Πανεπιστήμιο Θεσσαλίας

Μεταπτυχιακή Εργασία

Οπτικοακουστική Επεξεργασία Φωνής

Σταύρος Μπουγλός

Επιβλέπων:
Αναπληρωτής Καθηγητής
Γεράσιμος Ποταμιάνος

Για το Πρόγραμμα Μεταπτυχιακών Σπουδών Επιστήμη και Τεχνολογία Υπολογιστών, Τηλεπικοινωνιών και Δικτύων

στο

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Φεβρουάριος 2014

Ευχαριστίες

Θα ήθελα πρώτα από όλους να ευχαριστήσω τον καθηγητή κ.Ποταμιάνο Γεράσιμο για την επίβλεψη αυτής της διπλωματικής, για την άριστη συνεργασία που είχαμε και για τον χρόνο που μου αφιέρωσε. Ακόμα θα ήθελα να ευχαριστήσω θερμά τον Κωνσταντίνο Κυρίτση για την συμβολή του όπου χρειάστηκε. Τέλος θα ήθελα να ευχαριστήσω την οικογένειά μου για την υποστήριξη που μου προσέφεραν και μου προσφέρουν.

Περίληψη

Τα πρώτα συστήματα αναγνώρισης ομιλίας από υπολογιστή έγιναν χρησιμοποιώντας αποκλειστικά την ακουστική πληροφορία της ανθρώπινης ομιλίας. Όμως στην ανθρώπινη ομιλία ένα μέρος της πληροφορίας προέρχεται από την κίνηση του προσώπου του ομιλητή. Έτσι, για την βελτίωση των συστημάτων αναγνώρισης ομιλίας που λειτουργούν υπό θόρυβο, είναι φυσικό να επιχειρείται και η χρήση της πληροφορίας που μπορεί να εξαχθεί από την κίνηση του προσώπου του κάθε ομιλητή.

Ο σκοπός αυτής της διπλωματικής εργασίας είναι η αξιολόγηση της χρήσης των οπτικών χαρακτηριστικών σε συνδυασμό με τα ακουστικά χαρακτηριστικά για την βελτίωση της απόδοσης ενός συστήματος αναγνώρισης ομιλίας σε συνθήκες θορύβου.

Για αυτό το σκοπό, χρησιμοποιήθηκε μια οπτικοακουστική βάση δεδομένων πολλών ομιλητών που πρόφεραν συνεχή και μεονωμένα ψηφία σε εύκολες συνθήκες. Αρχικά χρησιμοποιήθηκε μόνο ο ήχος των βίντεο χωρίς θόρυβο και επιτυγχάνθηκε αναγνώριση ομιλίας με ακρίβεια 96%. Χρησιμοποιώντας μόνο τα οπτικά χαρακτηριστικά των βίντεο, η αναγνώριση ομιλίας γινόταν με ακρίβεια 32% . Στη συνέχεια χρησιμοποιήθηκε το ίδιο σύστημα υπό συνθήκες θορύβου και έγινε χρήση οπτικών χαρακτηριστικών των βίντεο για την βελτίωση της αναγνώρισης. Τελικά , με την οπτικοακουστική αναγνώριση ομιλίας η ακρίβεια της αναγνώρισης ήταν αρκετά μεγαλύτερη από ότι επιτυγχάνθηκε με κάθε μέθοδο χωριστά.

Ακόμα,έγιναν πειράματα οπτικής ανίχνευσης ομιλίας και διερευνήθηκε η δυνατότητα χρήσης των ίδιων μεθόδων σε οπτικοακουστικές βάσεις που προσομοιώνουν πραγματικές συνθήκες ομιλίας, όπως πολλοί ομιλητές σε δελτίο ειδήσεων και ένας ομιλητής σε περιβάλλον αυτοκινήτου.

Τέλος, με τα οπτικά χαρακτηριστικά που είχαν ήδη εξαχθεί, έγιναν και πειράματα οπτικής ανίχνευσης φωνητικής δραστηριότητας.

Abstract

Traditionally automatic speech recognition systems were based on acoustic information obtained from human speech alone. When a person speaks, though part of the information comes from his face movement. Thus, in order to improve automatic speech recognition in noisy environments, a useful idea could be the use of visual information extracted from the speaker's face movement.

The objective of this thesis is the evaluation of the combined use of acoustic and visual characteristics in order to improve the performance of an automatic speech recognition system under noise conditions.

For this reason, a speaker independent audiovisual database has been used with multiple speakers uttering connected and isolated digits. Initially, only the noiseless sound files of the database were used and the achieved recognition was 96%. Using only the visual characteristics extracted from the videos, the achieved recognition was 32%. Then,audio noise was added in order to deteriorate the recognition conditions and decrease the accuracy levels so as to put in use the visual characteristics which were extracted from the videos. Using both audio and visual characteristics the achieved recognition accuracy was higher than the accuracy achieved with each method alone.

Additionally, the extracted visual characteristics, where used for Visual Activity Detection as well. Finally, using visual and audio characteristics combined, the Audio-Visual Speech Recognition accuracy levels were higher than those achieved using each type of characteristics alone.

Περιεχόμενα

Ευχαριστίες ii

Περίληψη iv

Abstract vi

Contents viii

List of Figures xi

List of Tables xiii

1 Εισαγωγή 1
- 1.1 Οπτικοακουστική αναγνώριση ομιλίας . . . 1
- 1.2 Σκοπός της εργασίας . . . 2
- 1.3 Οργάνωση και περιεχόμενο . . . 2
- 1.4 Επισκόπηση βιβλιογραφίας . . . 3

2 Η οπτικοακουστική βάση CUAVE 5
- 2.1 Εισαγωγή . . . 5
- 2.2 Περιγραφή βάσης CUAVE . . . 5
  - 2.2.1 Μεμονωμένοι ομιλητές . . . 6
  - 2.2.2 Ζεύγη ομιλητών . . . 7
- 2.3 Τεχνικά Χαρακτηριστικά . . . 8
- 2.4 Επισημείωση βάσης . . . 9
- 2.5 Ανίχνευση Φωνητικής Δραστηριότητας στην οπτικοακουστική βάση CUAVE . . . 9

3 Ακουστική Επεξεργασία Ομιλίας 10
- 3.1 Κρυφά Μαρκοβιανά Μοντέλα . . . 10
  - 3.1.1 Το εργαλειοσύνολο HTK . . . 10
- 3.2 Υλοποίηση συστήματος αναγνώρισης ομιλίας . . . 11
- 3.3 Προετοιμασία δεδομένων . . . 11
  - 3.3.1 Γραμματική και λεξιλόγιο . . . 12
- 3.4 Κωδικοποίηση δεδομένων . . . 13
  - 3.4.0.1 Χαρακτηριστικά MFCC . . . 15
  - 3.4.1 Χρονικές παράμετροι χαρακτηριστικών . . . 16
- 3.5 Ορισμός Κρυφών Μαρκοβιανών Μοντέλων . . . 17

3.5.1 Αριθμός καταστάσεων ανά λέξη . . . 19
3.6 Αρχικοποίηση παραμέτρων HMM . . . 19
3.6.1 Αρχικοποίηση με HCompV . . . 20
3.6.2 Αρχικοποίηση με HInit . . . 20
3.7 Επανεκτίμηση . . . 20
3.8 Αναγνώριση . . . 21

4 Οπτική Επεξεργασία Ομιλίας 22
4.1 Εισαγωγή . . . 22
4.1.1 Η βιβλιοθήκη OpenCV . . . 22
4.2 Διαδικασία εξαγωγής χαρακτηριστικών . . . 22
4.3 Ανίχνευση αντικειμένων με χρήση ταξινομητών Haar . . . 23
4.4 Ανίχνευση προσώπου . . . 25
4.4.1 Ανίχνευση Προσώπου με την μέθοδο Viola και Jones . . . 25
4.4.1.1 Εικόνα ολοκλήρωμα . . . 26
4.4.1.2 Χρήση αλγόριθμου AdaBoost . . . 27
4.4.1.3 Διαδοχικά συνδεδεμένοι ταξινομητές . . . 28
4.4.2 Ανίχνευση προσώπου με χρήση OpenCV . . . 29
4.5 Προετοιμασία εικόνων . . . 30
4.6 Ανίχνευση και εξαγωγή περιοχής ενδιαφέροντος . . . 32
4.7 Μετασχηματισμός DCT περιοχής ενδιαφέροντος . . . 33
4.8 Εγγραφή δεδομένων . . . 34

5 Συνένωση Χαρακτηριστικών 35
5.1 Εισαγωγή . . . 35
5.2 Συχνότητα δειγματοληψίας . . . 35
5.3 Μορφή αρχείων παραμέτρων HTK . . . 36
5.4 Εγγραφή δεδομένων . . . 37

6 Οπτική Ανίχνευση Φωνητικής Δραστηριότητας 38
6.1 Οπτική Ανίχνευση Φωνητικής Δραστηριότητας . . . 38
6.2 Οργάνωση δεδομένων . . . 39
6.3 Επεξεργασία δεδομένων . . . 40

7 Πειράματα-Αποτελέσματα 42
7.1 Επιλογή αρχείων και ονομασία ακολουθιών . . . 42
7.2 Οπτική Ανίχνευση Φωνητικής Δραστηριότητας . . . 42
7.3 Αυτόματη Αναγνώριση Ομιλίας . . . 43
7.3.1 Φωνητική Αναγνώριση Ομιλίας . . . 44
7.3.1.1 Μέθοδοι εκπαίδευσης για μεμονωμένα ψηφία . . . 44
7.3.2 Ακίνητοι ομιλητές . . . 44
7.3.3 Κινούμενοι ομιλητές . . . 46
7.3.3.1 Κινούμενοι ομιλητές σε μετωπική στάση . . . 46
7.3.3.2 Κινούμενοι ομιλητές σε στάση προφιλ . . . 47
7.3.3.3 Κινούμενοι ομιλητές σε συνεχόμενη ομιλία . . . 47
7.3.4 Οπτικοακουστική Αναγνώριση Ομιλίας . . . 50
7.3.4.1 Σύγκριση μεθόδων εκπαίδευσης . . . 51
7.3.5 Σύγκριση Ακουστικής-Οπτικοακουστικής Αναγνώρισης . . . 53

7.3.6 Διερεύνηση ως προς το πλήθος των DCT 55

8 Συμπεράσματα **57**
8.1 Συμπεράσματα 57
8.1.1 Μελλοντικές κατευθύνσεις έρευνας 58

A Κείμενα Ομιλητών **60**
A.1 Μεμονωμένοι Ομιλητές 60
A.2 Ζεύγη Ομιλητών 62

Βιβλιογραφία **64**

Κατάλογος σχημάτων

2.1 Μεμονωμένη ομιλήτρια (από [11]) 6
2.2 Ζεύγος Ομιλητών σε κακές συνθήκες οπτικής αναγνώρισης ομιλίας (από [11]) 8

3.1 Δίκτυο Λέξεων (από [11]) 13
3.2 Αρχείο παραμέτρων (από [11]) 14
3.3 Υπολογισμός MFCC (από [11]) 16
3.4 Βασική τοπολογία HMM (από 23) 17
3.5 Αρχείο μοντέλου HMM (από [11]) 18
3.6 Αποτελέσματα αναγνώρισης HTK 21

4.1 Εξαγωγή χαρακτηριστικών από την περιοχή του στόματος με σκοπό την Οπτική Αναγνώριση Ομιλίας 24
4.2 Ταξινομητές Haar (από [7]): a)Χαρακτηριστικά ακμών b) Χαρακτηριστικά γραμμών c) Χαρακτηριστικά περικλειόμενου κέντρου 25
4.3 Πίνακας Προστιθέμενου Εμβαδού 26
4.4 Διαδοχικά συνδεδεμένοι ταξινομητές (από [33]) 28
4.5 Ανίχνευση προσώπου και ματιών σε περιβάλλον δελτίου ειδήσεων (από [12]) 29
4.6 Ανίχνευση προσώπων σε περιβάλλον δελτίου ειδήσεων (από [12]) 30
4.7 Ανίχνευση προσώπου σε περιβάλλον αυτοκινήτου (από[13]) . . 31
4.8 Ανίχνευση προσώπων σε περιβάλλον της βάσης CUAVE (από [11]) 32
4.9 Κανονικοποίηση θέσης προσώπου 32
4.10 Μετασχηματισμός DCT και αντίστροφος μετασχηματισμός DCT για περιοχή ενδιαφέροντος 33

