητας, ο αλγόριθμος βρίσκει τους k κοντινότερους γείτονες και προσθέτει ένα καινούριο δείγμα τυχαία ανάμεσά τους. Η default τιμή της παραμέτρου k είναι 5, αλλά ο χρήστης μπορεί να την αλλάξει ανάλογα με το πλήθος των δειγμάτων που θέλει να δημιουργήσει. Το κύριο πλεονέκτημα αυτής της μεθόδου είναι ότι δημιουργεί καινούρια δείγματα για την κλάση μειονότητας αντί να πολλαπλασιάζει τα υπάρχοντα, το οποίο έχει σαν αποτέλεσμα την επέκταση της περιοχής της κλάσης.



Σχήμα 4.36: Τεχνιτά δείγματα με χρήση της μεθόδου SMOTE. [21]

4.3.5 Συνδυάζοντας την ακουστική πληροφορία και την πληροφορία από τα MRI

Για να συνδυάσουμε την ακουστική πληροφορία με την πληροφορία που εξάγεται από τα MRIs χρησιμοποιήσαμε τη μέθοδο Canonical Correlation Analysis (CCA) η οποία είναι μία σύνθετη μέθοδος της στατιστικής ανάλυσης που χρησιμοποιείται για τη μείωση της διάστασης των δεδομένων. Μία συνοπτική περιγραφή της μεθόδου είναι η ακόλουθη. Ας υποθέσουμε ότι έχουμε δύο σετ μεταβλητών X, Y, με

$$X = [X_1, X_2, ..., X_p]^T, Y = [Y_1, Y_2, ..., Y_q]^T$$

και $p \leq q$.

Στόχος μας είναι να βρούμε ένα γραμμικό συνδυασμό αυτών των μεταβλητών έτσι ώστε οι προβολές των μεταβλητών στον καινούριο χώρο να συσχετίζονται κατά το μέγιστο δυνατό. Ονομάζουμε U το γραμμικό συνδυασμό των μεταβλητών του συνόλου X και V το γραμμικό συνδυασμό των μεταβλητών του συνόλου Y.

Επομένως,

$$
\begin{aligned}
U_1 &= \alpha_{11}X_1 + \alpha_{12}X_2 + ... + \alpha_{1p}X_p \\
U_2 &= \alpha_{21}X_1 + \alpha_{22}X_2 + ... + \alpha_{2p}X_p \\
&\vdots \\
U_p &= \alpha_{p1}X_1 + \alpha_{p2}X_2 + ... + \alpha_{pp}X_p
\end{aligned}
$$

$$
\begin{aligned}
V_1 &= \beta_{11}Y_1 + \beta 12Y_2 + ... + \beta_{1q}Y_q \\
V_2 &= \beta_{21}Y_1 + \beta 22Y_2 + ... + \beta_{2q}Y_q \\
&\vdots \\
V_p &= \beta_{p1}Y_1 + \beta p2Y_2 + ... + \beta_{pq}Y_q
\end{aligned}
$$

Ορίζουμε

$$(U_i, V_i)$$

ως το i_{th} canonical variate ζεύγος (p canonical variate ζεύγη στο πλήθος). Θέλουμε να υπολογίσουμε τους συντελεστές α, β οι οποίοι μεγιστοποιούν τη συσχέτιση μεταξύ του κάθε ζευγαριού.

Η canonical συσχέτιση μεταξύ του i_{th} ζευγαριού ορίζεται ως

$$R_i = \frac{cov(U_i, V_i)}{\sqrt{var(U_i) * var(V_i)}}$$

όπου η διακύμανση του U_i είναι

$$var(U_i) = \sum_{k=1}^{p}\sum_{l=1}^{p}(\alpha_{ik} * \alpha_{il} * cov(X_k, X_l))$$

και η διακύμανση του V_i είναι

$$var(V_i) = \sum_{k=1}^{p}\sum_{l=1}^{q}(\beta_{ik} * \beta_{il} * cov(Y_k, Y_l))$$

Η συνδιακύμανση του κάθε ζεύγους υπολογίζεται ως

$$cov(U_i, V_i) = \sum_{k=1}^{p}\sum_{l=1}^{q}(\alpha_{ik} * \beta_{il} * cov(X_k, Y_l))$$

Για τη μεγιστοποίηση της canonical συσχέτισης, εξετάζουμε όλα τα ζεύγη ένα προς ένα. Για το πρώτο ζεύγος (U_1, V_1) θέλουμε να βρούμε τους συντελεστές $\alpha_{11}, \alpha_{12}, ..., \alpha_{1p}$ και $\beta_{11}, \beta_{12}, ..., \beta_{1q}$ οι οποίοι μεγιστοποιούν την canonical συσχέτιση R_1 του πρώτου ζεύγους. Για να υπάρχει μοναδική λύση στο πρόβλημα, προσθέτουμε κάποιους περιορισμούς. Οι περιορισμοί είναι ότι η διακύμανση των μεταβλητών του πρώτου ζεύγους πρέπει να ισούται με μονάδα.

$$var(U_1) = 1$$

$$var(V_1) = 1$$

Για το δεύτερο ζεύγος (U_2, V_2), η διαδικασία εύρεσης των παραμέτρων $\alpha_{21}, \alpha_{22}, ..., \alpha_{2p}$ και $\beta_{21}, \beta_{22}, ..., \beta_{2q}$ ώστε να μεγιστοποιούν την canonical συσχέτιση R_2 είναι σχεδόν η ίδια με αυτήν που ακολουθήσαμε στο πρώτο ζεύγος (U_1, V_1). Ωστόσο, πρέπει να προσθέσουμε μερικούς ακόμα περιορισμούς για να εξακολουθήσουμε να έχουμε μοναδική λύση. Οι επιπλέον περιορισμοί είναι ότι οι ομάδες (U_1, U_2), (V_1, V_2), (U_1, V_2) και (U_2, V_1) πρέπει να είναι ασυσχέτιστες.

Οι περιορισμοί για το ζεύγος (U_2, V_2) είναι:

$$var(U_2) = 1$$

$$var(V_2) = 1$$

$$cov(U_1, U_2) = 0$$

$$cov(V_1, V_2) = 0$$

$$cov(U_1, V_2) = 0$$

$$cov(U_2, V_1) = 0$$

Γενικά, οι περιορισμοί για το i_{th} ζεύγος (U_i, V_i) ώστε να υπολογίσουμε τους συντελεστές $\alpha_{i1}, \alpha_{i2}, ..., \alpha_{ip}$ και $\beta_{i1}, \beta_{i2}, ..., \beta_{iq}$ που μεγιστοποιούν τη συσχέτιση R_i είναι:

$$\begin{aligned} cov(U_1, U_i) &= 0 \\ cov(U_2, U_i) &= 0 \\ &\vdots \\ cov(U_{i-1}, U_i) &= 0 \end{aligned}$$

$$
\begin{aligned}
cov(V_1, V_i) &= 0 \\
cov(V_2, V_i) &= 0 \\
&\vdots \\
cov(V_{i-1}, V_i) &= 0
\end{aligned}
$$

$$
\begin{aligned}
cov(U_1, V_i) &= 0 \\
cov(U_2, V_i) &= 0 \\
&\vdots \\
cov(U_{i-1}, V_i) &= 0
\end{aligned}
$$

$$
\begin{aligned}
cov(U_i, V_1) &= 0 \\
cov(U_i, V_2) &= 0 \\
&\vdots \\
cov(U_i, V_{i-1}) &= 0
\end{aligned}
$$

$$
\begin{aligned}
var(U_i) &= 1 \\
var(V_i) &= 1
\end{aligned}
$$

Έχοντας υπολογίζει όλα τα R, κρατάμε τους συντελεστές $\alpha_{i1}, \alpha_{i2}, ..., \alpha_{ip}$ και $\beta_{i1}, \beta_{i2}, ..., \beta_{iq}$ από το i_{th} ζεύγος (U_i, V_i) με το υψηλότερο R_i. Ωστόσο, κάποιος μπορεί να κρατήσει τους συντελεστές από περισσότερα ζεύγη (U,V) αν για παράδειγμα δύο R έχουν αρκετά μεγάλη τιμή.

Σε αυτό το κεφάλαιο παρουσιάσαμε με λεπτομέρεια τους διάφορους αλγορίθμους που θα χρησιμοποιήσουμε για την πραγματοποίηση των πειραμάτων της έρευνάς μας. Επιπλέον, δώσαμε μια κατεύθυνση σχετικά με τη μεθοδολογία που θα ακολουθήσουμε και τις δυσκολίες που προέκυψαν στα πειράματά μας καθώς και τον τρόπο με τον οποίο οι διάφορες μέθοδοι μας βοήθησαν να τα αντιμετωπίσουμε.

Κεφάλαιο 5

Πειράματα

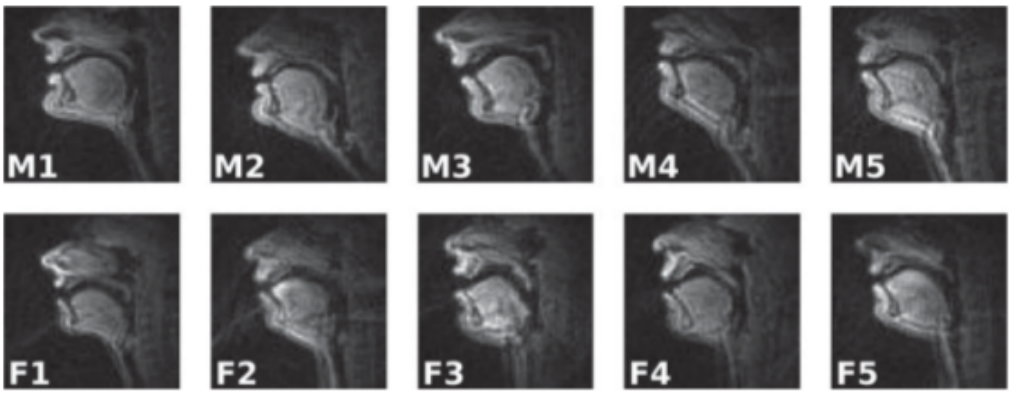
Σε αυτό το κεφάλαιο θα περιγράψουμε τη βάση δεδομένων που χρησιμοποιήσαμε καθώς και τα πειράματα που εκτελέσαμε. Τα πειράματα αποτελούνται από τρία μέρη: Πειράματα που αφορούν τη ακουστική πληροφορία, πειράματα που αφορούν την πληροφορία παραγωγής ομιλίας και πειράματα που αφορούν το συνδυασμό των προαναφερθέντων. Για την εκτέλεση των πειραμάτων χρησιμοποιήθηκαν ποικίλοι τρόποι εκπαίδευσης και αλγόριθμοι οι οποίοι αναφέρονται στη συνέχεια.

5.1 Βάση δεδομένων rtMRI-TIMIT

Η βάση δεδομένων rt-MRI TIMIT αποτελείται από 460 προτάσεις που έχουν επιλεχθεί από το σύνολο των προτάσεων της βάσης δεδομένων TIMIT. Οι προτάσεις έχουν επιλεχθεί έτσι ώστε να περιλαμβάνουν όλα τα φωνήματα που εμφανίζονται στα αμερικανικά Αγγλικά, σε ένα μεγάλο εύρος φωνολογικού περιεχομένου. Συνολικά, η βάση περιλαμβάνει ηχογραφήσεις από 10 ομιλητές. Πέντε από τους ομιλητές είναι άνδρες, ηλικίας $26-33$, ενώ οι υπόλοιποι είναι γυναίκες, ηλικίας $23-46$. Οι άνδρες ομιλητές έχουν το διακριτικό $M1-M5$ ενώ οι γυναίκες $F1-F5$ (Σχήμα 5.1). Οι ομιλητές έχουν επιλεχθεί έτσι ώστε να προέρχονται από διαφορετικές περιοχές, ώστε να συμπεριληφθούν στη βάση όσο περισσότερες διαφορετικές προφορές γίνεται. Η πλήρης λίστα με τα δημογραφικά χαρακτηριστικά των ομιλητών φαίνεται στο 5.2

Η βάση δεδομένων περιλαμβάνει τα audio καθώς και τα MRI recordings των ομιλητών που εκφωνούν τις 460 προτάσεις του TIMIT corpus. Για κάθε ομιλητή, η βάση δεδομένων περιέχει τα αυδιο, τα articulatory και τα audio-articulatory αρχεία των 460 προτάσεων οι οποίες είναι μοιρασμένες σε 92 αρχεία των πέντε προτάσεων το καθένα.

Η διάρκεια των αρχείων είναι περίπου ανάμεσα σε 20 και 30 δευτερόλεπτα. Η συχνότητα δειγματοληψίας του ήχου είναι $20000Hz$ και η συχνότητα δειγματοληψίας του video είναι 23 πλαίσια το δευτερόλεπτο. Η ανάλυση των πλαίσιο του video είναι 68×68 pixels. Στην αρχή κάθε αρχείου ακούγεται ένα χαρακτηριστικό beep πριν



Σχήμα 5.1: Εικόνες rtMRI από τους δέκα ομιλητές της USC-TIMIT database (πάνω σειρά: άνδρες, κάτω σειρά: γυναίκες). [1]

ξεκινήσει να μιλάει ο ομιλητής.

5.2 Προεπεξεργασία δεδομένων

Σε αυτό το στάδιο, ακούσαμε κάθε αρχείο ήχου για να βεβαιωθούμε αν ο εκάστοτε ομιλητής έλεγε αυτό που έπρεπε να πει σύμφωνα με το corpus ή όχι. Αρχεία που ο ομιλητής έκανε σαρδάμ ή που περιείχαν άλλες ανωμαλίες (όπως προφορά άλλων φωνημάτων π.χ. 'eeee' στο ενδιάμεσο του αρχείου) αφαιρέθηκαν χειροκίνητα (όχι μόνο τα audio αρχεία αλλά και τα articulatory και audio-articulatory αρχεία). Τα αρχεία που παρέμειναν και που αφαιρέθηκαν ανά ομιλητή φαίνοντα στον πίνακα 5.2. Όσον αφορά τα αρχεία ήχου, αφαιρέσαμε 149 αρχεία ενώ παραμείνανε 771 από τα 920 αρχεία ήχου που περιείχε αρχικά η βάση. Προφανώς αυτά τα αρχεία διαγράφηκαν (και διατηρήθηκαν) και από τα articulatory και audio- articulatory αρχεία. Συνολικά αφαιρέσαμε $3 * 149 = 447$ αρχεία ενώ κρατήσαμε $3 * 771 = 2313$ αρχεία από το σύνολο των $920 * 3 = 2760$ αρχείων.

Το επόμενο βήμα είναι η δημιουργία δύο alignment αρχείων για κάθε αρχείο ήχου. Το πρώτο από τα δύο περιλαμβάνει alignments των λέξεων(.wrd), ενώ το δεύτερο περιλαμβάνει alignments των φωνημάτων(.phn). Συνολικά $2 * 771 = 1542$ καινούρια αρχεία δημιουργήθηκαν. Και τα δύο είδη αρχείων δημιουργήθηκαν με χρήση του προγράμματος SailAlign [68].

ID	Gender	Race	Age	Birthplace
M1	Male	White	29	Buffalo, NY
M2	Male	White	33	Ann Arbor, MI
M3	Male	White	26	Madison, WI
M4	Male	White	26	St. Louis, MO
M5	Male	White	27	Mammoth, CA
W1	Female	White	23	Commack, NY
W2	Female	White	32	Westfield, IN
W3	Female	White	20	Palos Verdes, CA
W4	Female	White	46	Pittsburgh, PA
W5	Female	White	25	Brawley, CA

Σχήμα 5.2: Δημογραφικά χαρακτηριστικά των συμμετεχόντων.

Ομιλητής	Αφαιρέθηκαν	Παρέμειναν
M1	14	78
M2	11	81
M3	10	82
M4	14	78
M5	30	62
F1	20	72
F2	15	77
F3	18	74
F4	5	87
F5	12	80

Πίνακας 5.1: Αρχεία που διαγράφηκαν και που διατηρήθηκαν για κάθε ομιλητή.