7.1 Ακίνητοι ομιλητές:Σύγκριση για εκπαίδευση υπό θόρυβο-εκπαίδευση χωρίς θόρυβο/έλεγχος με θόρυβο 45
7.2 Κινούμενοι ομιλητές σε μετωπική στάση-εκπαίδευση χωρίς θόρυβο 47
7.3 Κινούμενοι ομιλητές σε στάση προφιλ 48
7.4 Κινούμενοι Ομιλητές σε συνεχόμενη ομιλία 49
7.5 Συνένωση χαρακτηριστικών-Σύγκριση για εκπαίδευση υπό θόρυβο-εκπαίδευση χωρίς θόρυβο/έλεγχος με θόρυβο 51
7.6 Συνένωση χαρακτηριστικών-Σύγκριση για εκπαίδευση υπό θόρυβο-εκπαίδευση χωρίς θόρυβο/έλεγχος με θόρυβο 52
7.7 Σύγκριση Ακουστικής-Οπτικοακουστικής Αναγνώρισης-εκπαίδευση χωρίς θόρυβο 53

7.8 Σύγκριση Fusion με Σύγκριση Ακουστικής-Οπτικοακουστικής Αναγνώρισης 54
7.9 Διερεύνηση ως προς το πλήθος των DCT 56

Κατάλογος πινάκων

5.1 Τύποι Παραμέτρων 36
5.2 Προσδιοριστικά Παραμέτρων 37

7.1 Επισημείωση βάσης CUAVE 43
7.2 Αποτελέσματα Οπτικής Ανίχνευσης Φωνητικής Δραστηριότητας 43
7.3 Ακίνητοι ομιλητές:Σύγκριση για εκπαίδευση υπό θόρυβο-εκπαίδευση χωρίς θόρυβο/έλεγχος με θόρυβο 45
7.4 Κινούμενοι ομιλητές σε μετωπική στάση-εκπαίδευση χωρίς θόρυβο 47
7.5 Κινούμενοι ομιλητές σε στάση προφιλ 48
7.6 Κινούμενοι Ομιλητές σε συνεχόμενη ομιλία 49
7.7 Συνένωση χαρακτηριστικών-Σύγκριση για εκπαίδευση υπό θόρυβο-εκπαίδευση χωρίς θόρυβο/έλεγχος με θόρυβο 52
7.8 Σύγκριση Ακουστικής-Οπτικοακουστικής Αναγνώρισης-εκπαίδευση χωρίς θόρυβο 53
7.9 Σύγκριση Ακουστικής-Οπτικοακουστικής Αναγνώρισης 54
7.10 Διερεύνηση ως προς το πλήθος των DCT 55

Κεφάλαιο 1

Εισαγωγή

1.1 Οπτικοακουστική αναγνώριση ομιλίας

Ο βασικός τρόπος επικοινωνίας μεταξύ των ανθρώπων είναι ο λόγος. Όταν ένα άτομο ομιλεί, η κύρια πληροφορία προκύπτει από την φωνή του. Όμως η κίνηση του προσώπου και ιδιαίτερα η κίνηση των χειλιών και της ευρύτερης περιοχής του στόματος επίσης εμπεριέχει πληροφορία. Όσο οι ακουστικές συνθήκες γίνονται χειρότερες, τόσο πιο έντονα φαίνεται η χρησιμότητα της οπτικής πληροφορίας στην κατανόηση του λόγου. Η ταυτόχρονη χρήση ακουστικής και οπτικής πληροφορίας γίνεται από τους ανθρώπους χωρίς ιδιαίτερη προσπάθεια και η σημασία της οπτικής πληροφορίας όσον αφορά την καταληπτότητα του λόγου γίνεται αντιληπτή διαισθητικά από τον καθένα όταν στο περιβάλλον υπάρχει θόρυβος που καλύπτει την φωνή του ομιλητή.

Η Ψηφιακή Αναγνώριση Ομιλίας είναι ένας σχετικά νέος τομέας έρευνας όπου ερευνώνται μέθοδοι για την αναγνώριση ομιλίας από υπολογιστή. Όπως και στην επικοινωνία μεταξύ ανθρώπων, σε πραγματικές συνθήκες, πολύ συχνά υπάρχουν δυσκολίες όπως ο θόρυβος υποβάθρου και η ύπαρξη πολλών ομιλητών. Επομένως είναι λογικό να χρησιμοποιείται η επιπλέον οπτική πληροφορία που μπορεί να εξαχθεί από τα οπτικά χαρακτηριστικά του βίντεο για την βελτίωση της απόδοσης των συστημάτων αναγνώρισης ομιλίας.

1.2 Σκοπός της εργασίας

Στην παρούσα διπλωματική πραγματεύεται το θέμα της αυτόματης αναγνώρισης ομιλίας (ASR) χρησιμοποιώντας μια βάση δεδομένων με βίντεο και ήχο που έχει δημιουργηθεί σε εργαστηριακές συνθήκες.

Ο σκοπός αυτής της διπλωματικής είναι η δημιουργία ενός οπτικοακουστικού συστήματος αυτόματης αναγνώρισης ομιλίας, η αξιολόγηση της απόδοσής του και η σύγκρισή του με το σύστημα που χρησιμοποιεί χαρακτηριστικά που εξήχθησαν αποκλειστικά από το ηχητικό σήμα.

1.3 Οργάνωση και περιεχόμενο

Το περιεχόμενο των επόμενων έξι κεφαλαίων (Κεφάλαιο 2 έως Κεφάλαιο 8) έχει ως εξής:

- Στο **κεφάλαιο 2** περιγράφεται η βάση CUAVE (Clemson University Audio Visual Experiments Database) [11] που χρησιμοποιήθηκε. Υπάρχουν εικόνες με παραδείγματα από βίντεο με μεμονωμένους ομιλητές αλλά και με ζεύγη ομιλητών.
- Στο **κεφάλαιο 3** παρουσιάζεται η διαδικασία σχεδιασμού ενός συστήματος αναγνώρισης ομιλίας που χρησιμοποιεί μόνο ηχητικά χαρακτηριστικά. Περιγράφονται τα βήματα της διαδικασίας που χρειάζονται χρησιμοποιώντας το εργαλειοσύνολο HTK [21]. Γίνεται ανάλυση των προδιαγραφών που πρέπει να ικανοποιεί για να τη χρήση του με τη βάση CUAVE και δίνεται συνοπτικά η θεωρία πάνω στην οποία στηρίχθηκαν τα βήματα υλοποίησης.
- Στο **κεφάλαιο 4** περιγράφεται η διαδικασία εξαγωγής από τα βίντεο, οπτικών χαρακτηριστικών από την ευρύτερη περιοχή του στόματος των ομιλητών. Στη συνέχεια αυτά τα χαρακτηριστικά θα χρησιμοποιηθούν στο σύστημα που περιγράφεται στο κεφάλαιο 3 με σκοπό την Οπτική Αναγνώριση Ομιλίας.
- Στο **κεφάλαιο 5** περιγράφεται η διαδικασία που πρέπει να ακολουθηθεί για την χρήση των ακουστικών χαρακτηριστικών σε συνδυασμό με τα οπτικά χαρακτηριστικά.

- Στο **κεφάλαιο 6** παρουσιάζεται μια προσέγγιση στο πρόβλημα της οπτικής ανίχνευσης φωνητικής δραστηριότητας με τη χρήση οπτικών χαρακτηριστικών που εξάγονται από την περιοχή χειλιών-σαγονιού. Το ζητούμενο είναι ο προσδιορισμός ύπαρξης ή απουσίας σήματος ομιλίας
- Το **κεφάλαιο 7** ξεκινάει με την περιγραφή και τα αποτελέσματα των πειραμάτων που έγιναν πάνω στην Οπτική ανίχνευση φωνητικής δραστηριότητας. Στη συνέχεια δίνονται τα αποτελέσματα των πειραμάτων που έγιναν χρησιμοποιώντας ηχητικά χαρακτηριστικά, οπτικά χαρακτηριστικά και ο συνδυασμός των δύο. Ακόμα γίνεται σύγκριση μεταξύ της μεθόδου αναγνώρισης ομιλίας μόνο με ηχητικά χαρακτηριστικά και της μεθόδου αναγνώρισης με συνδυασμό ηχητικών και οπτικών χαρακτηριστικών.
- Τέλος, στο **κεφάλαιο 8** βρίσκονται τα συμπεράσματα που προέκυψαν από αυτή τη διπλωματική και προτείνονται μερικές μελλοντικές κατευθύνσεις έρευνας.

1.4 Επισκόπηση βιβλιογραφίας

Η αναγνώριση ομιλίας με τη χρήση Κρυφών Μαρκοβιανών Μοντέλων (Hidden Markov Models-HMMs) αρχικά άνηκε αποκλειστικά στο ερευνητικό πεδίο ακουστικών σημάτων [14], [16], [17]. Όμως, τα συστήματα που είναι βασισμένα αποκλειστικά στα ακουστικά σήματα επηρεάζονται από τον ακουστικό θόρυβο που υπάρχει στο περιβάλλον.

Καθώς οι τεχνικές όρασης υπολογιστή αναπτύσσονταν ξεκίνησε η προσπάθεια χρήσης τους για την αντιμετώπιση της πρόκλησης που εισάγει η ύπαρξη θορύβου στην αναγνώριση ομιλίας [15], [24]. Οι μέθοδοι που χρησιμοποιούνται ποικίλουν αλλά η ομοιότητά τους έγκειται στο γεγονός ότι από την κίνηση του προσώπου εξάγονται οπτικά χαρακτηριστικά για να χρησιμοποιηθούν σε συνδυασμό με τα ακουστικά χαρακτηριστικά. Οι διάφορες μέθοδοι μπορεί να περιλαμβάνουν διάβασμα χειλιών (lip reading)[8],[10], πληροφορία βάθους προσώπου (facial depth information) [3], ή εξαγωγή χαρακτηριστικών από την ευρύτερη περιοχή του προσώπου [9].

Άλλες περιοχές έρευνας σχετικές με την οπτικοακουστική επεξεργασία ομιλίας είναι η αναγνώριση συναισθημάτων από την ομιλία [4] και η οπτικοακουστική ανίχνευση συχρονίας λόγου (Audio Visual Activity Detection) [14].

Τέλος, στον τομέα της οπτικοακουστικής επεξεργασίας φωνής εξετάζονται και οι περιπτώσεις όπου ο θόρυβος υπάρχει στο βίντεο [2], αλλά στην παρούσα διπλωματική δεν εξετάζεται αυτή η περίπτωση.

Κεφάλαιο 2

Η οπτικοακουστική βάση CUAVE

2.1 Εισαγωγή

Στην οπτικοακουστική επεξεργασία ομιλίας χρησιμοποιούνται οπτικά χαρακτηριστικά σε συνδυασμό με ακουστικά χαρακτηριστικά. Ο συνδυασμός αυτός, οπτικών και ακουστικών χαρακτηριστικών, επιχειρείται κυρίως γιατί η επεξεργασία ομιλίας μπορεί να γίνει ανθεκτικότερη στο θόρυβο.

Η βάση CUAVE (Clemson University Audio Visual Experiments Database) [11] δημιουργήθηκε στο Πανεπιστήμιο Clemson με σκοπό την χρήση της σε πειράματα οπτικοακουστικής επεξεργασίας.Περιλαμβάνει μια ποικιλία από ομιλητές και σχεδιάστηκε έτσι ώστε να ικανοποιεί ένα σύνολο στόχων που θα αναφερθούν αμέσως, στην περιγραφή που ακολουθεί παρακάτω.

2.2 Περιγραφή βάσης CUAVE

Για τη δημιουργία της βάσης CUAVE, ομιλητές πρόφεραν ψηφία στα αγγλικά (από το μηδέν έως το εννιά), συνεχόμενα ή μεμονωμένα. Η βάση περιλαμβάνει συνολικά περισσότερους από 7000 λέξεις. Περιλαμβάνει μια ποικιλία από ομιλητές, έτσι ώστε να να υπάρχουν όσο το δυνατόν περισσότεροι διαφορετικοί τρόποι προφοράς , διαφορετικές φυσιογνωμίες και χαρακτηριστικά προσώπου, ακόμα και διαφορετικοί τόνοι δέρματος. Ακόμα, μερικοί ομιλητές φοράνε γυαλιά, καπέλα ή κοσμήματα και πιθανόν να έχουν γένια ή μουστάκι.

Σχήμα 2.1: Μεμονωμένη ομιλήτρια (από [11])

Το βασικό κριτήριο σχεδιασμού της βάσης CUAVE ήταν η δημιουργία μιας ευέλικτης και εύκολα διαθέσιμης βάσης που να επιτρέπει αντιπροσωπευτικά πειράματα που αφορούν στην οπτικοακουστική επεξεργασία ομιλίας.

Η βάση αποτελείται από δύο βασικά μέρη, ένα με μεμονωμένους ομιλητές και ένα με ζεύγη ομιλητών. Λεπτομέρειες περιγράφονται στις παρακάτω υποενότητες.

2.2.1 Μεμονωμένοι ομιλητές

Το πρώτο βασικό μέρος αποτελείται από 36 μεμονωμένους ομιλητές. Η επιλογή των ομιλητών δεν έχει γίνει με κάποια ιδιαίτερα στενά κριτήρια επιλογής, αλλά έχει γίνει προσπάθεια ώστε να υπάρχουν ομιλητές και των δύο φύλλων. Επιπλέον, οι ομιλητές έχουν διαφορετικούς τρόπους προφοράς, και διάφορες αποχρώσεις δέρματος καθώς επίσης και μια πληθώρα οπτικών χαρακτηριστικών όπως για παράδειγμα, γυαλιά, καπέλα και κοσμήματα.

Για το μέρος που περιλαμβάνει μεμονωμένους ομιλητές, για κάθε ομιλητή στο πλάνο περιλαμβάνεται η περιοχή των ώμων και του κεφαλιού. Όλα τα βίντεο είναι έγχρωμα και χωρίς κάποια βοηθήματα για τον διαχωρισμό χειλιών/προσώπων. Οι ομιλητές καταγράφηκαν να προφέρουν ακολουθίες μεμονωμένων και συνεχόμενων αριθμών στην αγγλική γλώσσα και με πέντε διαφορετικούς τρόπους που μπορούν να χρησιμοποιηθούν για να χωριστεί κάθε βίντεο σε πέντε σκέλη. Λεπτομέρειες αναφέρονται παρακάτω καθώς και στο παράτημα Α.

Αρχικά, στα πρώτο σκέλος προφέρουν 50 μεμονωμένα ψηφία ενώ στέκονται ακίνητοι και σε μετωπική στάση προσώπου (πόζα). Καθώς δεν χρησιμοποιήθηκε κάποιος μηχανισμός περιορισμού της κίνησης , σε όλες τις περιπτώσεις υπάρχει φυσική κίνηση σε διάφορους βαθμούς. Σε μερικές περιπτώσεις η ανεπιθύμητη κίνηση ήταν ανεπαίσθητη, ενώ σε άλλες ήταν τόσο έντονη που δεν μπορούσε να θεωρηθεί πλέον ακίνητος ο ομιλητής. Ακόμη, σε μερικές περιπτώσεις, οι ομιλητές δεν κοίταζαν απευθείας την κάμερα, αλλά πλάγια.Για αυτούς τους λόγους, χρειάστηκε να γίνει επιλογή των βίντεο που χρησιμοποιήθηκαν για την οπτική επεξεργασία ομιλίας. Σε κάποιες περιπτώσεις, προφέρουν τα ψηφία χωρίς καθόλου παύση αλλά μπορεί να αντιμετωπιστεί και σαν να αποτελείται από μεμονωμένα ψηφία, καθώς υπάρχουν τα κατάλληλα αρχεία επισημείωσης (label files). Έτσι, είναι ευκολότερες οι συνθήκες εκπαίδευσης και δοκιμών.

Στο δεύτερο σκέλος, κάθε ομιλητής κινείται στο χώρο ενώ προφέρει τα ψηφία. Έτσι, κάθε ομιλητής κινείται αριστερά-δεξιά, μπρος και πίσω και γυρίζοντας το κεφάλι, ενώ προφέρει συνολικά 30 ψηφία.

Στο τρίτο σκέλος κάθε ομιλητής παραμένει ακίνητος και γυρισμένος προφίλ προφέρει από 10 ψηφία για κάθε πλευρά.

Στα τελευταία δυο σκέλη, προφέρουν συνεχόμενα ψηφία γυρισμένοι κατά μέτωπο. Συγκεκριμένα στο τέταρτο σκέλος προφέρουν τρεις τηλεφωνικούς αριθμούς ενώ στέκονται ακίνητοι και στο πέμπτο σκέλος προφέρουν άλλους τρεις τηλεφωνικούς αριθμούς ενώ κινούνται.

2.2.2 Ζεύγη ομιλητών

Το δεύτερο βασικό μέρος της βάσης περιλαμβάνει 20 ζεύγη ομιλητών και ο βασικός στόχος ήταν η χρήση της βάσης για μεθόδους αναγνώρισης ομιλίας

Σχήμα 2.2: Ζεύγος Ομιλητών σε κακές συνθήκες οπτικής αναγνώρισης ομιλίας (από [11])

που αφορούν σε πολλούς ομιλητές. Μπορεί για παράδειγμα να χρησιμοποιηθεί για αναγνώριση ενός ομιλητή, ή για αναγνώριση ομιλίας από δύο ομιλητές που μιλούν ταυτόχρονα, ή για ανίχνευση φωνητικής δραστηριότητας.

Για κάθε ομιλητή του κάθε ζεύγους υπάρχει ξεχωριστή επισημείωση για τον ομιλητή A και τον ομιλητή και υπάρχουν τρεις ακολουθίες για κάθε άτομο. Αρχικά μιλάει ο ομιλητής A και στη συνέχεια ο ομιλητής B, ενώ στη δεύτερη ακολουθία ψηφίων γίνεται ακριβώς το αντίστροφο. Στο τρίτο μέρος μιλούν και οι δύο ομιλητές ταυτόχρονα, προφέροντας διαφορετικές ακολουθίες αριθμών.

Επίσης στα βίντεο με τα ζεύγη ομιλητών υπάρχει ανεπιθύμητη κίνηση όταν θα έπρεπε να είναι ακίνητοι οι ομιλητές και πάλι με διαφορετική ένταση του φαινομένου σε κάθε περίπτωση.

2.3 Τεχνικά Χαρακτηριστικά

Τα βίντεο καταγράφηκαν σε ανάλυση 720x480 εικονοστοιχεία (pixel) ανά πλαίσιο (frame) με το πρότυπο NTSC στα 29,97 πλαίσια ανά δευτερόλεπτο (fps). Ο φωτισμός είναι ελεγχόμενος και ίδιος για όλα τα βίντεο και έχει χρησιμοποιηθεί πράσινο φόντο. Στην βάση περιλαμβάνονται συνολικά 36 ομιλητές , από τους οποίους 17 είναι θηλυκού γένους και 16 αρσενικού γένους, καθώς και 20 ζεύγη ομιλητών.

Για κάθε ζεύγος και για κάθε ομιλητή, τα δεδομένα συμπιέστηκαν σε ξεχωριστά αρχεία MPEG-2 , 5000 kbps με ήχο 16-bit, στερεοφωνικό με συχνότητα δειγματοληψίας 44.1 kHz. Ο ήχος για κάθε βίντεο είναι διαθέσιμος και σε αρχεία ".wav" , 16-bit, μονοφωνικά.

2.4 Επισημείωση βάσης

Η βάση CUAVE είναι διαθέσιμη μαζί με λεπτομερή αρχεία επισήμανσης στα οποία κάθε λέξη έχει επισημανθεί χωριστά. Τα δεδομένα έχουν επισημειωθεί με το χέρι, σε επίπεδο millisecond και σε κάθε ομιλητή ή ζεύγος ομιλητών αντιστοιχεί ένα αρχείο ".lab" που είναι συμβατό με το HTK [34].

2.5 Ανίχνευση Φωνητικής Δραστηριότητας στην οπτικοακουστική βάση CUAVE

Στην παρούσα διπλωματική διεξήχθησαν επίσης και πειράματα για Οπτική Ανίχνευση Φωνητικής Δραστηριότητας. Σε αυτά τα πειράματα επιχειρείται η ανίχνευση ενός ατόμου που μιλάει σε συνθήκες με παραπάνω από έναν ομιλητή ή γενικά σε σχετικά δύσκολες συνθήκες, χωρίς να είναι απαραίτητο να αναγνωρίζονται και οι λέξεις που προφέρονται. Έτσι, δεν απαιτείται πληροφορία που να αφορά στο περιεχόμενο του λόγου και για αυτό δεν χρειάζεται η επισήμανση να γίνεται τόσο αυστηρά, όσο για την αναγνώριση ομιλίας. Για αυτό το λόγο, για την αξιολόγηση της μεθόδου ανίχνευσης ομιλητή χρησιμοποιήθηκε το πλαίσιο που θα περιγράψω παρακάτω.

Αφού η ανίχνευση ομιλητή αφορά στις περιπτώσεις που υπάρχουν περισσότεροι από ένας ομιλητές, στη CUAVE αφορά μόνο στο μέρος με τα 22 ζεύγη ομιλητών. Η επισημείωση αρκεί να έχει γίνει σε επίπεδο frame όπως στο [25]. Έτσι, κάθε frame έχει επισημείωση 0,1 ή 2 που υποδεικνύει αντιστοίχως αν στο συγκεκριμένο frame δεν μιλάει κανένας ομιλητής (0), ο αριστερός ομιλητής (1) ή ο ομιλητής που βρίσκεται στο δεξί μέρος του frame. Για να επισημανθεί μια σειρά από frames ως "0" έπρεπε τουλάχιστον 25 συνεχόμενα frames να αντιστοιχούν σε σιωπή. Αυτή η τιμή (25) αντιστοιχεί στο όριο μεταξύ μιας απλής διακοπής στο λόγο και της "πραγματικής" σιωπής.

Κεφάλαιο 3

Ακουστική Επεξεργασία Ομιλίας

3.1 Κρυφά Μαρκοβιανά Μοντέλα

Ένα Κρυφό Μαρκοβιανό Μοντέλο(HMM) είναι η αναπαράσταση μιας Μαρκοβιανής Διαδικασίας η οποία δεν μπορεί να είναι παρατηρήσιμη. Κάθε κατάσταση του μοντέλου χαρακτηρίζεται από δύο σύνολα πιθανοτήτων: την πιθανότητα μετάβασης και μια κατανομή πιθανότητας εξόδου είτε διακριτή είτε συνεχή που, με δεδομένη την κατάσταση καθορίζουν την δεσμευμένη πιθανότητα εκπομπής κάποιου από τα σύμβολα εξόδου ή κάποιου συνεχούς τυχαίου διανύσματος.

3.1.1 Το εργαλειοσύνολο HTK

Το HTK (Hidden Markov Model Toolkit) είναι ένα εργαλειοσύνολο που χρησιμοποιείται για την δημιουργία και τον χειρισμό Κρυφών Μαρκοβιανών Μοντέλων (Hidden Markov Model Toolkit-HTK). Χρησιμοποιείται κυρίως στην ψηφιακή αναγνώριση ομιλίας, ωστόσο έχει χρησιμοποιηθεί και σε πολυάριθμες άλλες εφαρμογές όπως σύνθεση ομιλίας,αναγνώριση χαρακτήρων κ.α.

Αποτελείται από ένα σύνολο ενοτήτων βιβλιοθηκών και εργαλείων διαθέσιμα και σε πηγαίο κώδικα C. Τα εργαλεία αυτά παρέχουν εξειδικευμένες δυνατότητες για ανάλυση λόγου, εκπαίδευση Κρυφών Μαρκοβιανών Μοντέλων, δοκιμές και ανάλυση αποτελεσμάτων. Το λογισμικό υποστηρίζει HMMs με συνεχείς Γκαουσιανές κατανομές αλλά και με διακριτές κατανομές, και

μπορεί αν χρησιμοποιηθεί για την δημιουργία σύνθετων συστημάτων Κρυφών Μαρκοβιανών Μοντέλων.

Στη συνέχεια ακολουθεί η περιγραφή της κατασκευής ενός συστήματος φωνητικής αναγνώρισης ψηφίων.

3.2 Υλοποίηση συστήματος σναγνώρισης ομιλίας

Στη βάση CUAVE οι ομιλητές προφέρουν ψηφία από το μηδέν ως το εννιά, επομένως αρκεί η δυνατότητα αναγνώρισης του συστήματος σε επίπεδο λέξεων. Η διαδικασία της υλοποίησης μπορεί να χωριστεί σε επιμέρους κομμάτια που θα αναλυθούν παρακάτω, ακολουθώντας όσο είναι δυνατόν την σειρά που αντιστοιχεί στην αλληλουχία εντολών του HTK που εκτελούνται και είναι :

1. *HCopy*-Μετατρέπει τα δεδομένα φωνής σε διανύσματα χαρακτηριστικών Mel Frequency Cepstral Coefficients (MFCC)
2. *HCompv*-Αρχικοποίηση μοντέλων 'flat start'
3. *HInit*-Αρχικοποίηση μοντέλων με HInit
4. *HRest*-Επανεκτίμηση των παραμέτρων του μοντέλου
5. *HVite*-Αλγόριθμος Viterbi για αναγνώριση
6. *HResults*-Αξιολόγηση αποτελεσμάτων αναγνώρισης

3.3 Προετοιμασία δεδομένων

Τα δεδομένα ομιλίας που χρειάζονται για την εκπαίδευση αλλά και την αναγνώριση είναι ήδη έτοιμα στην βάση CUAVE. Επίσης δίνονται τα αρχεία ".lab" όπου εκεί επισημαίνεται η αρχή και το τέλος κάθε λέξεις με ακρίβεια 100 ms. Στη συνέχεια μένει να καθοριστούν το λεξιλόγιο και η γραμματική του συστήματος αναγνώρισης.