5.3 Ηχητικά πειράματα

5.3.1 Αρχικό στάδιο προεπεξεργασίας

Τώρα είμαστε έτοιμοι να ξεκινήσουμε τα πειράματά μας, τη δημιουργία και μελέτη των συστημάτων αναγνώρισης ομιλίας. Σε αυτό το στάδιο χρησιμοποιούμε μόνο τα αρχεία ήχου για τα πειράματά μας. Υπάρχουν δύο λόγοι για τους οποίους το κάνουμε αυτό. Πρώτον, για να αποκτήσουμε μία αίσθηση/εκτίμηση για το που αναμένουμε να

κυμαίνονται τα αποτελέσματα του τελικού audio-articulatory συστήματος, λύνοντας ένα πιο απλό πρόβλημα. Δεύτερον, για να τεστάρουμε διάφορες παραμέτρους όπως η συχνότητα δειγματοληψίας και να καταλήξουμε ποιες τιμές δίνουν καλύτερα αποτελέσματα. Υποθέτουμε ότι οι τιμές που δίνουν τα καλύτερα αποτελέσματα στα audio συστήματα θα δίνουν επίσης τα καλύτερα αποτελέσματα και στα audio-articulatory συστήματα.

Για τον σκοπό των πειραμάτων μας χρησιμοποιήσαμε το εργαλείο Kaldi. Η συνταγή που δημιουργήσαμε βασίστηκε κατά κύριο λόγο στην TIMIT/s5 συνταγή του Kaldi για το σύνολο των πειραμάτων που εκτελέστηκαν στο Kaldi. Πρακτικά προσαρμόσαμε την υπάρχουσα συνταγή του Kaldi στις ανάγκες του δικού μας προβλήματος. Για την εκτέλεση των πειραμάτων απαιτούνται τα εξής αρχεία:

1. Ένα λεξικό (.dic) στο οποίο κάθε λέξη αντιστοιχίζεται στα αντίστοιχα φωνήματα. Για τα πειράματά μας χρησιμοποιήσαμε το λεξικό cmu.
2. Audio (.wav) αρχεία
3. Word alignment (.wrd) αρχεία
4. Phone alignment (.phn) αρχεία
5. Audio transcript (.txt) αρχεία

Δημιουργήσαμε πέντε ζεύγη ομιλητών για να εφαρμόσουμε 5-fold cross- validation. Κάθε ζεύγος αποτελείται από έναν άνδρα και μία γυναίκα ομιλητή (Πίνακας 5.3.1).

Αριθμός ζεύγους	Άνδρας ομιλητής	Γυναίκα ομιλήτρια
1	M1	F1
2	M2	F2
3	M3	F3
4	M4	F4
5	M5	F5

Πίνακας 5.2: Ζεύγη ομιλητών.

Σε αυτό το σημείο διαιρούμε το σύνολο δεδομένων σε τρία υποσύνολα, το train, το development (dev) και το test set. Ο διαχωρισμός για κάθε ένα από τα 5 folds φαίνεται στον Πίνακα 5.3.1.

Fold	Train set	Dev set	Test set
1	Ζεύγος 1 - 2 - 3	Ζεύγος 4	Ζεύγος 5
2	Ζεύγος 5 - 1 - 2	Ζεύγος 3	Ζεύγος 4
3	Ζεύγος 4 - 5 - 1	Ζεύγος 2	Ζεύγος 3
4	Ζεύγος 3 - 4 - 5	Ζεύγος 1	Ζεύγος 2
5	Ζεύγος 2 - 3 - 4	Ζεύγος 5	Ζεύγος 1

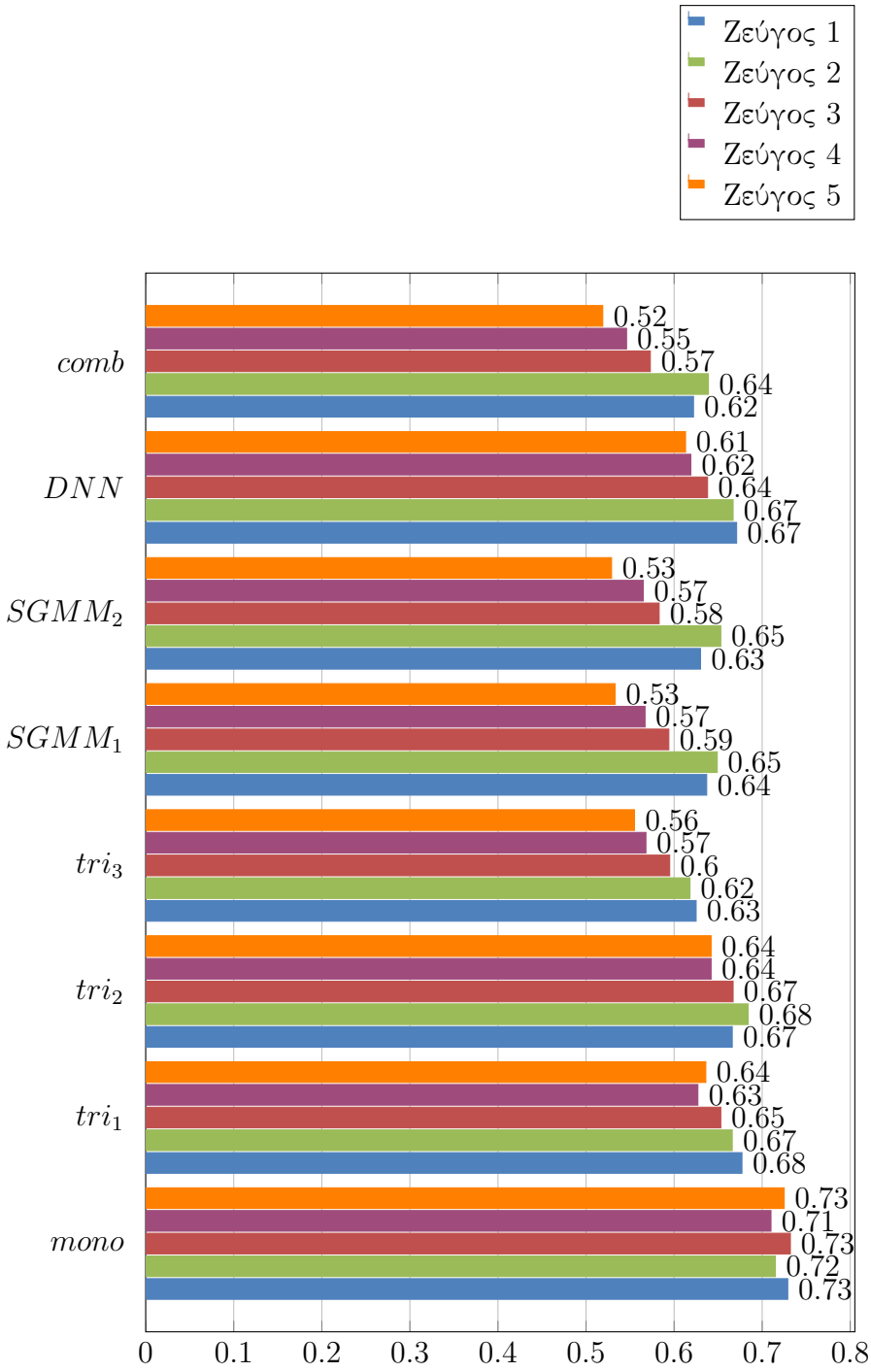
Πίνακας 5.3: Ομάδες ομιλητών.

Για την εκπαίδευση των συστημάτων αναγνώρισης ομιλίας, χρησιμοποιήθηκαν οι πιο κάτω μέθοδοι που προσφέρει το Kaldi, όπως αυτοί περιγράφησαν αναλυτικά στο κεφάλαιο 3:

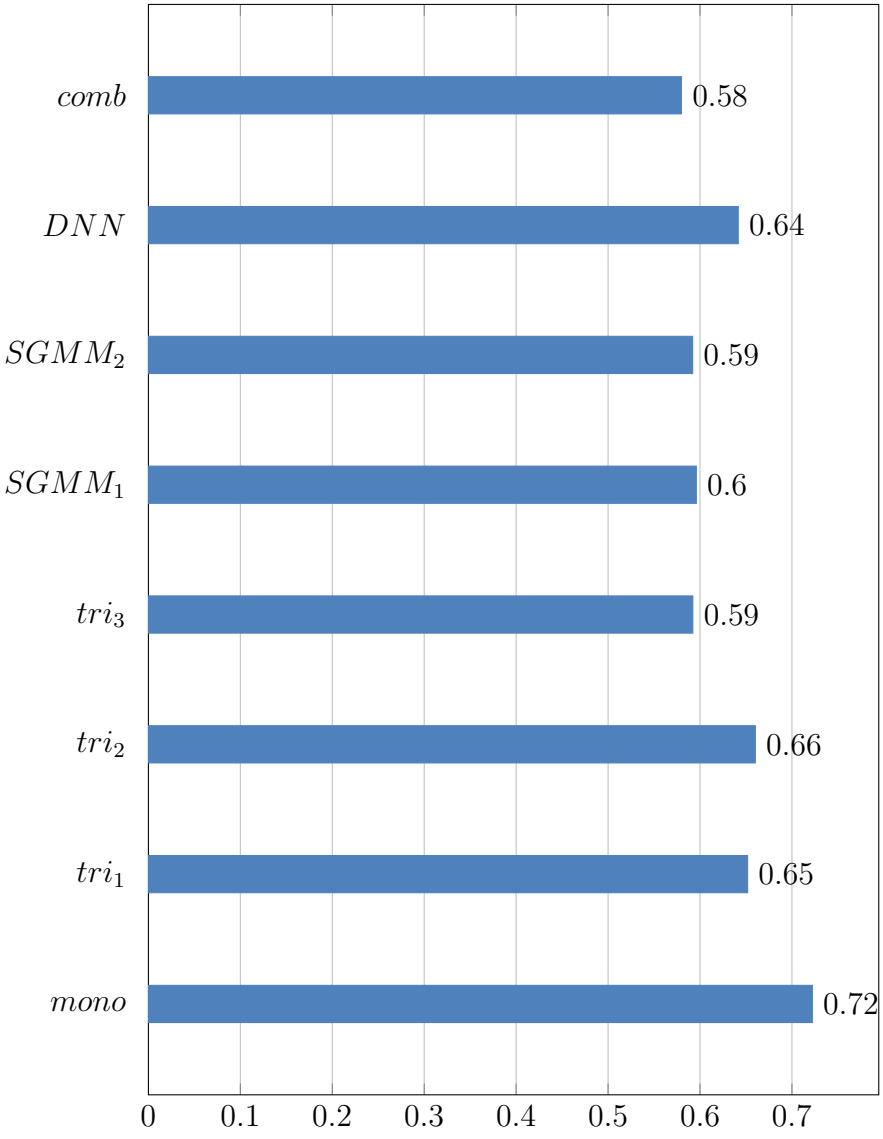
1. *mono*
2. tri_1
3. tri_2
4. tri_3
5. $SGMM_1$
6. $SGMM_2$
7. DNN
8. *comb*

Σαν χαρακτηριστικό εισόδου για για την εκπαίδευση των συστημάτων, χρησιμοποιήσαμε MFCCs χαρακτηριστικά με $10ms$ window sift.

Στα παρακάτω γραφήματα (5.3 , 5.4) παρουσιάζουμε το phone error rate ανά μέθοδο εκπαίδευσης και ανά ζεύγος καθώς και το κανονικοποιημένο phone error rate ανά μέθοδο εκπαίδευσης.



Σχήμα 5.3: Σφάλμα ανά μέθοδο εκπαίδευσης και ζεύγος

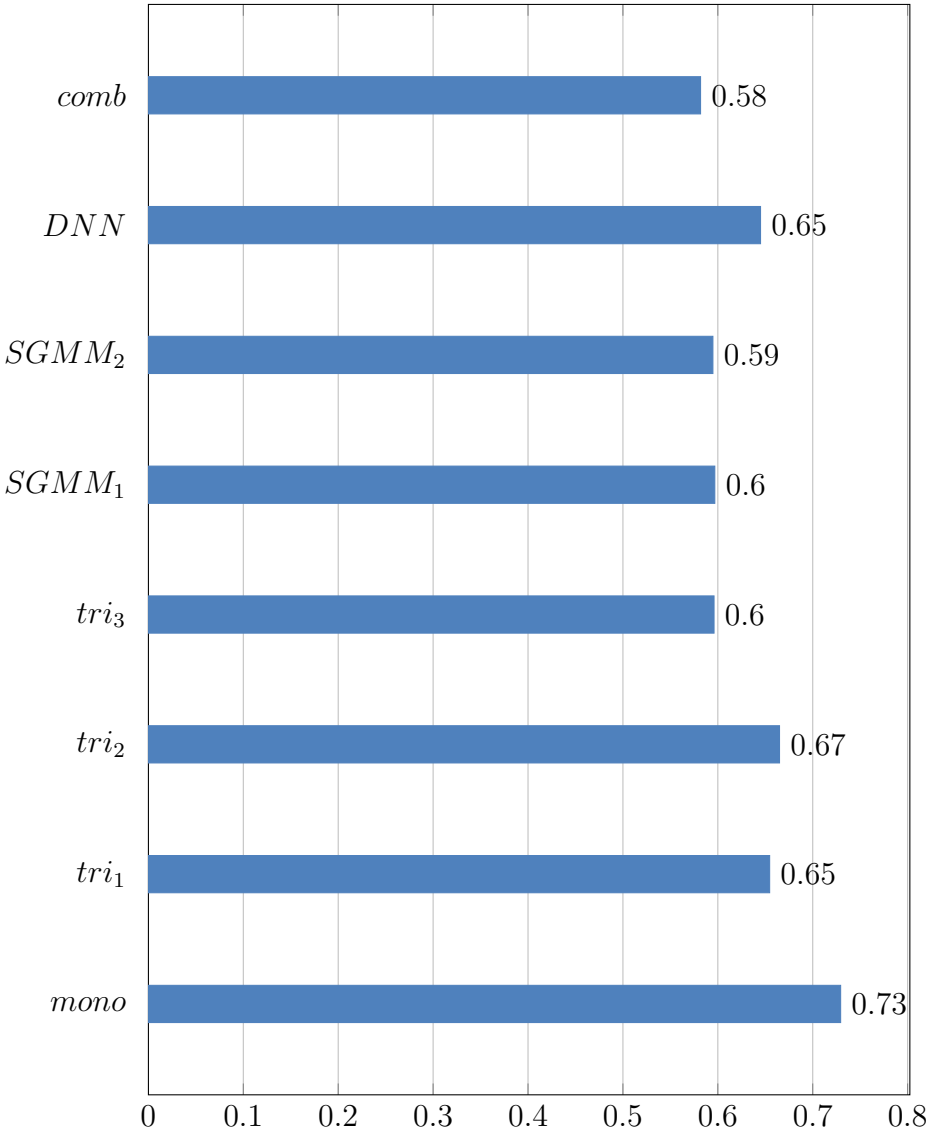


Σχήμα 5.4: Κανονικοποιημένο σφάλμα ανά μέθοδο εκπαίδευσης

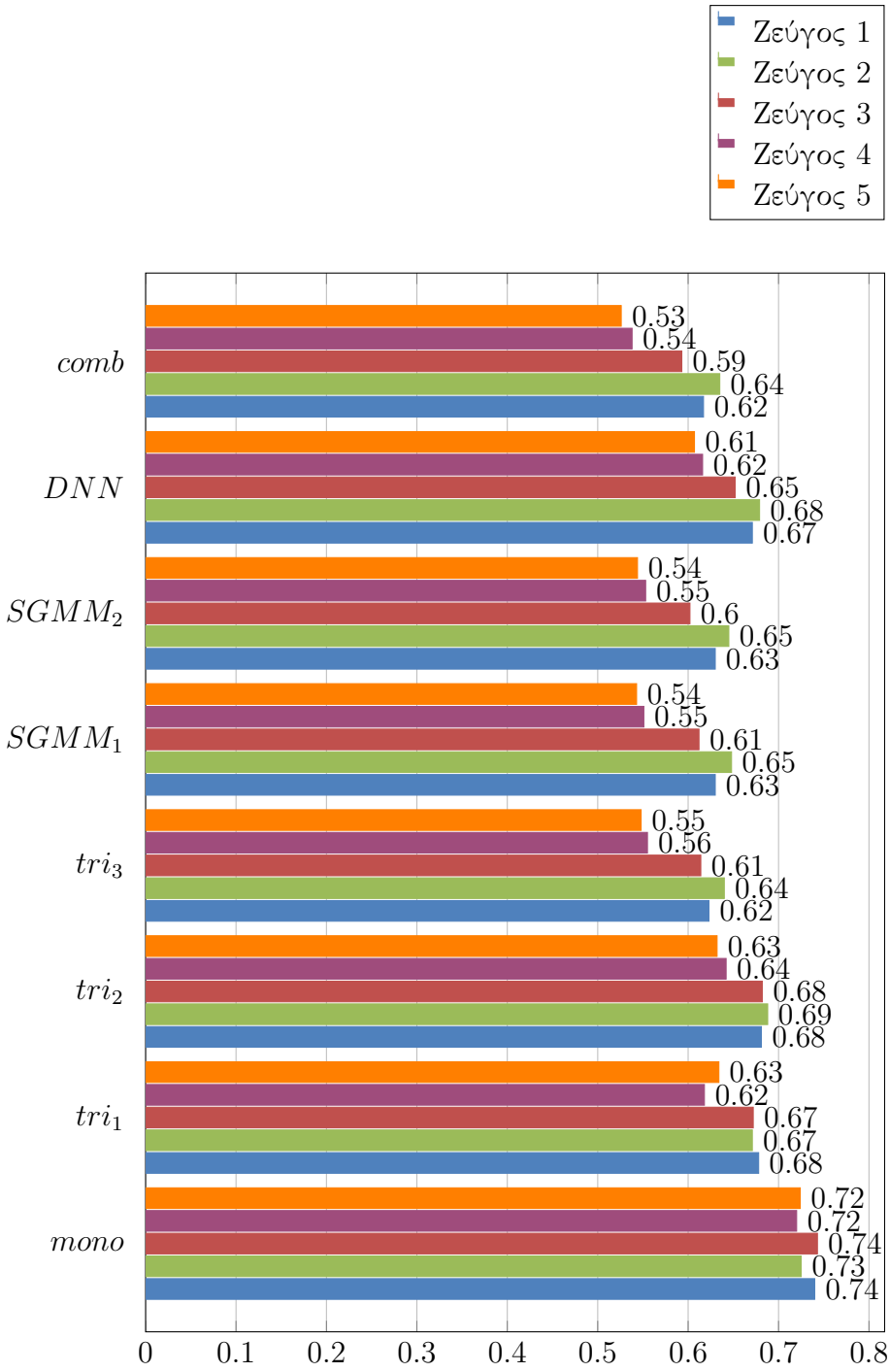
Σε αυτό το σημείο εισάγουμε τρία ανεξάρτητα στάδια προεπεξεργασίας δεδομένων έτσι ώστε να εκτιμήσουμε την αποτελεσματικότητα ορισμένων ιδεών. Σε κάθε ένα από τα στάδια, όλα τα αρχεία που δημιουργήσαμε προηγουμένως με το SailAlign τα ξαναδημιουργούμε από την αρχή. Για την επιπλέον προεπεξεργασία των δεδομένων που έγινε σε αυτά τα στάδια χρησιμοποιήσαμε το εργαλείο sox.

5.3.2 Πρώτο στάδιο

Σε αυτό το στάδιο κάνουμε υποδειγματοληψία των αρχείων ήχου. Η νέα συχνότητα είναι $16000Hz$ (η αρχική συχνότητα δειγματοληψίας είναι $20000Hz$). Στη συνέχεια προχωρήσαμε όπως και στο προηγούμενο στάδιο. Τα αποτελέσματα φαίνονται στα γραφήματα 5.5, 5.6 .



Σχήμα 5.5: Κανονικοποιημένο σφάλμα ανά μέθοδο εκπαίδευσης (Στάδιο 1)



Σχήμα 5.6: Σφάλμα ανά μέθοδο εκπαίδευσης και ζεύγος (Στάδιο 1)

5.3.3 Δεύτερο στάδιο

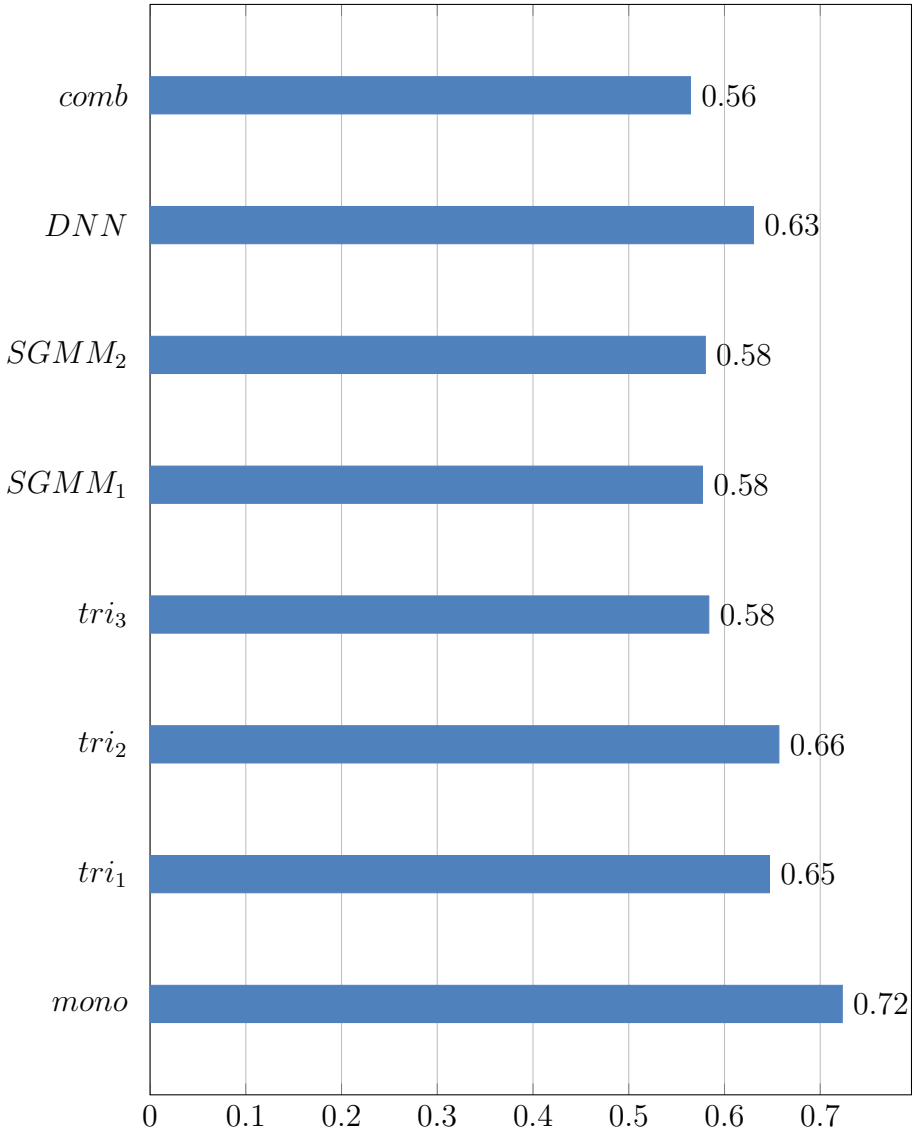
Σε αυτό το στάδιο αφαιρέσαμε το 'beep' από τα αρχικά αρχεία ήχου. Και πάλι συνεχίζουμε όπως και στο προηγούμενο στάδιο. Τα αποτελέσματα φαίνονται στα γραφήματα 5.7, 5.8 .

5.3.4 Τρίτο στάδιο

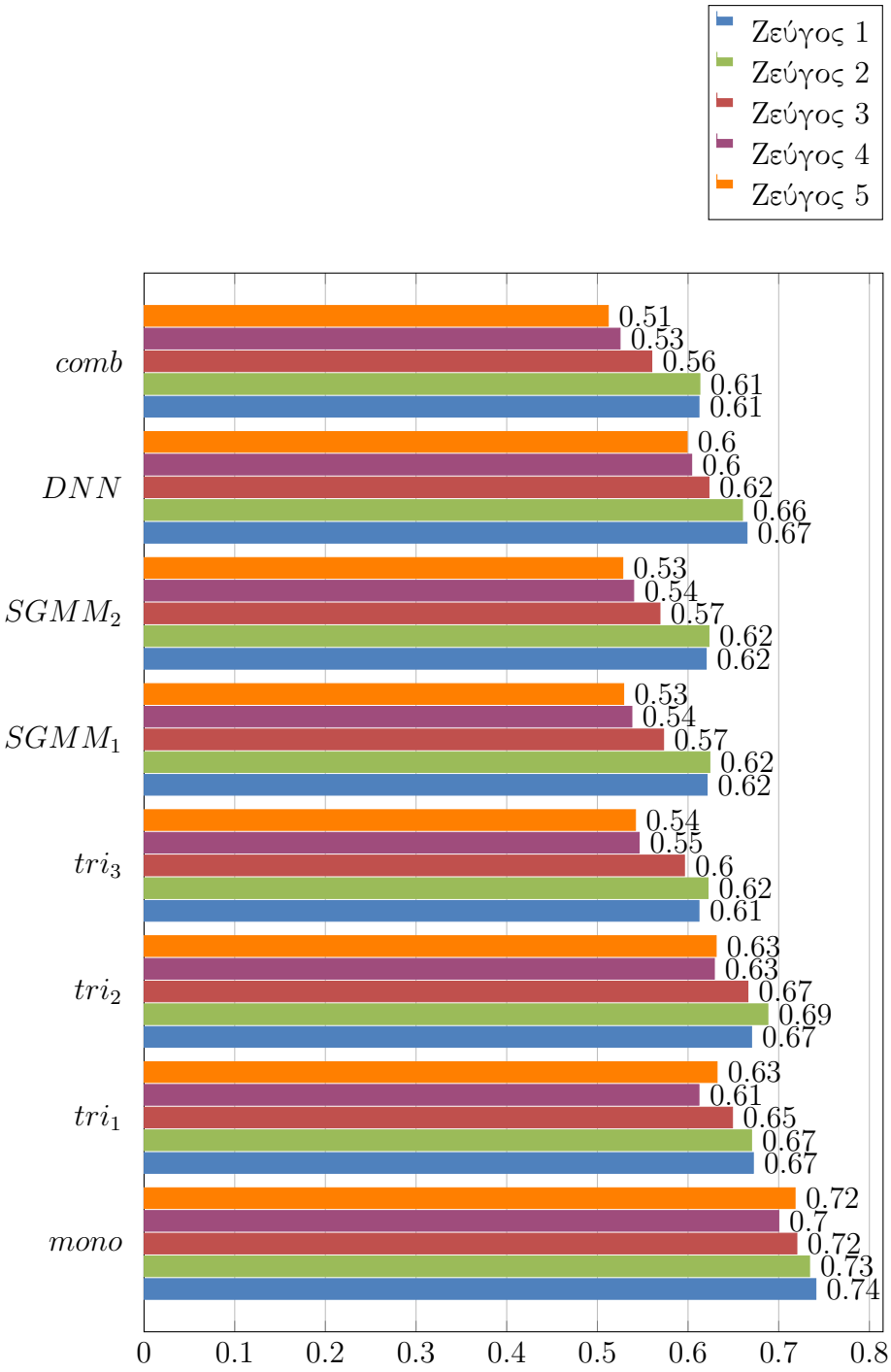
Στο τρίτο στάδιο, συνδυάζουμε τις ιδέες που εφαρμόσαμε στο στάδιο ένα και δύο. Αφαιρούμε το 'beep' από την αρχή των αρχείων και υποδειγματοληπτούμε τα audio αρχεία στα $16000Hz$. Τα αποτελέσματα φαίνονται στα σχήματα 5.9, 5.10.

5.3.5 Συμπεράσματα

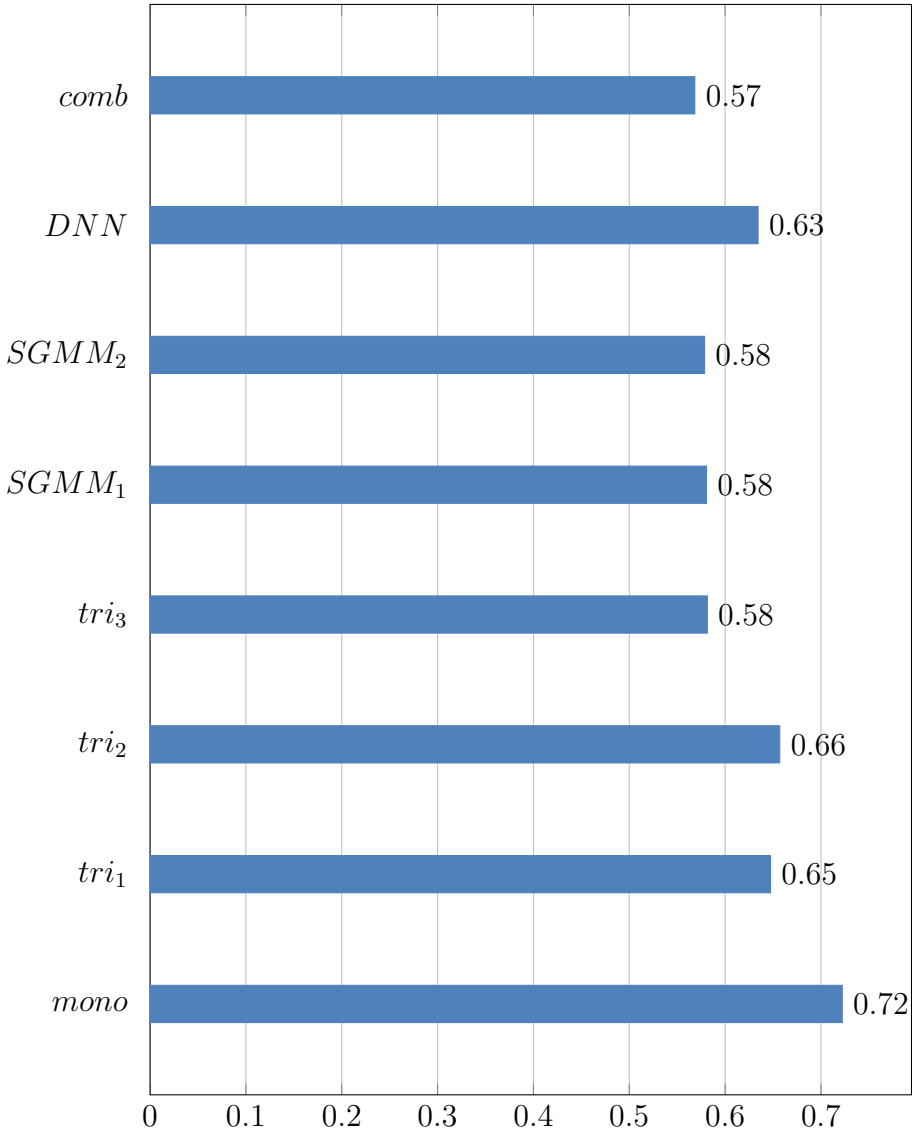
Σύμφωνα με το σχήμα 5.11, τα καλύτερα αποτελέσματα επιτεύχθηκαν στο στάδιο προεπεξεργασίας δύο, ακολουθούμενα από το στάδιο τρία, την αρχική προεπεξεργασία και τέλος το στάδιο ένα. Παρόλο που οι διαφορές των αποτελεσμάτων για τα διάφορα στάδια προεπεξεργασίας δεν ήταν ιδιαίτερα μεγάλες, επιλέξαμε να χρησιμοποιήσουμε την προεπεξεργασία του σταδίου δύο για τη συνέχεια των πειραμάτων μας.