3.3.1 Γραμματική και λεξιλόγιο

Εφόσον το σύστημα σχεδιάζεται για την αναγνώριση μόνο δέκα συνολικά λέξεων, το λεξιλόγιο είναι ιδιαίτερα μικρό και απλό. Αυτές οι δέκα λέξεις πρέπει να μπορούν να συνδυάζονται μεταξύ τους με οποιαδήποτε σειρά και για οποιοδήποτε αριθμό συνεχόμενων ψηφίων ώστε να μπορούν καλυφθούν οι προτάσεις που πρόφεραν οι ομιλητές και που φαίνονται με λεπτομέρεια στο Παράρτημα Α. Οι δυνατοί συνδυασμοί λέξεων αποτελούν την γραμματική του συστήματος αναγνώρισης και μπορεί να καθοριστεί από τον χρήστη με τη δημιουργία ενός αρχείου όπου θα είναι αποθηκευμένοι οι δυνατοί συνδυασμοί. Στο HTK, για την δημιουργία της γραμματικής του συστήματος, οι κάθετες γραμμές "|" αντιστοιχούν στο λογικό "OR", σε τετράγωνες παρενθέσεις "[]" περικλείονται τις λέξεις που μπορούν να χρησιμοποιηθούν προαιρετικά σε κάθε πρόταση, σε τριγωνικές παρενθέσεις "<>" περικλείονται οι εκφράσεις που μπορούν να χρησιμοποιηθούν μια φορά ή και περισσότερες επαναλήψεις, ενώ με παρενθέσεις "()" περικλείονται οι εκφράσεις που περιλαμβάνονται υποχρεωτικά. Επομένως για τη συγκεκριμένη περίπτωση, η γραμματική θα είναι

$ digit = one|two|three|four|five|six|seven|eight|nine|zero;
(< [sil] $ digit >)

Στην περίπτωση που θα χρειαζόταν να περιλαμβάνεται υποχρεωτικά σιωπή μετά από κάθε ψηφίο, τότε η παραπάνω έκφραση θα γινόταν":

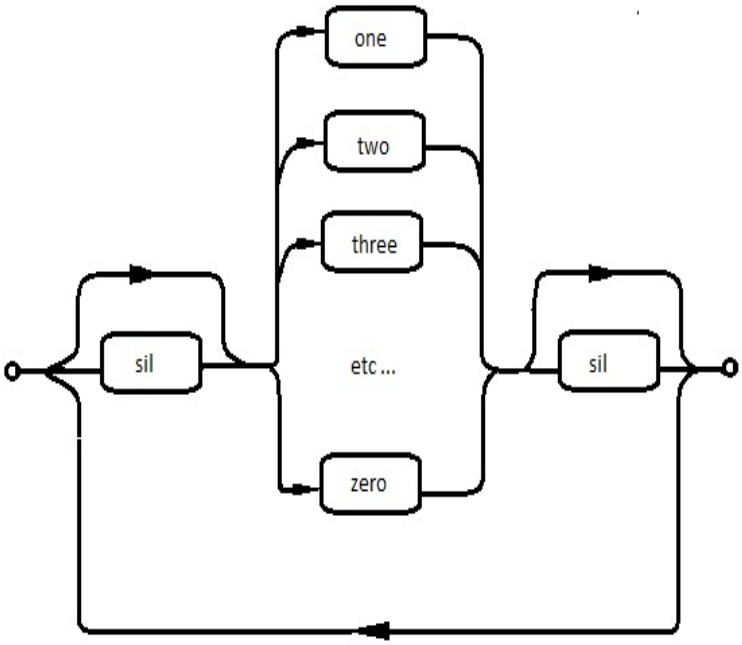
$ digit = one|two|three|four|five|six|seven|eight|nine|zero;
(< (sil) $ digit >)

Για τη συγκεκριμένη εφαρμογή αρκεί το λεξιλόγιο να είναι σε επίπεδο λέξης,και συνήθως αποθηκεύεται σε αρχείο με ονομασία dict.

Συνήθως, μετά την δημιουργία λεξιλογίου και γραμματικής, χρειάζεται να δημιουργηθεί και το δίκτυο λέξεων (word network) χρησιμοποιώντας την εντολή HParse. Όμως, στην περίπτωσή μας δίνεται με την CUAVE, επομένως δεν είναι απαραίτητο αυτό το βήμα.

Για έλεγχο ορθότητας του λεξιλογίου, αρκεί η εντολή

HSGen -l -n nsamples wdnet dict



Σχήμα 3.1: Δίκτυο Λέξεων (από [11])

όπου δημιουργεί έξοδο της μορφής
1. six sil eight sil three sil eight sil one sil five sil eight sil sil sil
2. sil seven sil five sil one sil one sil seven sil zero sil nine sil sil sil sil
3. six sil sil sil ...
κ.ο.κ.

δηλαδή προτάσεις με διάφορους πιθανούς συνδυασμούς λέξεων του λεξιλογίου (dict), χρησιμοποιώντας το δίκτυο λέξεων (wdnet) της CUAVE.

3.4 Κωδικοποίηση δεδομένων

Για την εκπαίδευση και την αναγνώριση με την χρήση του HTK δεν αρκεί τα δεδομένα να είναι σε αρχεία ήχου. Πρέπει από τις κυματομορφές των

```
SOURCEKIND = WAVEFORM
SOURCEFORMAT = WAVE
SOURCELABEL = HTK
TARGETKIND = MFCC_0_D_A
TARGETLABEL = HTK
NUMCHANS = 20
TARGETRATE = 100000
NUMCEPS = 12
WINDOWSIZE = 250000
USEHAMMING = T
PREEMCOEF = 0.97
NATURALWRITEORDER = T
```

Σχήμα 3.2: Αρχείο παραμέτρων (από [11])

αρχείων ήχου, να εξαχθούν ακολουθίες διανυσμάτων χαρακτηριστικών. Στα πλαίσια της διπλωματικής κρίθηκε καταλληλότερο η ανάλυση να γίνει με χρήση MFCC. Το εργαλείο HCopy του HTK που παίρνει σαν είσοδο τα αρχεία ήχου και εξάγει διανύσματα MFCC με βάση τις παραμέτρους που ορίζουμε στο αρχείο παραμέτρων (configuration file). Ένα τυπικό αρχείο παραμέτρων φαίνεται στο Σχήμα 3.2)

Στο παραπάνω αρχείο παραμέτρων, στις δύο πρώτες γραμμές δίνεται στο HTK πληροφορία για την μορφή του αρχείου εισόδου με τη παράμετρο NUMCEPS ορίζεται ότι από το αρχείο ήχου θα εξαχθούν 12 παράμετροι, που στη συγκεκριμένη περίπτωση, έχουν οριστεί να είναι MFCC (TARGETKIND=MFCC_0_D_A). Το πρόθεμα "_0" στο TARGETKIND σημαίνει ότι η πρώτη παράμετρος MFCC που θα υπολογιστεί είναι ανάλογη με την ολική ενέργεια του χρονικού παραθύρου . Με το "_D " (Delta) δίνεται η εντολή στο HTK να υπολογιστούν και οι παράμετροι Delta (πρώτοι χρονικοί παράγωγοι), ενώ με το "_A " (από το Acceleration ή Delta-Delta) θα εξαχθούν και οι δεύτερης τάξης χρονικοί παράγωγοι χρησιμοποιώντας τις παραμέτρους Delta. Επομένως θα εξαχθούν 39-διάστατα διανύσματα χαρακτηριστικών από κάθε παράθυρο (13 MFCC +13 Delta +13 Delta-Delta).

Στο HTK χρησιμοποιούνται χρονικές μονάδες μέτρησης 100 ns, επομένως η περίοδος δειγματοληψίας στο παραπάνω αρχείο είναι 10 ms (TARGETRATE). Ενώ με την παράμετρο WINDOWSIZE = 250000 ορίζεται ότι τα χρονικά παράθυρα στα οποία θα χωριστεί το αρχείο θα έχουν μήκος 25 ms.

Τέλος, με τις παραμέτρους USEHAMMING = T και PREEMCOEF = 0.97 ότι θα

χρησιμοποιηθεί για φιλτράρισμα παράθυρο Hamming με συντελεστή προέμφασης 0.97 ώστε να αυξηθεί η ενέργεια του σήματος στις υψηλότερες συχνότητες.

3.4.0.1 Χαρακτηριστικά MFCC

Στη συγκεκριμένη διπλωματική, για την αναγνώριση ομιλίας έχουν χρησιμοποιηθεί χαρακτηριστικά MFCC. Σχετίζονται με το real cepstrum (πραγματικό ανάφασμα) ενός παραθυροποιημένου σήματος μικρής διάρκειας που έχει προκύψει από τον FFT αυτού του σήματος.

Η ιδιαιτερότητά τους έγκειται κυρίως στο γεγονός ότι χρησιμοποιείται με μια μη-γραμμική κλίμακα συχνότητας. Αυτή η κλίμακα βασίζεται στην ανθρώπινη ακουστική ικανότητα και αυτή η ιδιότητά της είναι που την κάνει ιδιαίτερα χρήσιμη σε προβλήματα αναγνώρισης ομιλίας. Τα MFCCs υπολογίζονται με την χρήση τραπεζών φίλτρου (filterbanks). Τα filterbanks αποτελούνται από τριγωνικά φίλτρα που υπολογίζουν το φάσμα γύρω από κάθε κεντρική συχνότητα με αυξανόμενα εύρη φάσματος. Αφού καθοριστούν οι χαμηλότερες και οι υψηλότερες συχνότητες του filterbank κατανέμονται ομοιόμορφα στην κλίμακα Mel που δίνεται από τον τύπο (3.1)

$$m = 1127 \ln(1 + \frac{f}{700}) \tag{3.1}$$

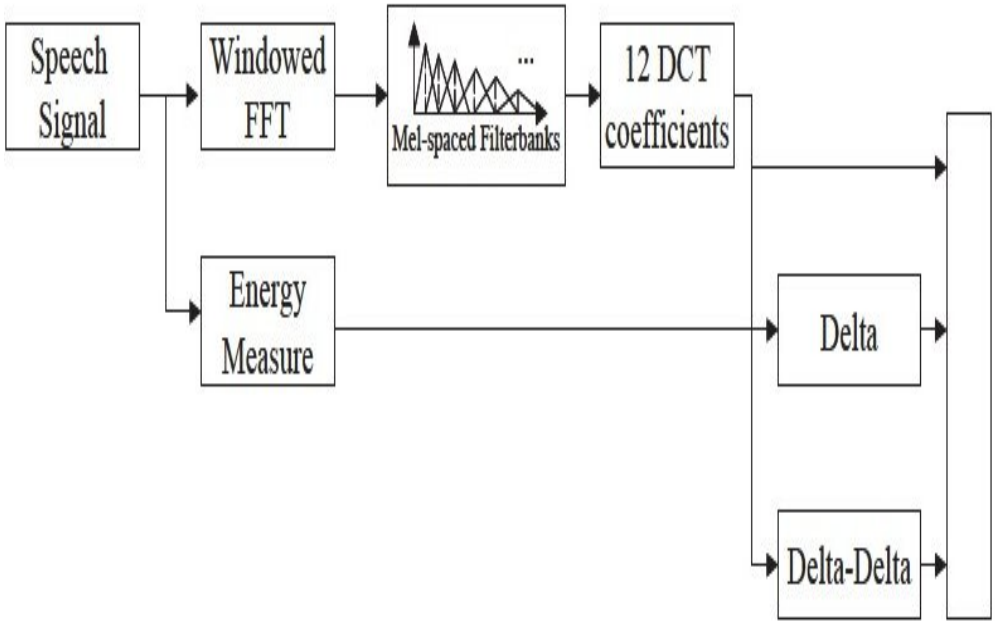
όπου στον τύπο (3.1) η συχνότητα f *hertz* σε m *mel* και στη συνέχεια υπολογίζεται ο λογάριθμος της ενέργειας της εξόδου κάθε φίλτρου.

Τα MFCCs είναι οι συντελεστές μετασχηματισμού συνημιτόνου (DCT) της ενέργειας εξόδου κάθε φίλτρου:

$$c[n] = \sum_{m=0}^{M-1} S[m] \cos(\pi n(m - 1/2)/M), 0 \leq n < M, \tag{3.2}$$

όπου S[m] είναι ο λογάριθμος της ενέργειας στην έξοδο κάθε φίλτρου και M ο αριθμός των φίλτρων.