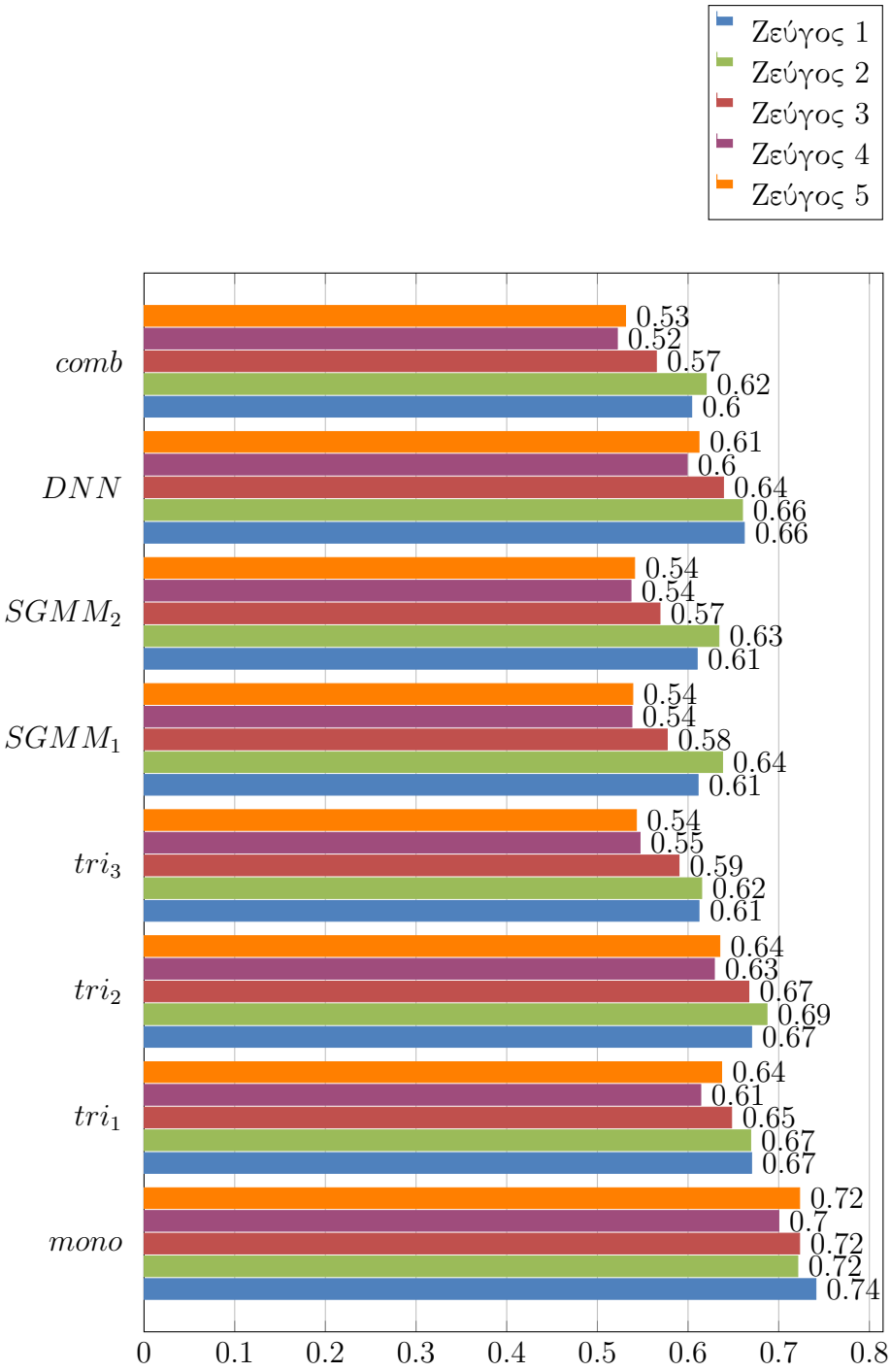
Σχήμα 5.7: Κανονικοποιημένο σφάλμα ανά μέθοδο εκπαίδευσης (Στάδιο 2)



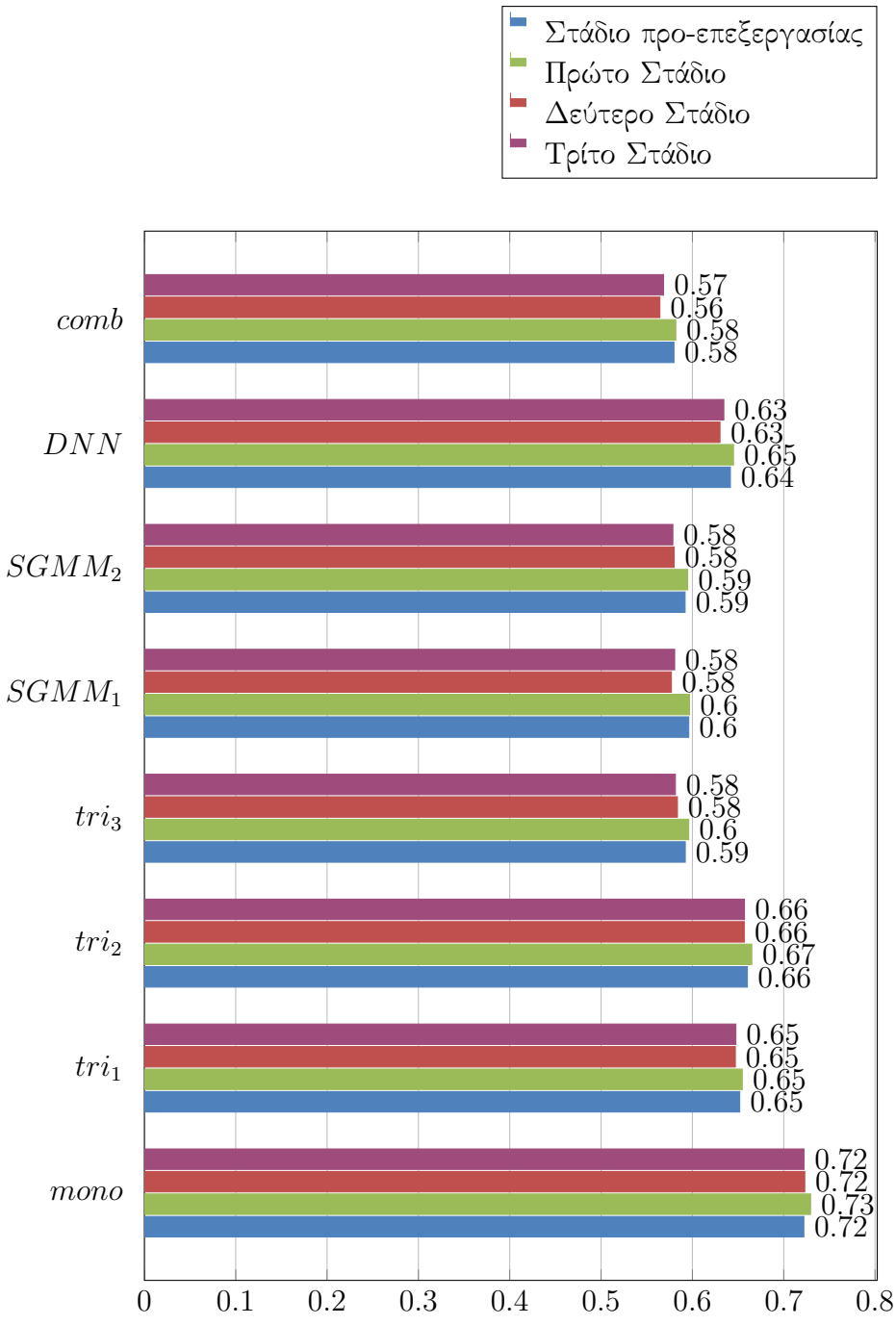
Σχήμα 5.8: Σφάλμα ανά μέθοδο εκπαίδευσης και ζεύγος (Στάδιο 2)



Σχήμα 5.9: Κανονικοποιημένο σφάλμα ανά μέθοδο εκπαίδευσης (Στάδιο 3)



Σχήμα 5.10: Σφάλμα ανά μέθοδο εκπαίδευσης και ζεύγος (Στάδιο 3)



Σχήμα 5.11: Κανονικοποιημένο σφάλμα ανά μέθοδο εκπαίδευση (ολικό)

5.4 Χρησιμοποιώντας την πληροφορία του συστήματος παραγωγής φωνής

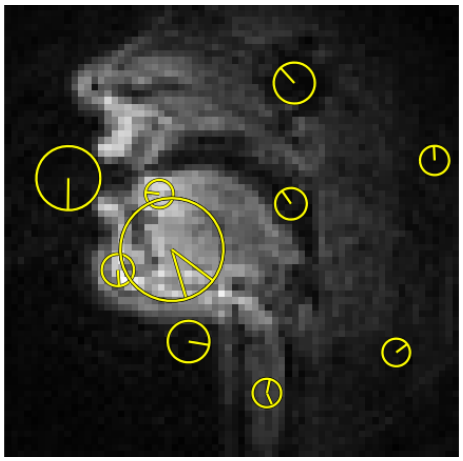
Το πρώτο πράγμα που κάναμε ήταν ο χωρισμός του κάθε video στα πλαίσια του. Κάθε video, αποτελείται από περίπου $600-800$ πλαίσια. Για κάθε πλαίσιο υπολογίζουμε τα SIFT χαρακτηριστικά. Ο αλγόριθμος SIFT παράγει περίπου $50-90$ key-points για κάθε πλαίσιο.

Για την εκπαίδευση των συστημάτων, είναι απαραίτητο να αντιστοιχίσουμε ένα διάνυσμα ανά εικόνα. Για το λόγο αυτό, χρησιμοποιούμε την τεχνική BoW όπως περιγράφηκε στο κεφάλαιο 4. Χρησιμοποιούμε τους SIFT descriptors από τις εικόνες σαν χαρακτηριστικό εισόδου. Για το labeling των εικόνων, χρησιμοποιούμε τα αρχεία phone alignment(.phn). Η συχνότητα του video είναι περίπου $23Hz$. Επομένως κάθε πλαίσιο διαρκεί περίπου $\frac{1}{23} = 0.043sec$. Κάθε πλαίσιο αντιστοιχίζεται στο ανάλογο φώνημα. Σε περίπτωση που δύο η περισσότερα φωνήματα (σύμφωνα με το .phn αρχείο) αναλογούν στο ίδιο πλαίσιο, θα επικρατήσει εκείνο το φώνημα που έχει τη μεγαλύτερη διάρκεια. Σε αυτό το στάδιο προκύπτει μία μικρή απώλεια πληροφορίας εξ΄ αιτίας της σχετικά χαμηλής συχνότητας δειγματοληψίας του video. Στην περίπτωσή μας, η απώλεια πληροφορίας προκύπτει επειδή σε κάποια φραμες αντιστοιχίζονται δύο φωνήματα (επιπλέον απώλεια πληροφορίας θα υπήρχε αν υπήρχαν φωνήματα με διάρκεια μικρότερη των $0.043sec$). Εξ΄ αιτίας του τρόπου με τον οποίο λύσαμε το συγκεκριμένο πρόβλημα, υπάρχουν περιπτώσεις που το φώνημα που αντιστοιχήθηκε σε κάποιο πλαίσιο δεν αντιστοιχεί στο συγκεκριμένο πλαίσιο, καθώς αυτό το πλαίσιο στη πράξη αντιστοιχεί στο μεταβατικό στάδιο από το ένα φώνημα στο άλλο. Πιθανοί τρόποι επίλυσης αυτού του προβλήματος είναι:

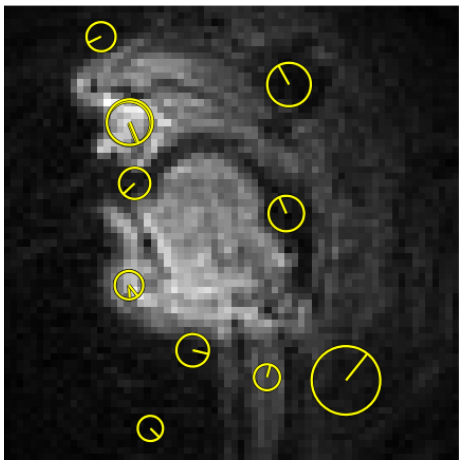
- Προσθήκη επιπλέον ετικετών για την περιγραφή καταστάσεων μετάβασης μεταξύ φωνημάτων
- Αφαίρεση των πλαίσια που αντιστοιχούν σε τέτοιες καταστάσεις

Στη περίπτωσή μας κανένας από αυτούς τους τρόπους δεν είναι εφαρμόσιμος. Αν χρησιμοποιήσουμε τον πρώτο τρόπο, δεν θα προκύψουν αρκετά δεδομένα για κάθε label για την εκπαίδευση των συστημάτων, καθώς το πλήθος δεδομένων που έχουμε δεν είναι τόσο μεγάλο. Εξ΄ αιτίας του σχετικά μικρού πλήθους δεδομένων που έχουμε, δεν έχουμε το περιθώριο να αφαιρέσουμε όλα τα πλαίσια για τα οποία υπάρχει διαμάχη σχετικά με το ποιο φώνημα θα τους αντιστοιχηθεί, καθώς συμβαίνει αρκετά συχνά. Με άλλα λόγια το γεγονός αυτό συμβαίνει αρκετά συχνά για να μην μπορούμε να αφαιρέσουμε τα συγκεκριμένα πλαίσια, αλλά όχι αρκετά συχνά ώστε να μπορούμε να ορίσουμε νέες ετικέτες για την κάθε περίπτωση. Υβριδικές λύσεις που συνδυάζουν και τις δύο προσεγγίσεις, για παράδειγμα να κάνουμε ξεχωριστές ετικέτες μόνο για τις περιπτώσεις που εμφανίζονται αρκετά συχνά ενώ τα υπόλοιπα πλαίσιο να τα αφαιρέσουμε. Όμως ούτε μία τέτοια προσέγγιση είναι εφικτή καθώς δεν έχουμε ούτε μία περίπτωση με αρκετά δεδομένα που θα μπορούσαμε να χρησιμοποιήσουμε ξεχωριστή

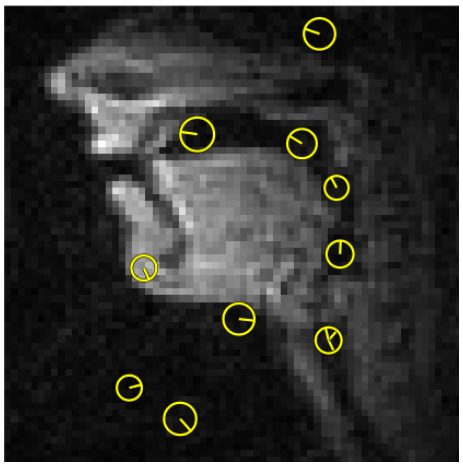
ετικέτα. Επιλέξαμε να συνεχίσουμε τα πειράματά μας χωρίς να εφαρμόσουμε κάτι από τα παραπάνω, καθώς καταλήξαμε στο συμπέρασμα πως μία μικρή απώλεια πληροφοριών είναι η καλύτερη επιλογή που έχουμε.



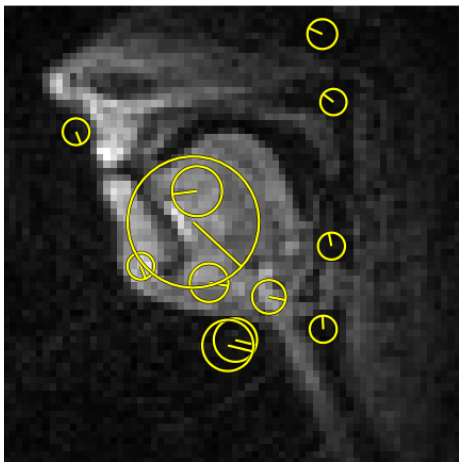
f3: ah φώνημα.



f3: ey φώνημα.



m3: ah φώνημα.



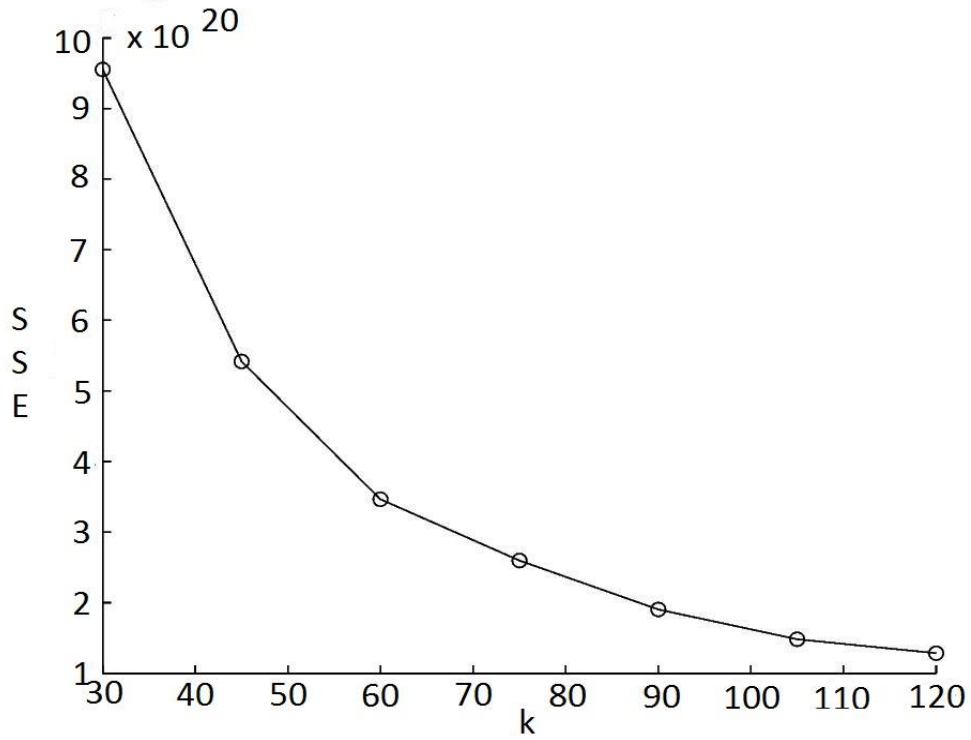
m3: ey φώνημα.

Σχήμα 5.12: SIFT περιγραφείς.

Σε αυτό το σημείο ένας απρόσμενος περιορισμός προέκυψε. Στο στάδιο k-means clustering της μεθόδου BoW, τα μηχανήματά μας είχαν ανεπάρκεια μνήμης RAM. Τα δεδομένα μας είναι περίπου $771(videos) \times 700(frames) \times 70(SIFT\ descriptors) = 37779000$ διανύσματα με 128 float αριθμούς το καθένα(περίπου 180GB δεδομένων χωρίς να λαμβάνουμε υπ' οψην μας τους πίνακες αποθήκευσης των ετικετών). Επομένως αλλάξαμε τα πλάνα μας από speaker independent μοντέλα σε speaker dependent. Επιλέξαμε να μελετήσουμε έναν άνδρα και μία γυναίκα ομιλητή ($m3$, $f3$). Πιθανοί

τρόποι να επιλυθούν προβλήματα σαν αυτό που προέκυψε σε μας συζητιούνται στον επίλογο.

Τώρα πρέπει να καθορίσουμε την τιμή της παραμέτρου k για τον αλγόριθμο k-means clustering. Χρησιμοποιήσαμε τη μέθοδο elbow. Οι τιμές της παραμέτρου k που εξετάσαμε είναι από $k = 30$ έως $k = 120$ με βήμα 15. Το (k, SSE) διάγραμμα φαίνεται παρακάτω (Σχήμα 5.13):



Σχήμα 5.13: $k - SSE$ γράφημα για την επιλογή τις τιμής του k για BoW $k-$means ομαδοποίηση. Για $k > 85$, η αύξηση της τιμής του k έχει σαν αποτέλεσμα αμελητέα μείωση στο SSE.

Εξετάζοντας το διάγραμμα παρατηρούμε ότι για $k > 90$ δεν υπάρχει αξιοσημείωτη μείωση στο SSE όταν αυξάνεται το k. Μία καλή τιμή της παραμέτρου k φαίνεται να είναι κάτι λιγότερο από ενενήντα, έτσι διαλέγουμε $k = 85$ (Σχήμα 5.13).

Το επόμενο στάδιο είναι η δημιουργία των ιστογραμμάτων. Επιλέξαμε τα ιστογράμματα που θα δημιουργήσουμε να είναι δυαδικά. Επιπροσθέτως, δημιουργήσαμε δύο επιπλέον σετ ιστογραμμάτων. Στο πρώτο, εφαρμόσαμε soft normalization ανά διάσταση των φράσεων, αφαιρώντας το μέσο όρο και στη συνέχεια διαιρώντας με το διπλάσιο της τυπικής απόκλισης (συν μία μικρή ποσότητα e για την περίπτωση που η τυπική απόκλιση είναι μηδέν). Στο δεύτερο, εκτός από την προηγούμενη κανονικοποίηση, κανονικοποιήσαμε τα ιστογράμματα και ανά δείγμα έτσι ώστε το άθροισμα

των τετραγώνων κάθε δείγματος να ισούται με μονάδα. Μόνο για το συγκεκριμένο στάδιο, αφαιρέσαμε τα ιστογράμματα που είχαν αντιστοιχηθεί στα φωνήματα *sil* και *sp*.

Για να επιλέξουμε ποιο είδος κανονικοποίησης δουλεύει καλύτερα, θα εκπαιδεύσουμε SVMs classifiers. Ένα νέο πρόβλημα προκύπτει ξανά. Παρόλο που αφαιρέσαμε τα φωνήματα *sil* και *sp* που αντιστοιχούν περίπου στο 40% των δεδομένων, εξακολουθεί να υπάρχει μεγάλη ανισορροπία δεδομένα ανάμεσα στις εναπομείναντες κλάσεις (39). Για την επίλυση του προβλήματος, εφαρμόσαμε τη μέθοδο SMOTE. Η παράμετρος που καθορίζει το πλήθος των γειτόνων ήταν διαφορετική για κάθε κλάση. Ο στόχος μας ήταν, μετά τη δημιουργία των απαραίτητων συνθετικών δειγμάτων σε κάθε κλάση, να έχουμε το ίδιο πλήθος δεδομένων σε κάθε κλάση, όσα έχει δηλαδή η κυρίαρχη κλάση. Για το φώνημα z ωστόσο, δεν ήταν δυνατόν να δημιουργήσουμε τόσα πολλά δείγματα με την εφαρμογή της SMOTE στα αρχικά δεδομένα, ακόμα και όταν χρησιμοποιήσαμε το μέγιστο πλήθος γειτόνων. Η διαφορά μεταξύ του επιθυμητού και του εφικτού πλήθους δεδομένων του φωνήματος z είναι περίπου 7%. Μιας και η διαφορά είναι μικρή συνεχίσαμε τα πειράματά μας χωρίς να δημιουργήσουμε ακόμα περισσότερα δείγματα.

Οι SVMs classifiers εκπαιδεύτηκαν ως εξής. Χρησιμοποιήσαμε RBF kernel με $\gamma = 0.02$, παράμετρο κόστους $C = 1$ για κάθε κλάση. 90% των δεδομένων χρησιμοποιήθηκε για εκπαίδευση και το υπόλοιπο 10% για testing. Για την υλοποίηση του multi-label SVM training χρησιμοποιήθηκε η μέθοδος One-Against- All με χρήση του εργαλείου libsvm στο περιβάλλον του Matlab. Όπως αναμενόταν, τα καλύτερα αποτελέσματα προήλθαν από τα ιστογράμματα που ήταν δύο φορές κανονικοποιημένα. Τα αποτελέσματα φαίνονται στον Πίνακα 5.4.

	$original$	$normalize_1$	$normalize_2$
Σφάλμα	98%	94.3%	**93.5%**

Πίνακας 5.4: Επιλογή της καλύτερης μεθόδου κανονικοποίησης.

Επιλέγουμε να συνεχίσουμε τα πειράματα μας με τα διπλά κανονικοποιημένα ιστογράμματα. Πλέον είμαστε έτοιμοι να εκπαιδεύσουμε τα συστήματά μας χρησιμοποιώντας μόνο τα articulatory χαρακτηριστικά. Καθώς ο τελικός μας στόχος είναι να χρησιμοποιήσουμε audio-articulatory χαρακτηριστικά, πρέπει να λύσουμε ένα πρόβλημα συμβατότητας που θα προκύψει αργότερα.

5.4.1 Αυξάνοντας τη συχνότητα των articulatory χαρακτηριστικών

Η συχνότητα των χαρακτηριστικών ήχου είναι $100Hz$ ενώ η συχνότητα των articulatory χαρακτηριστικών είναι $23Hz$ (η ίδια με το video frame rate). Επιλέγουμε να αντιγράψουμε τα articulatory χαρακτηριστικά όσες φορές χρειάζεται, προκειμένου να ταιριάξουν οι δύο συχνότητες. Αντιγράφουμε κάθε δείγμα τέσσερις φορές και στη συνέχεια αντιγράφουμε μία φορά κάθε δωδέκατο δείγμα. Υπάρχουν και άλλες ιδέες για την επίλυση αυτού του προβλήματος. Κάποιος θα μπορούσε να εφαρμόσει γραμμική παρεμβολή ανάμεσα στα υπάρχοντα articulatory χαρακτηριστικά για να δημιουργήσει καινούρια. Μια τέτοια προσέγγιση όμως δεν θα είχε νόημα στη πράξη, καθώς οι SIFT descriptors από συνεχόμενες εικόνες δεν έχουν γραμμική σχέση, συνεπώς αυτή η ιδέα απορρίφθηκε. Για τον ίδιο λόγο απορρίφθηκε και η ιδέα να εφαρμόσουμε γραμμική παρεμβολή στις εικόνες και να υπολογίσουμε στις καινούριες εικόνε τους SIFT descriptors.

Διαιρούμε τα δεδομένα μας ως εξής: 85% για το *train* σετ, 7% για το *dev* σετ και 8% για το *test* σετ. Καθώς δεν έχουμε αρκετά δεδομένα για να εκπαιδεύσουμε speaker dependent μοντέλα χρησιμοποιώντας DNNs, θα χρησιμοποιήσουμε HMMs για τη δημιουργία των συστημάτων μας. Συνοπτικά αναφέρουμε πως από τα τέσσερα συστήματα που εκπαιδεύσαμε, το πρώτο είναι ένα monophone σύστημα ενώ τα υπόλοιπα είναι triphone συστήματα. Για τα triphone συστήματα, το πρώτο χρησιμοποιεί MFCCA (MFCC+CCA) μαζί με την πρώτη και τη δεύτερη παράγωγο σαν διάνυσμα εισόδου. Το δεύτερο εφαρμόζει δύο μετασχηματισμούς (Linear Discriminant Analysis (LDA) και Maximum Likelihood Linear Transform (MLLT)) επιπλέον στο διάνυσμα εισόδου ενώ το τρίτο εφαρμόζει μαζί με όσα εφάρμοσαν τα προηγούμενα συστήματα Speaker Adaptive Training (SAT). Τα αποτελέσματα του phone error rate φαίνονται στον Πίνακα 5.4.1 για τους δύο ομιλητές. Χρησιμοποιήσαμε 7−fold cross-validation για να κάνουμε validate τα αποτελέσματα.

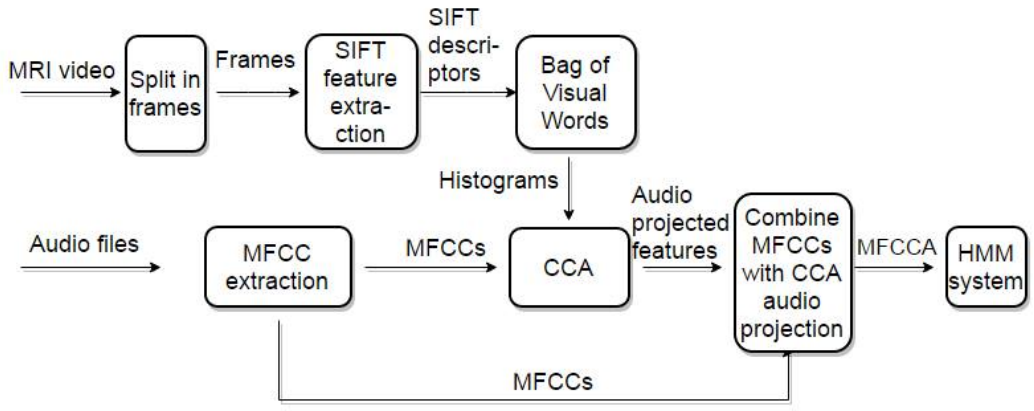
Articulatory	*mono*	tri_1	tri_2	tri_3
$f3$ ομιλητής	75.23%	85.15%	79.32%	79.87%
$m3$ ομιλητής	79.58%	84.71%	80.45%	78.06%

Πίνακας 5.5: Cross-validated αποτελέσματα αναγνώρισης ομιλίας για τους ομιλητές $f3$ και $m3$.

5.5 Ηχητικά - Articulatory πειράματα

Σε αυτό το σημείο, χρησιμοποιούμε 13-διάστατα χαρακτηριστικά audio (MFCCs με χρήση 25 ms window με 10 ms μετατόπιση) και 85-διάστατα (double normalized) articulatory χαρακτηριστικά. Καθώς είναι ιδιαίτερα δύσκολο να αποκτάμε MRI δεδομένα για εφαρμογές στην πραγματική ζωή, υποθέτουμε ότι η articulatory πληροφορία είναι διαθέσιμη μόνο κατά το στάδιο της εκπαίδευσης του συστήματός μας. Εφαρμόζουμε CCA στα train (και dev) δεδομένα και βρίσκουμε τον audio πίνακα μετασχηματισμού με διαστάσεις 13×13. Μετασχηματίζουμε τα audio χαρακτηριστικά του train και του test σετ και ενώνουμε τα MFCCs με τις CCA audio προβολές ώστε να δημιουργήσουμε τα MFCCA χαρακτηριστικά. Η ιδέα πίσω από αυτή την προσέγγιση είναι πως ο θόρυβος στο πεδίο της εικόνας είναι σε μεγάλο βαθμό ασυσχέτιστος με τον θόρυβο στο πεδίο του ήχου. Επομένως, με το εφαρμόσουμε τέτοιου είδους προβολές, διατηρούμε κατά κύριο λόγο την πληροφορία από το σήμα μας χωρίς τον θόρυβο. Ο audio πίνακας μετασχηματισμού που υπολογίζεται μέσω της CCA χρησιμοποιείται για το μετασχηματισμό των ακουστικών χαρακτηριστικών, τόσο του train όσο και του test σετ. Για περαιτέρω βελτίωση των αποτελεσμάτων, συνδυάζουμε τις CCA προβολές των audio χαρακτηριστικών με τα αρχικά αυδιο χαρακτηριστικά ώστε να δημιουργήσουμε τα MFCC-articulatory χαρακτηριστικά ή αλλιώς MFCCA [69], για βέλτιστα αποτελέσματα, καθώς δεν είναι όλη η ασυσχέτιστη πληροφορία που αφαιρείται θόρυβος [38].

Η πλήρης αρχιτεκτονική του συστήματος φαίνεται στο Σχήμα 5.14



Σχήμα 5.14: Multi-view σύστημα για τη βάση δεδομένων rt-MRI-TIMIT.

Για την εκπαίδευση των συστημάτων διαιρούμε τα δεδομένα και χρησιμοποιούμε τις ίδιες τεχνικές με αυτές που περιγράψαμε στο προηγούμενο κεφάλαιο. Και πάλι χρησιμοποιούμε 7−fold cross-validation για να κάνουμε validate τα αποτελέσματα. Χρησιμοποιούμε MFCCA σαν χαρακτηριστικά εισόδου για την εκπαίδευση των audio-articulatory συστημάτων.

Τα πλήρη αποτελέσματα από όλα τα πειράματα για τα speaker dependent μοντέλα φαίνονται στους ακόλουθους πίνακες:

$f3$	$mono$	tri_1	tri_2	tri_3
MFCC	70.24%	67.06%	68.41%	67.87%
Articulatory	75.23%	85.15%	79.32%	79.87%
MFCCA	67.69%	**65.21%**	67.13%	66.41%

Πίνακας 5.6: Cross-validated αποτελέσματα αναγνώρισης ομιλίας για την ομιλήτρια $f3$.