Το πλεονέκτημα του υπολογισμού των MFCC χρησιμοποιώντας ενέργειες φίλτρων είναι η ανεκτικότητα σε θόρυβο και σε σφάλματα στον υπολογισμό του φάσματος.



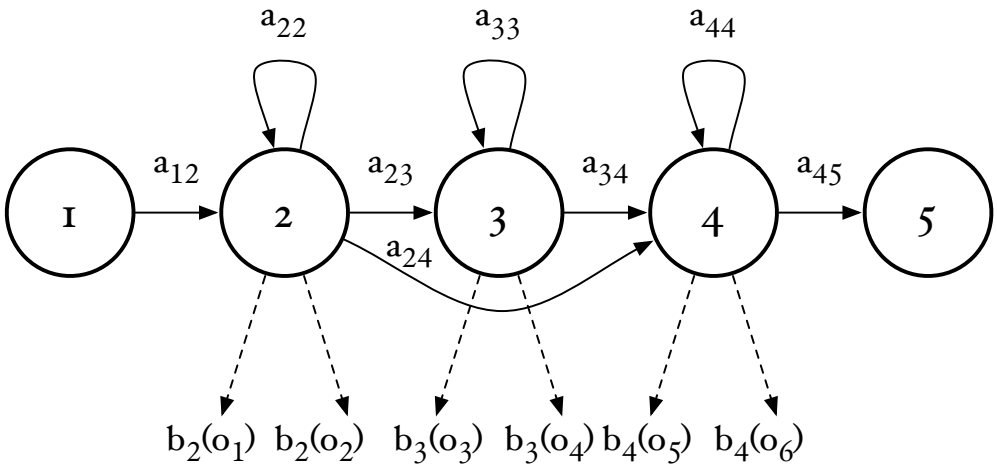
Σχήμα 3.3: Υπολογισμός MFCC (από [11])

3.4.1 Χρονικές παράμετροι χαρακτηριστικών

Προσθέτοντας παραγώγους (ως προς τον χρόνο) είναι δυνατόν να εξαχθεί πληροφορία από τις αλλαγές που γίνονται στα χαρακτηριστικά των συνεχόμενων πλαίσιων. Οι πρώτες παράγωγοι συνήθως λέγονται Δέλτα (Delta) και οι δεύτερες λέγονται Δέλτα-Δέλτα (Delta-Deltas) ή παράμετροι Επιτάχυνσης (Acceleration Parameters). Οι πρώτες παράγωγοι (Delta) υπολογίζονται από τον παρακάτω αναδρομικό τύπο (3.3) και οι δεύτερες(Delta-Delta) εφαρμόζοντας τον ίδιο τύπο στις πρώτες:

$$\Delta c[m] = \frac{\sum_{i=1}^{k} i(c[m+i] - c[[m-i])}{2\sum_{i=1}^{k} i^2}, m = 0, 1...M \tag{3.3}$$

Συνήθως στα συστήματα αναγνώρισης ομιλίας χρησιμοποιούνται διανύσματα 39 στοιχείων (13MFCC,13 Delta και 13 Delta-Delta).



Σχήμα 3.4: Βασική τοπολογία HMM (από [23])

3.5 Ορισμός Κρυφών Μαρκοβιανών Μοντέλων

Για να είναι δυνατή η ακουστική αναγνώριση των ψηφίων από το HTK πρέπει κάθε ακουστικό γεγονός (ψηφίο) να μοντελοποιηθεί καταλλήλως. Επομένως θα χρειαστούν έντεκα μοντέλα: ένα για κάθε ψηφίο από μηδέν έως εννιά και ένα για th σιωπή.

Αρχικά πρέπει να καθοριστεί η βασική τοπολογία των μοντέλων, ο αριθμών των καταστάσεων για κάθε ψηφίο και οι δυνατές μεταβάσεις μεταξύ των καταστάσεων. Όπως φαίνεται στο Σχήμα **??** στο HTK ή πρώτη και η τελευταία κατάσταση του HMM δεν εκπέμπουν διανύσματα χαρακτηριστικών. Υπάρχουν γιατί είναι απαραίτητες για εσωτερικές λειτουργίες του HTK. Οι πιθανότητες μετάβασης συμβολίζονται με a_{ij} και η εκπεμπόμενη συνάρτηση από κάθε κατάσταση, δηλαδή η πιθανότητα εκπομπής της παρατήρησης συμβολίζεται με b_i. Οι b_i είναι μίγμα Γκαουσιανών κατανομώ με διαγώνιους πίνακες.

Στο HTK κάθε HMM ορίζεται σε ένα αρχείο κειμένου (.txt) Για παράδειγμα, το μοντέλο για τα διαστήματα σιωπής αρκεί να έχει τρείς καταστάσεις και θα είναι όπως στο Σχήμα 3.5

Η περιγραφή του HMM "sil" περικλείεται από τα σύμβολα h "sil"<BeginHMM> (...) <EndHMM>. Στην τρίτη γραμμή, <VecSize> 39 <MFCC_D_A>, ορίζεται το μήκος και ο τύπος των διανυσμάτων χαρακτηριστικών, ενώ στην τέταρτη γραμμή, (<NumStates> 3) δηλώνεται ότι το συγκεκριμένο HMM έχει τρείς

```
~h "sil"
<BeginHMM>
<VecSize> 39 <MFCC_D_A>
<NumStates> 3
<State> 2
<Mean> 39
1.0  1.0  1.0...
1.0  1.0  1.0...
1.0  1.0  1.0...
 <Variance> 39
 1.0  1.0  1.0...
 <TransP>3
 1.0  1.0  1.0...
 0.0  0.5  0.5
 0.0  0.0  0.0
 <EndHMM>
```

Σχήμα 3.5: Αρχείο μοντέλου HMM (από [11])

καταστάσεις, μαζί με την πρώτη και την τελευταία που δεν εκπέμπουν διάνυσμα παρατήρησης.

Στη συνέχεια, μετά τη γραμμή <State> 2, ακολουθούν τα διανύσματα που εκπέμπονται από την κατάσταση 2. Στη παραπάνω περίπτωση οι συναρτήσεις παρατήρησης έχουν επιλεγεί να είναι Γκαουσιανές με διαγώνιους πίνακες. Έτσι για την περιγραφή τους αρκούν τα διανύσματα της μέσης τιμής και της μεταβλητότητας, δηλαδή τα διαγώνια στοιχεία του πίνακα αυτοσυσχέτισης. Η πρώτη και η τελευταία κατάσταση δεν χρειάζεται να υπάρχουν στο αρχείο αφού δεν εκπέμπουν.

Οι αρχικές τιμές δεν έχουν ιδιαίτερη σημασία καθώς θα αλλάξουν αμέσως στη συνέχεια. Για να οριστεί ο τύπος και η τοπολογία του HMM σημασία έχει το μήκος των διανυσμάτων μέσης τιμής και μεταβλητότητας καθώς και ο πίνακας μετάβασης (<TransP>).

Στο πίνακα μετάβασης με a_{ij} συμβολίζεται η μετάβαση από την κατάσταση i στην κατάσταση j. Οι μη μηδενικές τιμές του καθορίζουν τις δυνατές μεταβάσεις μεταξύ των καταστάσεων και αρχικοποιούνται σε τυχαίες τιμές, ενώ οι μηδενικές τιμές υποδηλώνουν ότι οι συγκεκριμένες μεταβάσεις δεν είναι δυνατές (αφού η πιθανότητα τους είναι μηδέν). Σε συστήματα αναγνώρισης ομιλίας τα HMM τυπικά έχουν μεταβάσεις από "αριστερά-προς-τα-δεξιά" (left-to-right) και ο πίνακας μετάβασης είναι άνω τριγωνικός.

3.5.1 Αριθμός καταστάσεων ανά λέξη

Ο αριθμός των καταστάσεων για κάθε λέξη θα μπορούσε να είναι ίδιος, αλλά για μεγαλύτερη ακρίβεια είναι καλύτερα να ποικίλει ανάλογα με το πλήθος των φωνημάτων της λέξης. Συνήθως για κάθε φώνημα της κάθε λέξης προστίθεται μια κατάσταση στο αντίστοιχο HMM. Επομένως για τον καθορισμό του πλήθους των καταστάσεων για κάθε λέξη χρειάζεται να γνωρίζουμε τον αριθμό των φωνημάτων της. Για το λεξιλόγιο της CUAVE, από το φωνητικό λεξιλόγιο BEEP [32] έχουμε:

zero z ia r ow
one w ah n
two t uw
three th r iy
four f ao/f ao r
five f ay v
six s ih k s
seven s eh v n
eight ey t
nine n ay n
sil sil

Επομένως τα zero, six και seven που έχουν από τέσσερα φωνήματα χρειάζονται 12 καταστάσεις (χωρίς την πρώτη και την τελευταία που δεν εκπέμπουν), τα one, three, four, five, και nine που έχουν από τρία φωνήματα χρειάζονται από εννέα καταστάσεις και τα υπόλοιπα δύο (two και eight) χρειάζονται από έξι καταστάσεις.

3.6 Αρχικοποίηση παραμέτρων HMM

Πριν την διαδικασία εκπαίδευσης , οι παράμετροι των HMM πρέπει να αρχικοποιηθούν έτσι ώστε ο αλγόριθμος εκπαίδευσης να συγκλίνει γρήγορα και με ακρίβεια. Για αυτό το σκοπό χρησιμοποιήθηκαν τα εργαλεία του HTK, HCompV και HInit.

3.6.1 Αρχικοποίηση με HCompV

Χρησιμοποιώντας την HCompv σε ένα σύνολο δεδομένων, υπολογίζεται η ολική μέση τιμή και συμμεταβλητότητα και στη συνέχεια, δίνει αυτές τις τιμές σε όλες τις Γκαουσιανές σε κάθε HMM. Δηλαδή, κάθε κατάσταση του HMM έχει τα ίδια διανύσματα μέσης τιμής και μεταβλητότητας που έχουν υπολογιστεί χρησιμοποιώντας όλα τα δεδομένα εκπαίδευσης. Αυτός ο τύπος αρχικοποίησης λέγεται "flat start".

3.6.2 Αρχικοποίηση με HInit

Άλλο ένα εργαλείο που υπάρχει στο HTK και μπορεί να χρησιμοποιηθεί για αρχικοποίηση είναι το HInit. Διαφέρει από το HCompV στο γεγονός ότι είναι στα δεδομένα εκπαίδευσης είναι απαραίτητο να έχουν επισημανθεί η αρχή και το τέλος κάθε λέξης. Το HInit αρχικοποιεί το HMM "ευθυγραμμίζοντας" χρονικά τα δεδομένα εκπαίδευσης με έναν αλγόριθμο Viterbi και χρησιμοποιώντας τα label files.

Η αρχή λειτουργίας της HInit στηρίζεται στην ιδέα ότι κάθε HMM είναι μια "γεννήτρια" διανυσμάτων χαρακτηριστικών λόγου. Κάθε παράδειγμα εκπαίδευσης μπορεί να περιγραφεί σαν την έξοδο του HMM του οποίου οι παράμετροι πρόκειται να υπολογιστούν. Έτσι, αν ήταν γνωστή η κατάσταση που "εξέπεμψε" κάθε διάνυσμα των δεδομένων εκπαίδευσης, τότε οι άγνωστες μέσες τιμές και μεταβλητότητες θα μπορούσαν να υπολογιστούν χρησιμοποιώντας την μέση τιμή όλων των διανυσμάτων που σχετίζονται με κάθε κατάσταση.

3.7 Επανεκτίμηση

Στη συνέχεια, γίνεται επανεκτίμηση των παραμέτρων με την χρήση HRest. Είναι μια επαναληπτική διαδικασία εκπαίδευσης, που εκτελείται για να υπολογιστούν οι βέλτιστες τιμές των παραμέτρων των HMM. Για κάθε HMM πρέπει να επαναληφθεί η ίδια διαδικασία αρκετές φορές. Σε κάθε επανάληψη το εργαλείο HRest χρησιμοποιεί την πιθανοφάνεια των δεδομένων για τον υπολογισμό της διαφοράς μεταξύ των βημάτων. Καθώς προχωράει η διαδικασία εκπαίδευσης,και γίνεται σύγκλιση με κάθε επανάληψη, η διαφορά της πιθανοφάνειας των δεδομένων μεταβάλλεται όλο και λιγότερο. Όταν μεταξύ

```
---------------------------- Overall Results--------------------------------
SENT: \%Correct=0.00 [H=0, S=3, N=3]
WORD: \%Corr=63.91, Acc=59.40 [H=85, D=35, S=13, I=6, N=133]
================================================================
```

Σχήμα 3.6: Αποτελέσματα αναγνώρισης HTK

δύο επαναλήψεων δεν μεταβάλλεται αυτό το μέτρο σύγκρισης, η διαδικασία σταματάει.