$m3$	$mono$	tri_1	tri_2	tri_3
MFCC	67.70%	63.74%	65.90%	64.46%
Articulatory	79.58%	84.71%	80.45%	78.06%
MFCCA	66.03%	**61.73%**	64.30%	63.47%

Πίνακας 5.7: Cross-validated αποτελέσματα αναγνώρισης ομιλίας για τον ομιλητή $m3$.

5.6 Επιπλέον πειράματα με χρήση DNNs

Προσπαθήσαμε να χρησιμοποιήσουμε $DNNs$ για το speaker dependent μοντέλο του $f3$ με διάφορες τιμές πολλών παραμέτρων για την περαιτέρω βελτίωση της απόδοσης των articulatory-ακουστικών συστημάτων. Οι παράμετροι που εξετάσαμε είναι: ο αριθμός των hidden layers, ο μέγιστος αριθμός επαναλήψεων, η διάσταση του hidden layer και η επεξεργασία των χαρακτηριστικών εισόδου. Τα αποτελέσματα φαίνονται παρακάτω

Όπως μπορούμε να δούμε τα καλύτερα αποτελέσματα προήλθαν από το DNN με 1 hidden layer, 20 επαναλήψεις και 2048 διαστάσεις hidden layer, χωρίς να επηρεάζει το αποτέλεσμα αν τα χαρακτηριστικά ήταν επεξεργασμένα ή όχι. Παρόλο που το αποτέλεσμα είναι καλύτερο από τα articulatory αποτελέσματα, δεν είναι καν κοντά στα αποτελέσματα προερχόμενα από άλλες μεθόδους εκπαίδευσης με articulatory-ακουστικά χαρακτηριστικά. Ακόμα και τα συστήματα ήχου εκπαιδευμένα με MFCCs έχουν καλύτερη απόδοση. Θεωρούμε ότι το πλήθος των δεδομένων είναι ο κύριος λόγος για άλλη μια φορά. Αφού δεν υπήρχε καμία βελτίωση με χρήση $DNNs$ για τον ομιλητή $f3$,δεν ελέγξαμε την απόδοση των $DNNs$ για τον ομιλητή $m3$.

Hidden layers	Iterations	Dimensions	Feature Processing	Result
6	20	2048	Yes	78.8% (Default)
1	20	2048	Yes	74.8%
2	20	2048	Yes	75.3%
1	10	2048	Yes	74.8%
1	20	2048	No	76.2%
2	20	2048	No	80.6%
1	20	128	Yes	75.6%
1	20	4096	Yes	77.9%
1	20	3072	Yes	76.1%
6	10	2048	Yes	77.9%
6	30	2048	Yes	75.9%

Πίνακας 5.8: Αποτελέσματα των πειραμάτων $DNNs$ για διάφορες τιμές των παραμέτρων για την ομιλήτρια $f3$ χρησιμοποιώντας ηχητικά - articulatory χαρακτηριστικά.

Κεφάλαιο 6

Συμπεράσματα

6.1 Ανακεφαλαίωση

Τα αποτελέσματα δείχνουν ότι ακόμα και με χαμηλής ποιότητας εικόνες MRI το αποτέλεσμα της αναγνώρισης μπορεί να βελτιωθεί αν χρησιμοποιήσουμε articulatory δεδομένα. Βλέπουμε ότι παρά το γεγονός ότι η επίδοση δεν ήταν ικανοποιητική χρησιμοποιώντας μόνο articulatory χαρακτηριστικά, ο συνδυασμός με τα MFCCs απέδωσε αρκετά ικανοποιητικά. Αυτό ήταν αναμενόμενο καθώς εχει δειχθεί στο παρελθόν από πολλές εργασίες σε διάφορα σύνολα δεδομένων όπου τα articulatory χαρακτηριστικά βελτιώνουν την αναγνώριση. Στην περίπτωσή μας, μπορούμε να δούμε ότι και τα τέσσερα συστήματα έχουν βελτιωμένη απόδοση περίπου 1.75% και 1.5% με τη μεγαλύτερη βελτίωση να προέρχεται από το *monophone* και $triphone_1$ σύστημα για τον $f3$ και $m3$ αντίστοιχα. Παρόλο που το $triphone_1$ σύστημα είχε την καλύτερη απόδοση και για τους δύο ομιλητές, περιμέναμε επίσης ένα από τα triphone συστήματα να έχει τη μεγαλύτερη βελτίωση σε απόδοση και στις δύο περιπτώσεις αλλά για τον $f3$ το monophone δούλεψε καλύτερα συγκρίνοντας τα MFCC με τα MFCCA αποτελέσματα. Μια πιθανή αιτία ίσως είναι τα λιγότερα δεδομένα για τον $f3$ καθώς πιο πολύπλοκα συστήματα γενικότερα χρειάζονται περισσότερα δεδομένα για τη σωστή λειτουργία τους σε σχέση με τα πιο απλά. Αυτό υποστηρίζεται ακόμα περισσότερο από το γεγονός ότι τα *monophone* συστήματα είχαν τα καλύτερα αποτελέσματα μεταξύ των άλλων articulatory συστημάτων.

Για τα συστήματα ήχου παρατηρούμε ότι τα $SGMM$, $DNNs$ και $Hybrid$ συστήματα έχουν την καλύτερη απόδοση. Αναμέναμε ότι τα DNNs θα ξεπερνούσαν σε απόδοση όλες τις άλλες μεθόδους εκπαίδευσης. Ωστόσο αυτό δε συνέβη. Μια πιθανή εξήγηση μπορεί να είναι το μικρό πλήθος των δεδομένων που διαθέτουμε, γεγονός που δυσκολεύει το DNN στο να συλλάβει τη συσχέτιση μεταξύ των χαρακτηριστικών.

6.2 Συμπεράσματα

Οι κύριες συμβολές αυτής της διπλωματικής εργασίας για τη βελτιστοποίηση των συστημάτων αναγνώρισης ομιλίας είναι:

- Μελέτη της σχετικής βιβλιογραφίας και παρουσίαση των δημοφιλέστερων τάσεων στο χώρο.
- Παρουσίαση του μαθηματικού υπόβαθρου και των κύριων ιδεών από τους αλγορίθμους που χρησιμοποιήθηκαν.
- Διάφορα πειράματα στην rtMRI-TIMIT βάση δεδομένων.
- Σκεφτήκαμε επίσης να συνδυάσουμε τις ιδέες μας με το έργο άλλων ερευνητών. Πιο συγκεκριμένα, σκεφτήκαμε να εφαρμόσουμε μάσκες ή τεχνικές εύρεσης ακμών με σκοπό να απομονώσουμε τις περιοχές ενδιαφέροντος, για παράδειγμα, τη στοματική κοιλότητα, και να εφαρμόσουμε τον αλγόριθμο SIFT μόνο σε αυτές τις περιοχές. Με αυτόν τον τρόπο αποκόπτουμε τους descriptors που είναι σε περιοχές εκτός ενδιαφέροντος, που είναι επιπλέον βήμα αποθορυβοποίησης. Δεν κάναμε μια τέτοια προσέγγιση γιατί προκύπτουν επιπρόσθετα προβλήματα. Για παράδειγμα,το κεφάλι του ομιλητή δεν είναι τελείως σταθερό καθ' όλη τη διάρκεια της ηχογράφησης,συνεπώς σημαντικά μέρη του κεφαλιού μπορεί μερικές φορές να πέφτουν εκτός των ορίων της μάσκας. Η εφαρμογή των τεχνικών εύρεσης ακμών είναι δύσκολη επίσης καθώς οι εικόνες έχουν χαμηλή ανάλυση. Για όλους αυτούς τους λόγους επιλέξαμε να εφαρμόσουμε τον αλγόριθμο SIFT σε όλη την εικόνα και να εφαρμόσουμε τεχνικές αποθορυβοποίησης σε επόμενο στάδιο. Η εφαρμογή μασκών χρειάζεται επιπλέον ανθρώπινη ερμηνεία για τον καθορισμό ενός σταθερού σημείου (συνήθως τη βάση της μύτης) για να εφαρμοστεί η μάσκα. Η δική μας προσέγγιση δε χρειάζεται καμία ανθρώπινη παρέμβαση.
- Βελτιωμένα αποτελέσματα της προσέγγισης μας συγκριτικά με τη συνηθισμένη προσέγγιση.

6.3 Περαιτέρω ιδέες και επεκτάσεις

Υπήρχαν πολλές ιδέες άξιες εξέτασης οι οποίες συνέβησαν κατά τη διάρκεια αυτής της διπλωματικής. Θα αναφέρουμε τις πιο ενδιαφέρουσες μεταξύ αυτών έτσι ώστε να επισημάνουμε τις πιο υποσχόμενες ερευνητικές επιλογές.

- Στα πειράματά μας χρησιμοποιήσαμε SIFT descriptors για να επωφεληθούμε από την articulatory πληροφορία των MRI εικόνων. Κάποιος μπορεί να χρησιμοποιήσει διαφορετικούς τύπους descriptors. Μια ενδιαφέρουσα επιλογή θα ήταν οι SURF ή οι PCA-SIFT descriptors. Το κύριο πλεονέκτημα αυτών των descriptors είναι ότι είναι λιγότερο δαπανηροί συγκριτικά με τον αλγόριθμο

SIFT. Όσο αυξάνεται το κόστος του αλγορίθμου,τόσο αυξάνεται και η απόδοση. Στην περίπτωσή μας όμως,όπου το πρόβλημα δεν είναι τόσο πολύπλοκο αφού οι εικόνες έχουν χαμηλή ποιότητα, λιγότερο πολύπλοκοι αλγόριθμοι μπορεί να έχουν παρόμοια ή καλύτερα αποτελέσματα.

- Μια άλλη επιλογή είναι να εφαρμόσουμε την ιδέα των tri-phone συστημάτων στο BoW στάδιο. Αφού λάβαμε το 85 dimension χαρακτηριστικό ανά εικόνα και το κανονικοποιήσαμε με δύο τρόπους (η κανονικοποίηση είναι προαιρετική σε αυτό το στάδιο),μπορούμε να ενώσουμε το προηγούμενο,το τωρινό και το επόμενο ιστόγραμμα για να κατασκευάσουμε ένα 255 dimension ιστόγραμμα του παρόντος φωνήματος (μπορούμε να εφαρμόσουμε ξανά κανονικοποίηση με δύο τρόπους στο σημείο αυτό). Οι νέοι άξονες μπορούν να χρησιμοποιηθούν σαν τελικά articulatory χαρακτηριστικά.

- Είχαμε δύο ενδιαφέρουσες ιδέες στο k-means μέρος του BoW σταδίου για να χειριστούμε το μεγάλο πλήθος δεδομένων.

    1. Η πρώτη ήταν να μη χρησιμοποιήσουμε όλο το σύνολο δεδομένων αλλά μερικά δείγματα όλων των φωνημάτων από κάθε ομιλητή για να εφαρμόσουμε k-means ομαδοποίηση.Με αυτόν τον τρόπο το μέγεθος του συνόλου δεδομένων θα μειωνόταν και θα ήταν πιθανώς εύχρηστο από τα μηχανήματά μας. Επιλέγουμε να μη χρησιμοποιήσουμε αυτήν την ιδέα εξαιτίας του αυξημένου θορύβου στο σύνολο δεδομένων,κάτι που είναι πολύ πιθανό να κάνει την πλειοψηφία των εναπομείναντων δειγμάτων στο σύνολο δεδομένων να είναι θορυβώδη.

    2. Η δεύτερη ιδέα ήταν μια προσπάθεια να μετατρέψουμε τον αλγόριθμο k-means έτσι ώστε να τρέχει παράλληλα. Πρώτα τρέχουμε τον k-means για καθένα από τους δέκα ομιλητές ξεχωριστά και κρατάμε τα κέντρα. Επίσης εφαρμόζουμε σε κάθε ομαδοποιημένο κέντρο το κέντρο ενός κυκλικού patch με άπειρη ακτίνα,διαιρούμενο σε οκτώ (μπορούμε να επιλέξουμε και άλλους αριθμούς) ίσους τομείς (τόξα). Βρίσκουμε το κέντρο μάζας των σημείων που έχουν ανατεθεί στη συγκεκριμένη κλάση για κάθε τομέα. Για κάθε τομέα αποθηκεύουμε των αριθμό των δειγμάτων και το κέντρο μάζας. Έπειτα επιλέγουμε στην τύχη έναν ομιλητή ως οδηγό. Για κάθε επόμενο ομιλητή κάνουμε τα ακόλουθα: Υπολογίζουμε τις αποστάσεις μεταξύ των κέντρων όλων τον ομιλητών και των κέντρων του οδηγού-ομιλητή και ομαδοποιούμε τα κοντινότερα κέντρα. Η αντιστοιχία μεταξύ των κέντρων του οδηγού ομιλητή και των κέντρων του κάθε ενός από τους υπόλοιπους ομιλητές θα πρέπει να είναι ένα προς ένα. Αντιστοιχούμε τα κέντρα όλων των ομιλητών στα κοντινότερα κέντρα του ομιλητή-οδηγού. Πλέον έχουμε τόσες ομάδες κέντρων όσο και το πλήθος των κλάσεων και η κάθε ομάδα έχει τόσα στοιχεία όσα και το πλήθος των ομιλητών. Για κάθε μία από αυτές τις ομάδες, υπολογίζουμε το κέντρο μάζας βασιζόμενοι στους τομείς των πατσηες, χρησιμοποιώντας σα συντελεστή βαρύτητας το πλήθος των

δεδομένων που αντιστοιχούν σε κάθε σημείο του patch. Τα τελικά κέντρα μάζας είναι η τελική έξοδος του τροποποιημένου αλγορίθμου k-means.

Κάποιος μπορεί να χρησιμοποιήσει άλλες τεχνικές clustering αντί για k-means clustering όπως τον DB-scan αλγόριθμο. Εμείς όμως επιλέγουμε να ακολουθήσουμε στην κλασσική BoW pipeline προσέγγιση.

Παρότι προτείνουμε μερικές τεχνικές για την απόκτηση μοντέλων ανεξάρτητα του ομιλητή, πρέπει να γίνει περισσότερη δουλειά προς την κατεύθυνση της γενίκευσης αυτών των αποτελεσμάτων σε ολόκληρη τη βάση δεδομένων rtMRI. Η διαχείριση με κατάλληλο τρόπο όλων των διαθέσιμων πληροφοριών είναι το δύσκολο κομμάτι καθώς, στην προσέγγιση μας, αυτός ήταν ο βασικός περιορισμός που αντιμετωπίσαμε και μας έκανε να αλλάξουμε την προσέγγιση μας και να στραφούμε σε μοντέλα εξαρτώμενα από τον ομιλητή. Τα αποτελέσματά μας είναι υποσχόμενα συνεπώς πιστεύουμε ότι η δημιουργία μοντέλων ανεξάρτητα του ομιλητή με βελτιωμένα αποτελέσματα και με χρήση της προσέγγισής μας είναι πολύ πιθανή.

Παράρτημα Α΄

MRI-TIMIT corpus

This was easy for us.
Jane may earn more money by working hard.
She is thinner than I am.
Bright sunshine shimmers on the ocean.
Nothing is as offensive as innocence.

Why yell or worry over silly items.
Where were you while we were away?
Are your grades higher or lower than Nancy's.
He will allow a rare lie.
Will robin wear a yellow lily.

Swing your arm as high as you can.
Before Thursday's exam review every formula.
The museum hires musicians every evening.
A roll of wire lay near the wall.
Carl lives in a lively home.

Alimony harms a divorced man's wealth.
She wore warm fleecy woolen overalls.
Alfalfa is healthy for you.
When all else fails use force.
Although always alone we survive.

Only lawyers love millionaires.
Did dad do academic bidding?
Help greg to pick a peck of potatoes.
A good attitude is unbeatable.
Coconut cream pie makes a nice dessert.

Don't do Charlie's dirty dishes.
Help celebrate your brother's success.
Only the most accomplished artists obtain popularity.
Critical equipment needs proper maintenance.
Young people participate in athletic activities.