3.8 Αναγνώριση

Για την διαδικασία αναγνώρισης , αρχικά το σήμα εισόδου μετατρέπεται σε διανύσματα MFCC με την χρήση HCopy όπως έγινε προηγουμένως με τα δεδομένα εκπαίδευσης. Στη συνέχεια τα διανύσματα που προέκυψαν από το προς αναγνώριση φωνητικό σήμα, επεξεργάζονται με τον αλγόριθμο Viterbi ώστε να συγκριθεί με ποια HMM ταιριάζουν περισσότερο. Για αυτό το σκοπό καλείται η συνάρτηση Hvite, που παράγει σαν έξοδο ένα αρχείο results.mlf. Για να αξιολογηθούν τα αποτελέσματα της διαδικασία χρειάζεται η κλήση μιας τελευταίας συνάρτησης, της HResults. Η έξοδός της είναι της μορφής:

όπου N είναι ο συνολικός αριθμός των επισημειωμένων λέξεων που υπάρχουν στα αρχεία μεταγραφής που χρησιμοποιείται σαν αναφορά, S τα σφάλματα αντικατάστασης(subtitution errors), D τα σφάλματα διαγραφής(deletion errors) και I τα σφάλματα παρεμβολής(insertion errors).

Ο υπολογισμός της ακρίβειας της αναγνώρισης λέξεων γίνεται σύμφωνα με τον τύπο (3.4)

$$RecognitionAccuracy(\%) = \frac{N - D - S - I}{N} \times 100\% \tag{3.4}$$

Το ποσοστό της ακρίβειας (Accuracy(%)) που υπολογίζεται με τον τύπο (3.4) είναι αντιπροσωπευτικό της απόδοσης του συστήματος αναγνώρισης και στη συνέχεια, στα πειράματα, σε αυτό θα επικεντρώσουμε την προσοχή μας.

Κεφάλαιο 4

Οπτική Επεξεργασία Ομιλίας

4.1 Εισαγωγή

Σε αυτό το κεφάλαιο περιγράφεται η διαδικασία εξαγωγής των οπτικών χαρακτηριστικών που θα χρησιμοποιηθούν για την αναγνώριση ομιλίας.Η περιοχή ενδιαφέροντος(Region of Interest-ROI) για την εξαγωγή χαρακτηριστικών είναι η ευρύτερη περιοχή του στόματος και όχι μόνο τα χείλη και η διαδικασία περιγράφεται παρακάτω.

4.1.1 Η βιβλιοθήκη OpenCV

Για την ανίχνευση προσώπων και την εξαγωγή της ROI χρησιμοποιήθηκε η βιβλιοθήκη OpenCV [27]. Αποτελείται από μια σειρά συναρτήσεων και δομών δεδομένω με κύριο προσανατολισμό την όραση υπολογιστών. Είναι γραμμένη σε στη γλώσσα C++ και η κύρια διεπαφή της είναι σε C++, ενώ υποστηρίζεται ακόμα και η παλιότερη διεπαφή της σε C, που είναι λιγότερο εύχρηστη. Ακόμα, υάρχουν διαθέσιμες διεπαφές σε python, java και για Matlab/Octave. Τέλος, για την χρήση της υπάρχουν διαθέσιμες πηγές στο διαδίκτυο [21],[26],[27] καθώς και βιβλία [21].

4.2 Διαδικασία εξαγωγής χαρακτηριστικών

Για την εξαγωγή χαρακτηριστικών από την περιοχή του στόματος με σκοπό την Οπτική Αναγνώριση Ομιλίας χρειάζεται αρχικά να διευκρινιστεί η ύπαρξη ή όχι προσώπου στην προς εξέταση εικόνα. Αφού ανιχνευτεί πρόσωπο,

στη συνέχεια από όλη την περιοχή του προσώπου θα πρέπει να απομονωθεί το κάτω μέρος του προσώπου και εκεί να οριστεί η περιοχή ενδιαφέροντος, από όπου θα γίνει η εξαγωγή χαρακτηριστικών. Επομένως, απαιτούνται τα παρακάτω βήματα:

1. Ανίχνευση προσώπου
2. Ανίχνευση ματιών (όταν χρειάζεται)
3. Κανονικοποίηση προσώπου (όπου είναι απαραίτητο)
4. Ανίχνευση στόματος και εξαγωγή περιοχής ενδιαφέροντος(ROI)
5. Εξαγωγή DCT από την περιοχή του στόματος
6. Συμπίεση DCT

Τα βήματα 2 και 3 δεν είναι πάντα απαραίτητα όπως θα εξηγήσω παρακάτω.

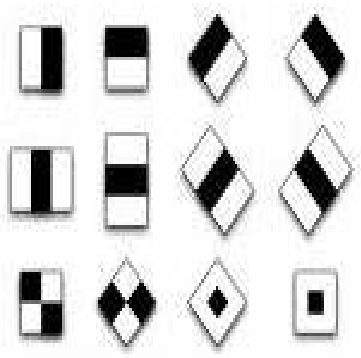
Στη συνέχεια θα περιγραφούν τα βήματα για την εξαγωγή χαρακτηριστικών, ξεκινώντας με την Ανίχνευση Αντικειμένων με ταξινομητές Haar , καθώς η ανίχνευση προσώπων είναι εφαρμογή αυτής της μεθόδου.

4.3 Ανίχνευση αντικειμένων με χρήση ταξινομητών Haar

Η Ανίχνευση Αντικειμένων με χρήση ταξινομητών Haar (Haar classifiers) προτάθηκε από τους Viola και Jones to 2001 [7]. Είναι μια αποτελεσματική και γρήγορη μέθοδος ανίχνευσης αντικειμένων βασισμένη στη μηχανική μάθηση όπου μια συνάρτηση διαδοχικής σύνδεσης (cascade function) εκπαιδεύεται από ένα πλήθος εικόνων με ύπαρξη του εν λόγω αντικειμένου(θετικά δείγματα) και εικόνων χωρίς την ύπαρξη του αντικειμένου(αρνητικά δείγματα) και στη συνέχεια χρησιμοποιείται για την ανίχνευση αντικειμένων σε άλλες εικόνες. Η μέθοδος βασίζεται στα χαρακτηριστικά τύπου Haar. Αξιοποιώντας αυτά τα χαρακτηριστικά αντί να χρησιμοποιούνται οι τιμές έντασης καθενός εικονοστοιχείου χρησιμοποιείται η αλλαγή της τιμής αντίθεσης μεταξύ γειτονικών ορθογώνιων ομάδων εικονοστοιχείων. Οι διαφορές αντίθεσης μεταξύ των ομάδων εικονοστοιχείων χρησιμοποιούνται για να προσδιοριστούν σχετικά φωτεινές και σκοτεινές περιοχές. Δύο ή τρείς γειτονικές

Σχήμα 4.1: Εξαγωγή χαρακτηριστικών από την περιοχή του στόματος με σκοπό την Οπτική Αναγνώριση Ομιλίας

ομάδες με σχετική διαφορά αντίθεσης μεταξύ τους αποτελούν ένα χαρακτηριστικό Haar. Για παράδειγμα, στο σχήμα (4.2a) υπολογίζεται ο μέσος όρος των εικονοστοιχείων που βρίσκονται μέσα στο μαύρο ορθογώνιο και ο μέσος όρος των εικονοστοιχείων που βρίσκονται στο άσπρο ορθογώνιο και στη συνέχεια οι δύο μέσοι όροι αφαιρούνται μεταξύ τους και προκύπτει η τιμή του χαρακτηριστικού Haar. Το μέγεθός των χαρακτηριστικών τύπου Haar μπορεί να μεταβάλλεται αυξομειώνοντας το μέγεθος των ομάδων και έτσι δίνεται η δυνατότητα να μπορούν να χρησιμοποιηθούν για την ανίχνευση αντικειμένων διαφόρων διαστάσεων. Haar χαρακτηριστικά όπως αυτά του σχήματος (4.2a) χρησιμοποιούνται για τον εντοπισμό ακμών(συνήθως στις άκρες των προς εντοπισμό αντικειμένων), χαρακτηριστικά όπως του σχήματος (4.2b) είναι πιο αποτελεσματικά για τον εντοπισμό γραμμών στο σώμα των αντικειμένων ενώ όπως αυτά του (4.2c) είναι καταλληλότερα για εντοπισμό τετραγώνων. Όλα τα παραπάνω μπορούν να περιστραφούν κατά 45



Σχήμα 4.2: Ταξινομητές Haar (από [7]): a)Χαρακτηριστικά ακμών b) Χαρακτηριστικά γραμμών c) Χαρακτηριστικά περικλειόμενου κέντρου

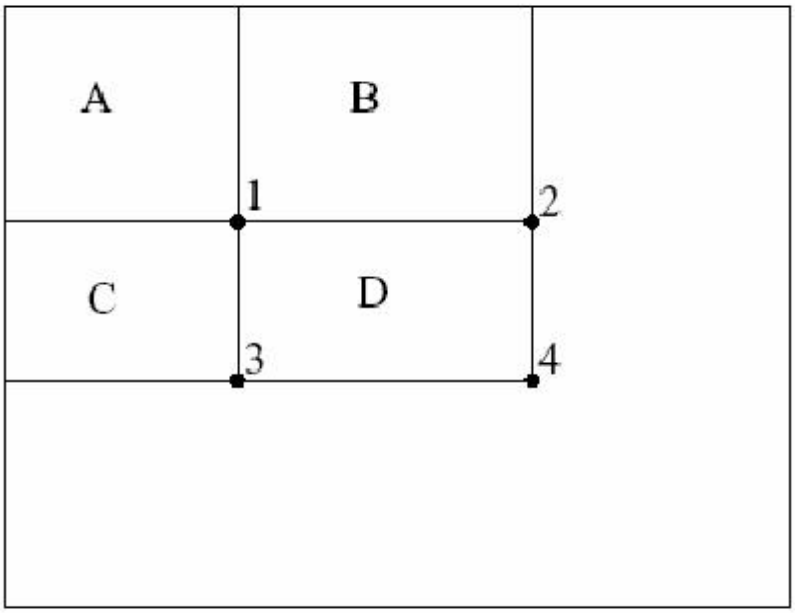
μοίρες για καλύτερη αναπαράσταση διαγώνιων σχημάτων. Έτσι αφού δεν γίνονται υπολογισμοί με τιμές έντασης καθενός εικονοστοιχείου ξεχωριστά , η μέθοδος των χαρακτηριστικών Haar έχει σημαντικά χαμηλότερο υπολογιστικό κόστος και μεγαλύτερη ταχύτητα υπολογισμού. Επομένως είναι μια χρήσιμη μέθοδος για την ανίχνευση προσώπων [6],[20].

4.4 Ανίχνευση προσώπου

Υπάρχουν διάφορες μέθοδοι για την ανίχνευση προσώπου όπου μπορεί για παράδειγμα να χρησιμοποιείται η διαφορά τόνου στο χρώμα του δέρματος, η ανίχνευση περιγραμμάτων, νευρωνικά δίκτυα, μετασχηματισμός Hough ή φίλτρα [9]. Αυτές οι μέθοδοι έχουν μεγάλο υπολογιστικό κόστος γιατί στην ουσία εκτελούνται πάνω στα εικονοστοιχεία της κάθε εικόνας για αυτό πιο αποτελεσματικά μπορεί να γίνει με την χρήση ταξινομητών Haar όπως έχει εφαρμοστεί από τους Viola & Jones [19].

4.4.1 Ανίχνευση Προσώπου με την μέθοδο Viola και Jones

Οι Viola & Jones [19] πρότειναν μια μέθοδο για ανίχνευση προσώπων που συνδυάζει χαρακτηριστικά τύπου Haar , ολοκλήρωμα εικόνας, τον αλγόριθμο AdaBoost και Διαδοχικά ταξινομημένους ταξινομητές.



Σχήμα 4.3: Πίνακας Προστιθέμενου Εμβαδού

4.4.1.1 Εικόνα ολοκλήρωμα

Για κάθε εικόνα ο αριθμός των χαρακτηριστικών τύπου Haar είναι πολύ μεγάλο και οι άθροιση όλων των τιμών των εικονοστοιχείων κάθε ορθογωνίου θα ήταν πολύ χρονοβόρα. Τα αθροίσματα μπορούν να υπολογιστούν πιο αποτελεσματικά με τη χρήση μιας βοηθητικής εικόνας , της εικόνας ολοκλήρωμα (integral image). Για την δημιουργία μιας εικόνας ολοκληρώματος χρειάζεται πρώτα ένας Πίνακας Προστιθέμενου Εμβαδού σχήμα (4.3). Ο πίνακας έχει ίδιες διαστάσεις με την εικόνα και σε κάθε σημείο (x,y) αντιστοιχεί μια τιμή που προκύπτει από το άθροισμα όλων των τιμών των εικονοστοιχείων που βρίσκονται πάνω και αριστερά από το εν λόγω σημείο.

Η τιμή του πίνακα προστιθέμενου εμβαδού στο σημείο (x,y) υπολογίζεται

$$S(x,y) - i(x,y) + s(x-1,y) + s(x,y-1) - s(x-1,y-1) \tag{4.1}$$

Και ο πίνακας μπορεί να κατασκευαστεί με ένα πέρασμα πάνω στην δοσμένη εικόνα .

Με αυτόν τον τρόπο τα αθροίσματα που χρειάζεται να υπολογιστούν για

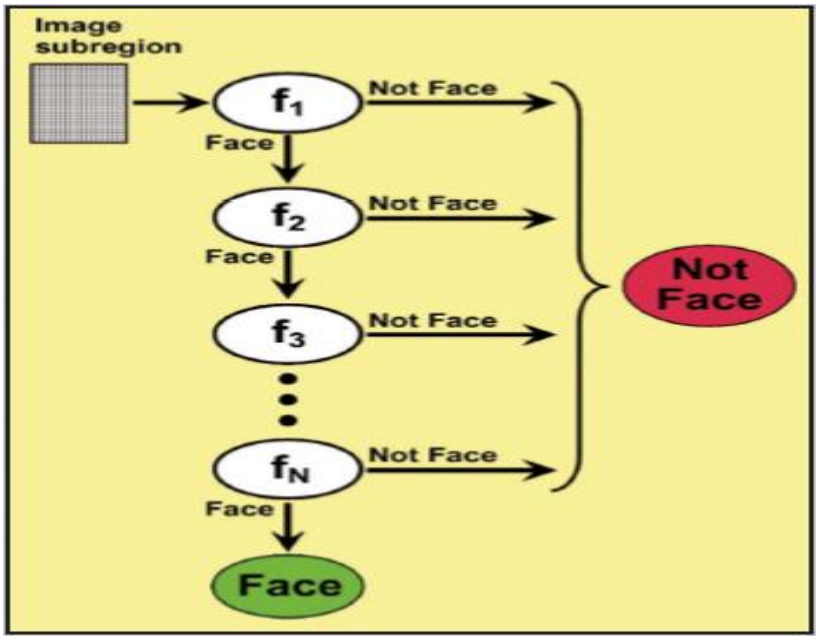
τους ταξινομητές τύπου Haar μπορούν να υπολογιστούν πιο εύκολα και γρήγορα χρησιμοποιώντας αναφορες στις τιμές του πίνακα, Έτσι για παράδειγμα στο σχήμα (4.3), το άθροισμα μέσα στο ορθογώνιο D μπορεί να υπολογιστεί αθροίζοντας τις τιμές στα σημεία 1 και 4 και αφαιρώντας τις τιμές των σημείων 2 και 3.

4.4.1.2 Χρήση αλγόριθμου AdaBoost

Χρησιμοποιώντας τα παραπάνω , τα χαρακτηριστικά μπορούν να υπολογιστούν πολύ ταχύτερα. Όμως λόγω του μεγάλου πλήθους τους ο υπολογισμός για το σύνολο των χαρακτηριστικών θα ήταν και πάλι πολύ χρονοβόρος. Έτσι δημιουργείται πλέον η ανάγκη να επιλεχθούν τα πλέον αποτελεσματικά χαρακτηριστικά ώστε να χρησιμοποιηθεί όσο το δυνατόν μικρότερος αριθμός χαρακτηριστικών στον τελικό ταξινομητή και επιπλέον τα χαρακτηριστικά θα πρέπει να είναι σημαντικά διαφοροποιημένα μεταξύ τους . Για αυτό το σκοπό χρησιμοποιήθηκε από τους Viola & Jones [7] ο αλγόριθμος AdaBoost που είναι ένας αλγόριθμος ενδυνάμωσης (boosting) και μπορεί να χρησιμοποιηθεί για την αύξηση απόδοσης οποιουδήποτε αλγόριθμου ταξινόμησης. Σε ένα μεγάλο αριθμό ταξινομητών αποδίδεται ένα μεγάλο βάρος όταν η λειτουργία ταξινόμησης είναι καλή και ένα μικρότερο βάρος όταν η λειτουργία τους δεν είναι καλή. Έτσι οι πιο αποτελεσματικοί αδύναμοι ταξινομητές (weak classifiers) συνδυάζονται για την δημιουργία του τελικού ισχυρού ταξινομητή(strong classifier) . Ένας αποτελεσματικός τρόπος αντιστοίχησης απλών ταξινομητών και χαρακτηριστικών , είναι η χρήση αδύναμων ταξινομητών που οι λειτουργίες ταξινόμησής τους εξαρτώνται από ένα μόνο χαρακτηριστικό.

Επομένως ο αλγόριθμος AdaBoost μπορεί να συνοψιστεί σε τρείς λειτουργίες:

- επιλογή των αποτελεσματικότερων χαρακτηριστικών από ένα μεγάλο σύνολο χαρακτηριστικών
- δημιουργία απλών ταξινομητών που βασίζονται σε ένα μόνο από τα επιλεγμένα χαρακτηριστικά και
- συνδυασμός των αδύναμων(απλών) ταξινομητών για την δημιουργία του ισχυρού ταξινομητή.

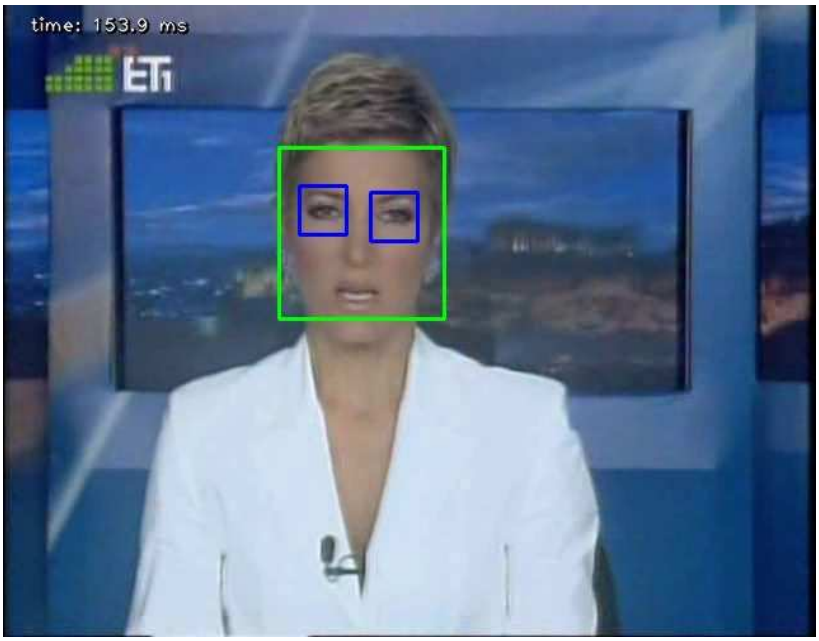


Σχήμα 4.4: Διαδοχικά συνδεδεμένοι ταξινομητές (από [33])

4.4.1.3 Διαδοχικά συνδεδεμένοι ταξινομητές

Για να αποφύγουν την χρήση μεγάλου αριθμού δειγμάτων που θα αύξανε πολύ τον χρόνο ανίχνευσης, οι Viola & Jones χρησιμοποίησαν διαδοχικούς αδύναμους ταξινομητές διαδοχικά συνδεδεμένους (cascade of clasisifiers). Κάθε εικόνα χωρίζεται σε παράθυρα ανίχνευσης όπου πάνω τους χρησιμοποιούνται οι ταξινομητές. Με την παραδοχή ότι στα περισσότερα υποπαράθυρα δεν υπάρχουν πρόσωπα, αρχικά χρησιμοποιούνται οι απλούστεροι και ταχύτεροι ταξινομητές που απορρίπτουν τα αρνητικά υποπαράθυρα ανίχνευσης. Όταν, για παράδειγμα, ο πρώτος ταξινομητής ανιχνεύσει θετικό δείγμα το δίνει ως είσοδο στον επόμενο ταξινομητή που είναι πιο πολύπλοκος και χρονοβόρος και αυτός με την σειρά του απορρίπτει τα αρνητικά δείγματα και τροφοδοτεί τον επόμενο ταξινομητή με μόνο τα θετικά δείγματα και η διαδικασία επαναλαμβάνεται με όλο και πιο πολύπλοκους και ακριβείς ταξινομητές και πάντα μόνο με τα θετικά δείγματα.

Έτσι, χρησιμοποιώντας αρχικά απλούς ταξινομητές που απορρίπτουν σε πολύ μικρό χρονικό διάστημα τα αρνητικά δείγματα πετυχαίνεται ταχύτητα ενώ χρησιμοποιώντας διαδοχικά όλο και πιο πολύπλοκους ταξινομητές αποφεύγονται οι εσφαλμένες θετικές ανιχνεύσεις (false positives).



Σχήμα 4.5: Ανίχνευση προσώπου και ματιών σε περιβάλλον δελτίου ειδήσεων (από [12])

4.4.2 Ανίχνευση προσώπου με χρήση OpenCV

Στη βιβλιοθήκη OpenCV υπάρχουν ήδη υλοποιημένες μέθοδοι για ανίχνευση αντικειμένων και προσώπων καθώς και έτοιμοι ταξινομητές για πρόσωπο, μάτια , στόμα κ.α. Οι ταξινομητές είναι σε μορφή XML και ήδη δίνονται κάποιοι μαζί με την OpenCV. Στην δική μας περίπτωση θα χρειαστούμε ταξινομητές για πρόσθια ανίχνευση και στο φάκελο της OpenCV υπάρχουν μερικοί ήδη διαθέσιμοι με την OpenCV.

Κάθε ένας δίνει ελαφρώς διαφορετικά αποτελέσματα ανάλογα με το περιβάλλον της εικόνας. Θα μπορούσαν να χρησιμοποιηθούν και όλοι μαζί για συνδυασμό αποτελεσμάτων ή μαζί με ταξινομητές για μάτια, μύτη, στόμα κλπ. Επιπλέον υπάρχουν διαθέσιμοι για κατέβασμα και άλλοι, που ίσως είναι καταλληλότεροι για τις διάφορες πιθανές συνθήκες [26].

Για την ανίχνευση, πρώτα χρειάζεται να φορτώσουμε τους κατάλληλους ταξινομητές και στη συνέχεια την εικόνα ή το βίντεο όπου θα γίνει η ανίχνευση.

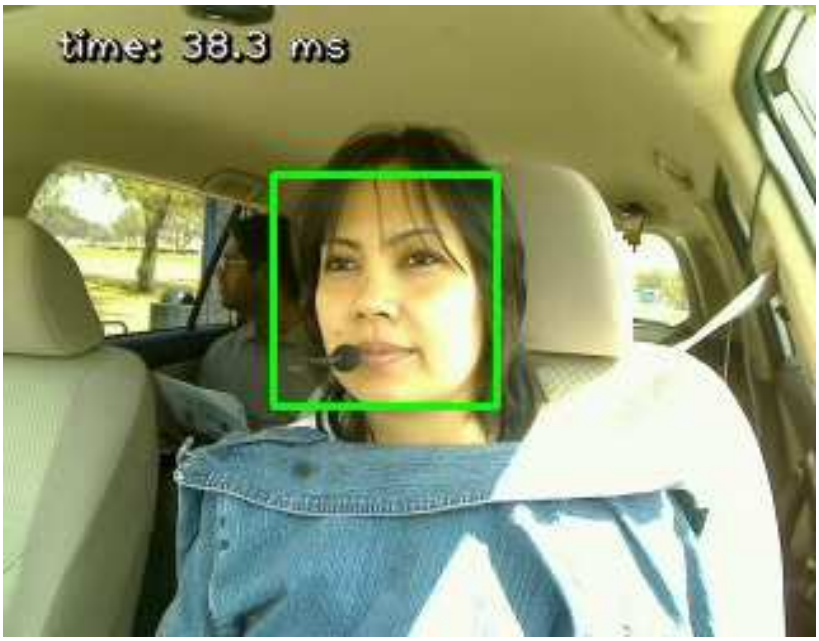
Μερικά αποτελέσματα ανίχνευσης φαίνονται παρακάτω.