Barb's gold bracelet was a graduation present.
Stimulating discussions keep students' attention.
Etiquette mandates compliance with existing regulations.
Biblical scholars argue history.
Elderly people are often excluded.

Basketball can be an entertaining sport.
Addition and subtraction are learned skills.
Grandmother outgrew her upbringing in petticoats.
At twilight on the twelfth day we'll have chablis.
Catastrophic economic cutbacks neglect the poor.

Ambidextrous pickpockets accomplish more.
Even a simple vocabulary contains symbols.
The eastern coast is a place for pure pleasure and excitement.
The lack of heat compounded the tenant's grievances.
Academic aptitude guarantees your diploma.

The prowler wore a ski mask for disguise.
We experience distress and frustration obtaining our degrees.
The singer's finger had a splinter.
The legislature met to judge the state of public education.
Chocolate and roses never fail as a romantic gift.

Any contributions will be greatly appreciated.
Continental drift is a geological theory.
Regular attendance is seldom required.
Challenge each general's intelligence.
We got drenched from the uninterrupted rain.

Last year's gas shortage caused steep price increases.
Upgrade your status to reflect your wealth.
Eat your raisins outdoors on the porch steps.
Porcupines resemble sea urchins.
Spring street is straight ahead.

Cliff's display was misplaced on the screen.
An official deadline cannot be postponed.
Fill that canteen with fresh spring water.
Gently place Jim's foam sculpture in the box.
Bagpipes and bongos are musical instruments.

Doctors prescribe drugs too freely.
Will you please describe the idiotic predicament.
It's impossible to deal with bureaucracy.
Good service should be rewarded by big tips.
My instructions desperately need updating.

Cooperation along with understanding alleviate dispute.
Cement is measured in cubic yards.
Primitive tribes have an upbeat attitude.
Flying standby can be practical if you want to save money.
It's hard to tell an original from a forgery.

'The thinker' is a famous sculpture.
The misprint provoked an immediate disclaimer.
A large household needs lots of appliances.
Cut a small corner off each edge.
Iguanas and alligators are tropical reptiles.

Masquerade parties tax one's imagination.
Penguins live near the icy antarctic.
Guess the question from the answer.
Medieval society was based on hierarchies.
Project development was proceeding too slowly.

Kindergarten children decorate their classrooms for all holidays.
Special task forces rescue hostages from kidnappers.
Call an ambulance for medical assistance.
He stole a dime from a beggar.
You must explicitly delete files.

A huge tapestry hung in her hallway.
Birthday parties have cupcakes and ice cream.
His scalp was blistered from today's hot sun.
She slipped and sprained her ankle on the steep slope.
The best way to learn is to solve extra problems.

His sudden departure shocked the cast.
Tugboats are capable of hauling huge loads.
A muscular abdomen is good for your back.
The cartoon features a muskrat and a tadpole.
The emblem depicts the acropolis all aglow.

Clasp the screw in your left hand.
The mango and the papaya are in a bowl.
Combine all the ingredients in a large bowl.
The misquote was retracted with an apology.
The coyote bobcat and hyena are wild animals.

Trespassing is forbidden and subject to penalty.
Encyclopedias seldom present anecdotal evidence.
A screwdriver is made from vodka and orange juice.
Objects made of pewter are beautiful.
Westchester is a county in New York.

Artificial intelligence is for real.
The emperor had a mean temper.
Lots of foreign movies have subtitles.
Angora cats are furrier than siamese.
He ate four extra eggs for breakfast.

We plan to build a new beverage plant.
Publicity and notoriety go hand in hand.
Pizzerias are convenient for a quick lunch.
December and january are nice months to spend in Miami.
Technical writers can abbreviate in bibliographies.

Scientific progress comes from the development of new techniques.
The clumsy customer spilled some expensive perfume.
The bungalow was pleasantly situated near the shore.
Agricultural products are unevenly distributed.
Pledge to participate in nevada's aquatic competition.

Which long article was opaque and needed clarification?
The sound of jennifer's bugle scared the antelope.
The willowy woman wore a muskrat coat.
Too much curiosity can get you into trouble.
Cyclical programs will never compile.

Correct execution of my instructions is crucial.
Most precincts had a third of the votes counted.
While waiting for Chipper, she crisscrossed the square many times.
Vietnamese cuisine is exquisite.
The previous speaker presented ambiguous results.

Mosquitoes exist in warm humid climates.
Scholastic aptitude is judged by standardized tests.
Orange juice tastes funny after toothpaste.
The water contained too much chlorine and stung his eyes.
Our experiment's positive outcome was unexpected.

Remove the splinter with a pair of tweezers.
The drunkard is a social outcast.
The government sought authorization of his citizenship.
As coauthors we presented our new book to the haughty audience.
As a precaution the outlaws bought gunpowder for their stronghold.

Her auburn hair reminded him of autumn leaves.
They remained lifelong friends and companions.
Curiosity and mediocrity seldom coexist.
Biologists use radioactive isotopes to study microorganisms.
Employee layoffs coincided with the company's reorganization.

Who took the kayak down the bayou?
How would you evaluate this algebraic expression?
The Mayan neoclassic scholar disappeared while surveying ancient ruins.
The diagnosis was discouraging however he was not overly worried.
The triumphant warrior exhibited naive heroism.

Whoever cooperates in finding Nan's cameo will be rewarded.
Severe myopia contributed to ron's inferiority complex.
Buying a thoroughbred horse requires intuition and expertise.
Does creole cooking use curry?
She encouraged her children to make their own halloween costumes.

We could barely see the fjords through the snow flurries.
Almost all colleges are now coeducational.
Rich looked for spotted hyenas and jaguars on the safari.
Thick glue oozed out of the tube.
Why else would Danny allow others to go?

The cat's meow always hurts my ears.
Did you buy any corduroy overalls?
Would a tomboy often play outdoors.
They often go out in the evening.
Who authorized the unlimited expense account?

Destroy every file related to my audits.
Serve the coleslaw after i add the oil.
Withdraw all phony accusations at once.
Straw hats are out of fashion this year.
Why buy oil when you always use mine?

They enjoy it when i audition.
Would you allow acts of violence.
How do oysters make pearls?
Draw each graph on a new axis.
Norwegian sweaters are made of lamb's wool.

Young children should avoid exposure to contagious diseases.
Ralph controlled the stopwatch from the bleachers.
Approach your interview with statuesque composure.
How much allowance do you get?
The causeway ended abruptly at the shore.

Even i occasionally get the Monday blues.
Military personnel are expected to obey government orders.
When peeling an orange it is hard not to spray juice.
Do you hear the sleigh bells ringing?
Rob sat by the pond and sketched the stray geese.

Michael colored the bedroom wall with crayons.
Only the best players enjoy popularity.
I gave them several choices and let them set the priorities.
The news agency hired a great journalist.
The morning dew on the spider web glistened in the sun.

The sermon emphasized the need for affirmative action.
The small boy put the worm on the hook.
How permanent are their records?
Try to recall the events in chronological order.
The most recent geological survey found seismic activity.

Cory attacked the project with extra determination.
You always come up with pathological examples.
Put the butcher block table in the garage.
How good is your endurance?
Keep the thermometer under your tongue.

Steph could barely handle the psychological trauma.
It's healthier to cook without sugar.
The viewpoint overlooked the ocean.
Are you looking for employment.
His failure to open the store by eight cost him his job.

Highway and freeway mean the same thing.
The paper boy bought two apples and three ices.
I itemize all accounts in my agency.
Clear pronunciation is appreciated.
The courier was a dwarf.

A doctor was in the ambulance with the patient.
Puree some fruit before preparing the skewers.
It's not easy to create illuminating examples.
The hallway opens into a huge chamber.
They all agree that the essay is barely intelligible.

How ancient is this subway escalator?
The cigarettes in the clay ashtray overflowed onto the oak table.
Reading in poor light gives you eyestrain.
I ate every oyster on Nora's plate.
The boston ballet overcame their funding shortage.

The gorgeous butterfly ate a lot of nectar.
By eating yogurt you may live longer.
Do they allow atheists in church?
My ideal morning begins with hot coffee.
The irate actor stomped away idiotically.

We are open every Monday evening.
The essay undeniably reflects our view ably.
Remember to allow identical twins to enter freely.
Do you have the yellow ointment ready?
Can the agency overthrow alien forces.

How oily do you like your salad dressing?
We saw eight tiny icicles below our roof.
The saw is broken so chop the wood instead.
Withdraw only as much money as you need.
Draw every outer line first then fill in the interior.

The jaw operates by using antagonistic muscles.
Do atypical farmers grow oats?
Are holiday aprons available to us.
Be careful not to plow over the flower beds.
Allow each child to have an ice pop.

The angry boy answered but didn't look up.
Cliff was soothed by the luxurious massage.
Steve wore a bright red cashmere sweater.
John's brother repainted the garage door.
The rose corsage smelled sweet.

To further his prestige he occasionally reads the Wall Street Journal.
Alice's ability to work without supervision is noteworthy.
The oasis was a mirage.
Cory and trish played tag with beach balls for hours.
The tooth fairy forgot to come when Roger's tooth fell out.

Planned parenthood organizations promote birth control.
Rich purchased several signed lithographs.
In every major cloverleaf traffic sometimes gets backed up.
In the long run it pays to buy quality clothing.
Brush fires are common in the dry underbrush of Nevada.

Weatherproof galoshes are very useful in seattle.
This brochure is particularly informative for a prospective buyer.
The avalanche triggered a minor earthquake.
These exclusive documents must be locked up at all times.
Please take this dirty table cloth to the cleaners for me.

Should giraffes be kept in small zoos?
If Carol comes tomorrow have her arrange for a meeting at two.
The two artists exchanged autographs.
I'd rather not buy these shoes than be overcharged.
Shaving cream is a popular item on Halloween.

Amoebas change shape constantly.
Tofu is made from processed soybeans.
The bluejay flew over the high building.
Cheap stockings run the first time they're worn.
Cottage cheese with chives is delicious.

A chosen few will become generals.
The meeting is now adjourned.
Shipbuilding is a most fascinating process.
The proof that you are seeking is not available in books.
The hood of the jeep was steaming in the hot sun.

My desires are simple: give me one informative paragraph on the subject.
Those answers will be straightforward if you think them through carefully first.
Drop five forms in the box before you go out.
If people were more generous there would be no need for welfare.
Bob found more clams at the ocean's edge.

That dog chases cats mercilessly.
The cranberry bog gets very pretty in autumn.
A big goat idly ambled through the farmyard.
The nearest synagogue may not be within walking distance.
The groundhog clearly saw his shadow but stayed out only a moment.

A leather handbag would be a suitable gift.
The fog prevented them from arriving on time.
The local drugstore was charged with illegally dispensing tranquilizers.
The full moon shone brightly that night.
Steve collects rare and novel coins.

Al received a joint appointment in the biology and the engineering departments.
Gregory and Tom chose to watch cartoons in the afternoon.
Chip postponed alimony payments until the latest possible date.
Count the number of teaspoons of soysauce that you add.
The big dog loved to chew on the old rag doll.

Todd placed top priority on getting his bike fixed.
An adult male baboon's teeth are not suitable for eating shellfish.
Often you'll get back more than you put in.
Gus saw pine trees and redwoods on his walk through Sequoia National Forest.
Bob bandaged both wounds with the skill of a doctor.

The dark murky lagoon wound around for miles.
Did Shawn catch that big goose without help?
Ducks have webbed feet and colorful feathers.
The high security prison was surrounded by barbed wire.
Take charge of choosing her bride's maids' gowns.

The frightened child was gently subdued by his big brother.
I know I didn't meet her early enough.
The barracuda recoiled from the serpent's poisonous fangs.
The patient and the surgeon are both recuperating from the lengthy operation.
I'll have a scoop of that exotic purple and turquoise sherbet.

The preschooler couldn't verbalize her feelings about the emergency conditions.
Many wealthy tycoons splurged and bought both a yacht and a schooner.
The new suburbanites worked hard on refurbishing their older home.
According to my interpretation of the problem, two lines must be perpendicular.
A connoisseur will enjoy this shellfish dish.

A lawyer was appointed to execute her will.
Dolphins are intelligent marine mammals.
Diane may splurge and buy a turquoise necklace.
The moisture in my eyes is from eyedrops, not from tears.
George seldom watches daytime movies.

The system may break down soon so save your files frequently.
I assume moisture will damage this ship's hull.
The annoying raccoons slipped into Phil's garden every night.
The cow wandered from the farmland and became lost.
Each untimely income loss coincided with the breakdown of a heating system part.

The gunman kept his victim cornered at gunpoint for three hours.
Will you please confirm government policy regarding waste removal.
The surplus shoes were sold at a discount price.
Lori's costume needed black gloves to be completely elegant.
Bob papered over the living room murals.

That noise problem grows more annoying each day.
Right now may not be the best time for business mergers.
That diagram makes sense only after much study.
A boring novel is a superb sleeping pill.
John cleans shellfish for a living.

Women may never become completely equal to men.
She always jokes about too much garlic in his food.
I just saw Jim near the new archeological museum.
Pam gives driving lessons on Thursdays.
Why charge money for such garbage?

We welcome many new students each year.
George is paranoid about a future gas shortage.
The carpet cleaners shampooed our oriental rug.
Please shorten this skirt for Joyce.
His shoulder felt as if it were broken.

Which church do the Smiths worship in?

The giant redwoods shimmered in the glistening sun.
Her right hand aches whenever the barometric pressure changes.
They own a big house in the remote countryside.
He picked up nine pairs of socks for each brother.

They all like long hot showers.
A young mouse scampered across the field and disappeared.
She uses both names interchangeably.
Calcium makes bones and teeth strong.
The fish began to leap frantically on the surface of the small lake.

Tim takes Sheila to see movies twice a week.
They assume no burglar will ever enter here.
Just drop notices in any suggestion box.
The taxicab broke down and caused a traffic jam.
We'll serve rhubarb pie after Rachel's talk.

Her wardrobe consists of only skirts and blouses.
Barb burned paper and leaves in a big bonfire.
Of course you can have another tunafish sandwich.
There was a gigantic wasp next to Irving's big top hat.
Laugh, dance and sing, if fortune smiles upon you.

I'd ride the subway but I haven't enough change.
Eating spinach nightly increases strength miraculously.
Butterscotch fudge goes well with vanilla ice cream.
Daphne's Swedish needlepoint scarf matched her skirt.
Irish youngsters eat fresh kippers for breakfast.

Move the garbage nearer to the large window.
A huge power outage rarely occurs.
Valley lodge yearly celebrates the first calf born.
Iris thinks this zoo has eleven Spanish zebras.
Those who teach values first abolish cheating.

Once you finish greasing your chain be sure to wash thoroughly.
Smash lightbulbs and their cash value will diminish to nothing.
Top zinnias rarely have crooked stems.
Movies never have enough villains.
Every cab needs repainting often.

A crab challenged me but a quick stab vanquished him.
A toothpaste tube should be squeezed from the bottom.
Those who are not purists use canned vegetables when making stew.
The fifth jar contains big juicy peaches.
The overweight charmer could slip poison into anyone's tea.

Each stag surely finds a big fawn.
The rich should invest in black zircons instead of stylish shoes.
Please sing just the club theme.
They used an aggressive policeman to flag thoughtless motorists.
Shell shock caused by shrapnel is sometimes cured through group therapy.

The advertising verse of plymouth variety store never changes.
Suburban housewives often suffer from the gab habit.
A lone star shone in the early evening sky.
The toddler found a clamshell near the camp site.
What is this large thing by the ironing board?

Thomas thinks a larger clamp solves the problem.
First add milk to the shredded cheese.
Spherical gifts are difficult to wrap.
Ralph prepared red snapper with fresh lemon sauce for dinner.
Roy ignored the spurious data points in drawing the graph.

The thick elm forest was nearly overwhelmed by Dutch Elm disease.
In developing film many toxic chemicals are used.
Is this seesaw safe.
Those thieves stole thirty jewels.
Aluminium cutlery can often be flimsy.

Those musicians harmonize marvellously.
Most young rabbits rise early every morning.
Beg that guard for one gallon of petrol.
Get a calico cat to keep the rodents away.
Tina turner is a pop singer.

That pickpocket was caught redhanded.
Mum strongly dislikes appetizers.
Her classical repertoire gained critical acclaim.
Did you eat lunch yesterday?
It's illegal to postdate a cheque.

Hispanic costumes are quite colourful.
Youngsters love corn candy as a treat.
Glucose and fructose are natural sugars found in fruit.
Few people live to be a hundred.
Tradition requires parental approval for underage marriage.

The easy going zoologist relaxed throughout the voyage.
The altruistic dowager helped many malnourished vagrants.
The haunted house was a hit due to outstanding audiovisual effects.
They all enjoy ice cream sundaes.
Jeff's toy gocart never worked.

Nonprofit organizations have frequent fundraisers.
Allow leeway here.
I honour my mum.
Rationalize all errors.
May I order a parfait after i eat dinner.

Herb's birthday frequently occurs on Thanksgiving.
Does Hindu ideology honour cows?
We apply auditory modelling to computer speech recognition.
Tornados often destroy aviaries.
Jeff thought you argued in favour of a centrifuge purchase.

Rock-and-roll music has a great rhythm.
We like blue cheese but Victor likes brie.
Please dig my potatoes up before the frost.
Rob made hungarian goulash for dinner and gooseberry pie for desert.
Trish saw hours and hours of movies this Saturday.

The speech symposium might begin on Monday.
Gremlins is yet another exciting movie by Steven Spielberg.
Growing well-kept gardens is very time consuming.
Is she going with you.
I took her word for it.

Children can consume many fruit rollups in one sitting.
People drink much water with horseradish relish.
Gwen grows green beans in her vegetable garden.
The football team coach has a watch as thin as a dime.
Seamstresses attach zips with a thimble needle and thread.

A moth zigzagged along the path through Otto's garden.
Coffee is grown on steep junglelike slopes in temperate zones.
That stinging vapour was caused by chloride vaporization.
Don't look for cheap valuables in a bank vault.
Which theatre shows Mother Goose?

Παράρτημα Β΄

Λίστα φωνημάτων

aa	ch	g	m	s	uw
ae	d	hh	n	sh	v
ah	dh	ih	ng	sil	w
ao	eh	iy	ow	sp	y
aw	er	jh	oy	t	z
ay	ey	k	p	th	zh
b	f	l	r	uh	-

Πίνακας Β΄.1: Τα συνολικά φωνήματα που χρησιμοποιήθηκαν

Παράρτημα Γ′

Αρχεία που αφαιρέθηκαν από κάθε ομιλητή

Τα ονόματα των ακουστικών και οπτικών αρχείων που αφαιρέθηκαν για κάθε ομιλητή είναι:

F1 Ομιλήτρια

071_075	131_135	171_175	236_240	326_330
086_090	136_140	196_200	246_250	331_335
096_100	151_155	201_205	251_255	421_425
121_125	166_170	231_235	316_320	436_440

Πίνακας Γ′.1: Αρχεία που αφαιρέθηκαν για την f1 ομιλήτρια (20 στο σύνολο)

F2 Ομιλήτρια

036_040	141_145	196_200	286_290	406_410
046_050	151_155	276_280	331_335	416_420
106_110	161_165	281_285	341_345	436_440

Πίνακας Γ′.2: Αρχεία που αφαιρέθηκαν για την f2 ομιλήτρια (15 στο σύνολο)

F3 Ομιλήτρια

006_010	076_080	181_185	261_265	331_335	396_400
011_015	081_085	216_220	271_275	346_350	401_405
046_050	086_090	256_260	281_285	386_390	446_450

Πίνακας Γ′.3: Αρχεία που αφαιρέθηκαν για την f3 ομιλήτρια (18 στο σύνολο)

F4 Ομιλήτρια

166_170	256_260	286_290	301_305	336_340

Πίνακας Γ′.4: Αρχεία που αφαιρέθηκαν για την f4 ομιλήτρια (5 στο σύνολο)

F5 Ομιλήτρια

031_035	116_120	261_265	311_315	386_390	426_430
036_040	256_260	271_275	326_330	416_420	456_460

Πίνακας Γ′.5: Αρχεία που αφαιρέθηκαν για την f5 ομιλήτρια (12 στο σύνολο)

M1 Ομιλητής

106_110	161_165	231_235	256_260	321_325	366_370	386_390
116_120	201_205	241_245	316_320	331_335	376_380	401_405

Πίνακας Γ′.6: Αρχεία που αφαιρέθηκαν για τον m1 ομιλητή (14 στο σύνολο)

M2 Ομιλητής

101_105	156_160	181_185	241_245	331_335	426_430
146_150	171_175	186_190	281_285	341_345	———

Πίνακας Γ′.7: Αρχεία που αφαιρέθηκαν για τον m2 ομιλητή (11 στο σύνολο)

M3 Ομιλητής

031_035	061_065	291_295	331_335	421_425
036_040	201_205	296_300	396_400	446_450

Πίνακας Γ′.8: Αρχεία που αφαιρέθηκαν για τον m3 ομιλητή (10 στο σύνολο)

M4 Ομιλητής

026_030	071_075	091_095	181_185	256_260	316_320	386_390
031_035	086_090	151_155	206_210	311_315	321_325	421_425

Πίνακας Γ′.9: Αρχεία που αφαιρέθηκαν για τον m4 ομιλητή (14 στο σύνολο)

M5 Ομιλητής

001_005	141_145	301_305	351_355	381_385
006_010	146_150	316_320	356_360	386_390
016_020	206_210	326_330	361_365	391_395
031_035	216_220	331_335	366_370	401_405
041_045	281_285	336_340	371_375	436_440
046_050	296_300	346_350	376_380	441_445

Πίνακας Γ′.10: Αρχεία που αφαιρέθηκαν για τον m5 ομιλητή (30 στο σύνολο)

Βιβλιογραφία

[1] Asterios Toutios and Shrikanth S Narayanan, "Advances in real-time magnetic resonance imaging of the vocal tract for speech science and technology research," *APSIPA Transactions on Signal and Information Processing*, vol. 5, pp. e6, 2016.

[2] JR Westbury, "X-ray microbeam speech production database user's handbook: Madison," *WI: Waisman Center, University of Wisconsin*, 1994.

[3] Lukasz Mik, Robert Wielgat, Daniel Krol, Rafal Jkedryka, Anita Lorenc, and Radoslaw Swikecinski, "Multimodal speech data acquisition with the use of ema, fast-speed video cameras and a dedicated microphone array," in *Mixed Design of Integrated Circuits and Systems, 2016 MIXDES-23rd International Conference*. IEEE, 2016, pp. 415–418.

[4] W Hardcastle, Wilf Jones, Colin Knight, Ann Trudgeon, and G Calder, "New developments in electropalatography: A state-of-the-art report," *Clinical Linguistics & Phonetics*, vol. 3, no. 1, pp. 1–38, 1989.

[5] Kate Saenko, Trevor Darrell, and James R Glass, "Articulatory features for robust visual speech recognition," in *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 2004, pp. 152–158.

[6] Kate Saenko, Karen Livescu, Michael Siracusa, Kevin Wilson, James Glass, and Trevor Darrell, "Visual speech recognition with loosely synchronized feature streams," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*. IEEE, 2005, vol. 2, pp. 1424–1431.

[7] Shuo-Yiin Chang, "Feature design for robust speech recognition: Nurture and nature," *Ph.D. thesis, Electrical Engineering & Computer Sciences, UC Berkeley*, 2016.

[8] Sigurdur Sigurdsson, Kaare Brandt Petersen, and Tue Lehn-Schiøler, "Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music," in *Seventh International Conference on Music Information Retrieval (ISMIR)*, 2006.

[9] Vibha Tiwari, "Mfcc and its applications in speaker recognition," *International journal on emerging technologies*, vol. 1, no. 1, pp. 19–22, 2010.

[10] Steve Young, Gunnar Evermann, D Kershaw, G Moore, J Odell, D Ollason, D Povey, V Valtchev, and P Woodland, "The htk-book 3.2," *Cambridge University, Cambridge, England*, 2002.

[11] Liang Lu, "Subspace gaussian mixture models for automatic speech recognition," *Ph.D. thesis, University of Edinburgh*, 2013.

[12] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[13] http://www.cs.cmu.edu/ 16385/Slides/8.2_Bag_of_Visual_Words.pdf.

[14] https://www.slideshare.net/SarahGuido/kmeans-clustering-with-scikitlearn.

[15] https://qph.ec.quoracdn.net/main-qimg-678795190794dd4c071366c06bf32115-c.

[16] http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html.

[17] https://msdn.microsoft.com/en-us/magazine/mt185575.aspx.

[18] http://horicky.blogspot.fr/2011/04/k-means-clustering-in-map-reduce.html.

[19] https://stats.stackexchange.com/questions/108617/what-do-the-variables-mean-in-the-svm-objective-function.

[20] http://upload.wikimedia.org/wikipedia/commons/1/1b/Kernel_Machine.png.

[21] https://i.stack.imgur.com/1KlNM.jpg.

[22] Paul W Schonle, Klaus Grabe, Peter Wenig, Jorg Hohne, Jorg Schrader, and Bastian Conrad, "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain and Language*, vol. 31, no. 1, pp. 26–35, 1987.

[23] Simon King, Joe Frankel, Karen Livescu, Erik McDermott, Korin Richmond, and Mirjam Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[24] Katrin Kirchhoff, Gernot A Fink, and Gerhard Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, no. 3, pp. 303–319, 2002.

[25] Shuangyu Chang, Mirjam Wester, and Steven Greenberg, "An elitist approach to automatic articulatory-acoustic feature classification for phonetic characterization of spoken language," *Speech Communication*, vol. 47, no. 3, pp. 290–311, Nov 2005.

[26] Nasir Ahmad, Sekharjit Datta, David Mulvaney, and Omar Farooq, "A comparison of visual features for audiovisual automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3939, 2008.

[27] George Papandreou, Athanassios Katsamanis, Vassilis Pitsikalis, and Petros Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 423–435, 2009.

[28] Soroush Vosoughi, "Improving automatic speech recognition through head pose driven visual grounding," in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2014, pp. 3235–3238.

[29] Zenzi M Griffin and Kathryn Bock, "What the eyes say about speaking," *Psychological science*, vol. 11, no. 4, pp. 274–279, 2000.

[30] Katrin Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments.," in *ICSLP*, 1998.

[31] Daniel PW Ellis, Rita Singh, and Sunil Sivadas, "Tandem acoustic modeling in large-vocabulary recognition," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*. IEEE, 2001, vol. 1, pp. 517–520.

[32] Hynek Hermansky, Daniel PW Ellis, and Sangita Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*. IEEE, 2000, vol. 3, pp. 1635–1638.

[33] Yajie Miao, Hao Zhang, and Florian Metze, "Towards speaker adaptive training of deep neural network acoustic models," *Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014.

[34] Samuel Thomas, Michael L Seltzer, Kenneth Church, and Hynek Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6704–6708.

[35] Dario Martin-Iglesias, J Bernal-Chaves, Carmen Pelaez-Moreno, Ascension Gallardo-Antolin, and Fernando Diaz-de Maria, "A speech recognizer based on multiclass svms with hmm-guided segmentation," in *International Conference*

on Nonlinear Analyses and Algorithms for Speech Processing. Springer, 2005, pp. 257–266.

[36] Xin He and Xian-Zhong Zhou, "Audio classification by hybrid support vector machine/hidden markov model," *World Journal of Modeling and Simulation*, vol. 1, no. 1, pp. 56–59, 2005.

[37] Raman Arora and Karen Livescu, "Multi-view cca-based acoustic features for phonetic recognition across speakers and domains," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[38] Sujeeth Bharadwaj, Raman Arora, Karen Livescu, and Mark Hasegawa-Johnson, "Multiview acoustic feature learning using articulatory measurements," in *Intl. Workshop on Stat. Machine Learning for Speech Recognition*. Citeseer, 2012.

[39] Mark Gales and Steve Young, "The application of hidden markov models in speech recognition," *Foundations and trends in signal processing*, vol. 1, no. 3, pp. 195–304, 2008.

[40] Lawrence R Rabiner and Biing-Hwang Juang, "Fundamentals of speech recognition," 1993.

[41] Reinhold Haeb-Umbach and Hermann Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*. IEEE, 1992, vol. 1, pp. 13–16.

[42] Kantilal Varichand Mardia, John T Kent, and John M Bibby, "Multivariate analysis," *Academic Press, 518 pp*, 1980.

[43] Cheng Li and Bingyu Wang, "Fisher linear discriminant analysis," 2014.

[44] Josef V Psutka, "Benefit of maximum likelihood linear transform (mllt) used at different levels of covariance matrices clustering in asr systems," in *International Conference on Text, Speech and Dialogue*. Springer, 2007, pp. 431–438.

[45] Juri Ganitkevitch, "Speaker adaptation using maximum likelihood linear regression," in *Rheinish-Westflesche Technische Hochschule Aachen, the course of Automatic Speech Recognition*. Citeseer, 2005.

[46] PC Woodland and Daniel Povey, "Large scale discriminative training for speech recognition," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[47] Daniel Povey, *Discriminative training for large vocabulary speech recognition*, Ph.D. thesis, University of Cambridge, 2005.

[48] Daniel Povey, Luksvs Burget, Mohit Agarwal, Pinar Akyazi, Kai Feng, Arnab Ghoshal, Ondvrej Glembek, Nagendra Kumar Goel, Martin Karafiat, Ariya Rastrow, et al., "Subspace gaussian mixture models for speech recognition," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4330–4333.

[49] Daniel Povey, Stephen M Chu, and Balakrishnan Varadarajan, "Universal background model based speech recognition," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4561–4564.

[50] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.

[51] Mark JF Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[52] Scott Axelrod, Vaibhava Goel, Ramesh Gopinath, Peder Olsen, and Karthik Visweswariah, "Discriminative estimation of subspace constrained gaussian mixture models for speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 172–189, 2007.

[53] Krystian Mikolajczyk and Cordelia Schmid, "A performance evaluation of local descriptors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[54] Krystian Mikolajczyk, Bastian Leibe, and Bernt Schiele, "Local features for object class recognition," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. IEEE, 2005, vol. 2, pp. 1792–1799.

[55] Bruce A. Draper Stephen o Hara, "Introduction to the bag of features paradigm for image classification and retrieval," in *Computing Research Repository*, 2011.

[56] Thorsten Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *European conference on machine learning*. Springer, 1998, pp. 137–142.

[57] Juan Ramos, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, 1997.

[58] Sam Scott and Stan Matwin, "Feature engineering for text classification," in *ICML*, 1999.

[59] Arnold WM Smeulders Uijlings, Jasper RR and Remko JH Scha, "Real-time bag of words, approximately," in *Proceedings of the ACM international Conference on Image and Video Retrieval.*, 2009.

[60] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, Jun 2008.

[61] David G Lowe, "Distinctive image and features from scale-invariant and keypoints," in *International Journal of Computer Vision*, 2004.

[62] Greg Hamerly and Charles Elkan, "Learning the k in k-means," *Advances in neural information processing systems*, vol. 16, pp. 281, 2004.

[63] Gustavo E. A. P. A. Batista and Ronaldo C. Prati, "A study of the behavior of several methods for balancing machine learning training data," in *ACM Sigkdd Explorations Newsletter*, 2004.

[64] Mahendra Sahare and Hitesh Gupta, "A review of multi-class classification for imbalanced data," in *International Journal of Advanced Computer Research*, 2012.

[65] Yuchun Tang, Yan-Qing Zhang, N.V. Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 1, pp. 281–288, Feb 2009.

[66] Nathalie Japkowicz and Shaju Stephen, "The class imbalance problem: A systematic study," *Intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[67] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer, "Smote: Synthetic and minority over-sampling and technique," in *Journal of Artificial Intelligence Research*, 2002.

[68] SailAlign: Robust and long speech-text alignment, "Appears in proc. of workshop on new tools and methods for very large scale research in phonetic sciences, jan," 2011.

[69] Raman Arora and Karen Livescu, "Multi-view cca-based acoustic features for phonetic recognition across speakers and domains," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7135–7139.