Σχήμα 4.6: Ανίχνευση προσώπων σε περιβάλλον δελτίου ειδήσεων (από [12])

Σε κάθε περίπτωση χρησιμοποιήθηκαν διάφοροι ταξινομητές, και όχι μόνο αυτοί που δίνονται με την OpenCV, ώστε να επιλεχθούν οι πιο αποτελεσματικοί για τα διάφορα περιβάλλοντα. Η επιλογή των καταλληλότερων ταξινομητών έγινε με την μέθοδο δοκιμής και σφάλματος.

4.5 Προετοιμασία εικόνων

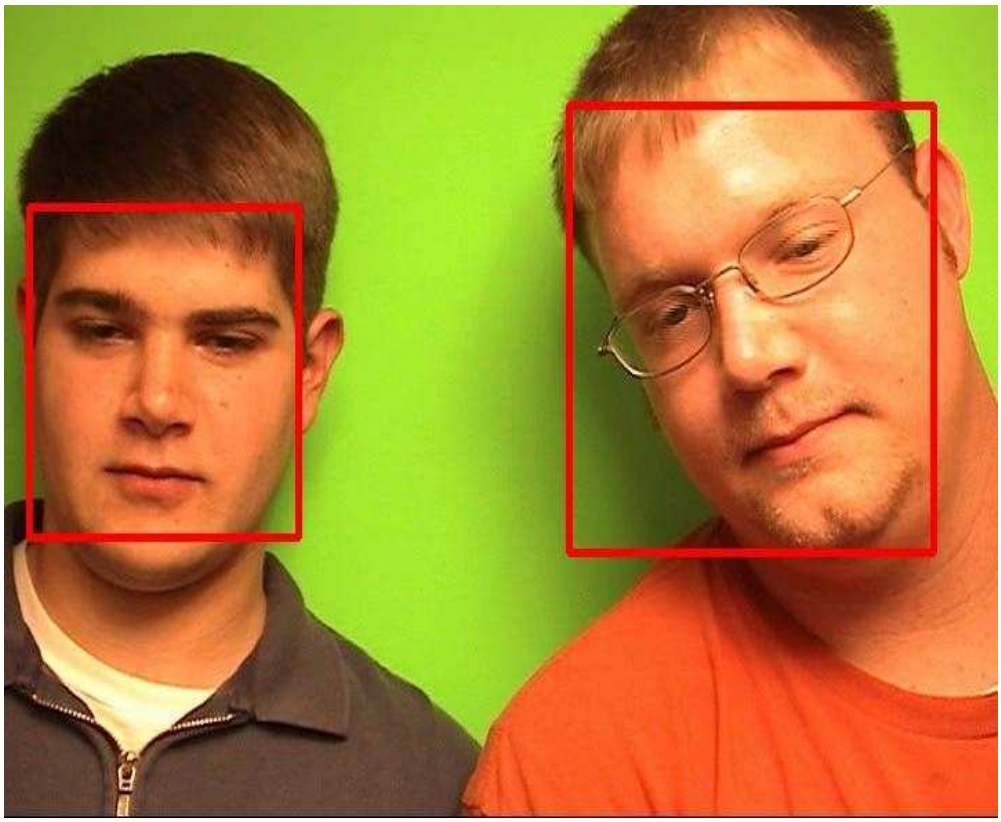
Για καλύτερα αποτελέσματα στο στάδιο της εκπαίδευσης, θα πρέπει τα ανιχνευμένα πρόσωπα να βρίσκονται σε σχετικά σταθερή θέση μέσα στην κάθε εικόνα, έτσι ώστε για παράδειγμα τα μάτια να βρίσκονται στις ίδιες συντεταγμένες, το μέγεθος του προσώπου να είναι πάντα το ίδιο , η κλίση του προσώπου σταθερή κ.ο.κ. Για αυτό το λόγο και επειδή κάποια άτομα δεν παραμένουν σταθερά καθώς προφέρουν τις ζητούμενες φράσεις, χρειάζεται να εκτελεστούν και τα βήματα 2 και 3 που αναφέρθηκαν στην αρχή και φαίνονται στο σχήμα (4.1). Όταν ο ομιλητής παραμένει σταθερός όσο προφέρει τις φράσεις, τα βήματα 2 και 3 μπορούν να παραληφθούν. Στο βήμα 2 , χρησιμοποιώντας ffmpeg [28] χωρίζουμε τα βίντεο σε frames, και στη συνέχεια με τον ίδιο τρόπο που ανιχνεύτηκαν τα πρόσωπα, βρίσκουμε τις συντεταγμένες των ματιών, χρησιμοποιώντας κατάλληλους ταξινομητές. Στη συνέχεια



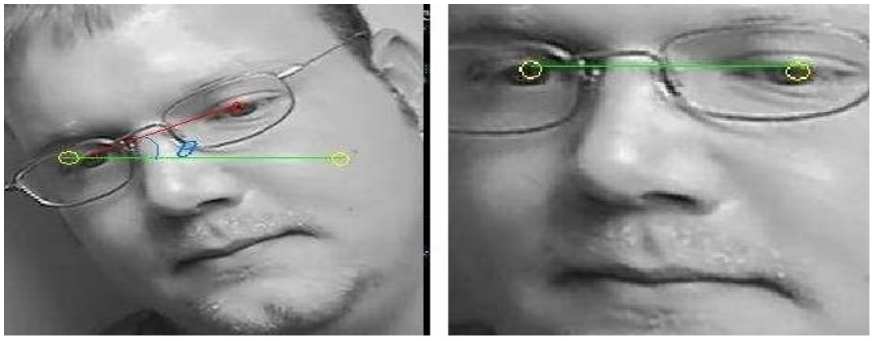
Σχήμα 4.7: Ανίχνευση προσώπου σε περιβάλλον αυτοκινήτου (από[13])

κάνουμε κανονικοποίηση των προσώπων που έχουν ανιχνευτεί, μετασχηματίζοντας την εικόνα κατάλληλα ώστε να βρίσκονται τα μάτια σε συγκεκριμένες συντεταγμένες. Εφόσον γνωρίζουμε τις κατάλληλες συντεταγμένες των ματιών, περιστρέφουμε ως προς την γωνία που έχει η ευθεία των ματιών με την οριζόντιο, και αλλάζουμε το μέγεθος της εικόνας ώστε τα μάτια να έχουν συγκεκριμένη απόσταση μεταξύ τους. Επειδή η ανίχνευση ματιών δεν ήταν πολύ αποτελεσματική και σε πολλές περιπτώσεις υπήρχε εσφαλμένη ανίχνευση , χρειαζόταν να γίνεται οπτικός έλεγχος στα frames του βίντεο και χειρωνακτική διόρθωση των συντεταγμένων. Στις περιπτώσεις που δεν γινόταν σωστά η ανίχνευση, χρησιμοποιούταν οι διορθωμένες συντεταγμένες για 50-100 frames κάθε φορά. Για αυτό το λόγο, δηλαδή επειδή αυτή η διαδικασία ήταν χρονοβόρα και όχι ιδιαίτερα αποτελεσματική όπου δεν ήταν δεν ήταν απαραίτητο παραλήφθηκαν τα βήματα 2 και 3.

Ακόμα, για την μέθοδο ανίχνευσης που περιγράφηκε παραπάνω δεν είναι απαραίτητη η πληροφορία του χρώματος, επομένως σε αυτό το βήμα μπορεί να γίνει και η μετατροπή της εικόνας σε κλίμακες του γκρι (greyscale) καθώς, όπως θα φανεί παρακάτω, είναι απαραίτητο οι εικόνες να είναι σε greyscale.



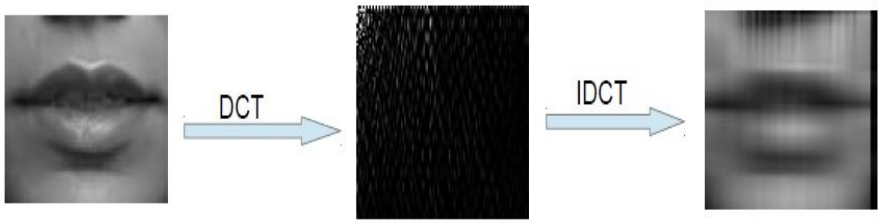
Σχήμα 4.8: Ανίχνευση προσώπων σε περιβάλλον της βάσης CUAVE (από [11])



Σχήμα 4.9: Κανονικοποίηση θέσης προσώπου

4.6 Ανίχνευση και εξαγωγή περιοχής ενδιαφέροντος

Για την Οπτική Αναγνώριση Ομιλίας πρέπει να εξαχθούν τα κατάλληλα χαρακτηριστικά από την περιοχή ενδιαφέροντος , που είναι η περιοχή του στόματος. Η εξαγωγή περιοχής ενδιαφέροντος γίνεται και πάλι με τον ίδιο τρόπο που έγινε η ανίχνευση του προσώπου και των ματιών. Στη συνέχεια θα



Σχήμα 4.10: Μετασχηματισμός DCT και αντίστροφος μετασχηματισμός DCT για περιοχή ενδιαφέροντος

πρέπει η εικόνα της περιοχής του στόματος να συμπιεστεί. Αυτό γίνεται με τον υπολογισμό του μετασχηματισμού DCT της περιοχής του στόματος. Δεν είναι απαραίτητο να αποθηκευτεί η περιοχή του στόματος σε διαφορετικά αρχεία, αλλά για να υπάρχει η δυνατότητα να ελεγχθεί η ορθότητα της επιλογής περιοχής ενδιαφέροντος, περικόπτω την περιοχή ενδιαφέροντος πριν την εφαρμογή DCT. Για αυτό το λόγο, σε αυτό το σημείο,μετατρέπεται το βίντεο σε greyscale και στη συνέχεια το χωρίζεται στα frames που το απαρτίζουν. Η μετατροπή σε grayscale είναι απαραίτητη για να μπορεί στη συνέχεια να γίνει ο μετασχηματισμός DCT. Επιλέχθηκε ο μετασχηματισμός DCT γιατί είναι ιδιαίτερα αποτελεσματικός για την συμπίεση εικόνων [18].

Η διαδικασία αυτή γίνεται αφού έχει χωριστεί το βίντεο σε frames και επαναλαμβάνοντας την διαδικασία που ακολουθήθηκε για την ανίχνευση προσώπου και την ανίχνευση ματιών , χρησιμοποιώντας αυτή τη φορά ταξινομητές εκπαιδευμένους για ανίχνευση στόματος.

4.7 Μετασχηματισμός DCT περιοχής ενδιαφέροντος

Το πλήθος των εικονοστοιχείων που αποτελούν την περιοχή του στόματος είναι τόσο μεγάλο που δεν θα μπορούσε να γίνει επιτυχής εκπαίδευση στη συνέχεια. Πρέπει με κάποιο τρόπο να μειωθούν δραστικά τα χαρακτηριστικά της εικόνας χάνοντας όσο το δυνατόν λιγότερη πληροφορία. Ένας κατάλληλος τρόπος συμπίεσης εικόνων είναι ο μετασχηματισμός DCT (Discrete Cosine Transform). Ο μετασχηματισμός DCT δεν μπορεί να εκτελεστεί σε έγχρωμες εικόνες για αυτό πρέπει να έχει προηγηθεί ο μετασχηματισμός των εικόνων σε greyscale.

Συνήθως στην επάνω αριστερή γωνία , όπου αντιστοιχούν οι χαμηλές συχνότητες συγκεντρώνονται οι συντελεστές με την μεγαλύτερη ενέργεια, και

επομένως με την σημαντικότερη πληροφορία. Επομένως κρατώντας μόνο μερικά στοιχεία από αυτή τη περιοχή μπορούμε να έχουμε αρκετή πληροφορία κρατώντας σημαντικά μικρότερο αριθμό χαρακτηριστικών.

Για παράδειγμα, κρατώντας μόνο τον επάνω αριστερά 4X4 υποπίνακα (16 στοιχεία) της Εικόνας 9 και μηδενίζοντας όλα τα υπόλοιπα στοιχεία του πίνακα που προκύπτει από τον μετασχηματισμό της αρχικής εικόνας (πχ 150X150,22500 στοιχεία) έχουμε αρκετά χαρακτηριστικά για να προχωρήσουμε στη εκπαίδευση του συστήματος.

Κάνοντας τον αντίστροφο μετασχηματισμό DCT στον παραπάνω υποπίνακα παίρνουμε την Εικόνα 10, από όπου φαίνεται πώς ο μετασχηματισμός DCT είναι μια αρκετά καλή επιλογή για συμπίεση εικόνων.

Στη συνέχεια οι τιμές DCT που εξήχθησαν από την παραπάνω διαδικασία θα χρησιμοποιηθούν για οπτική ανίχνευση φωνητικής δραστηριότητας(visual voice activity detection) και σε συνδυασμό με MFCC για Οπτικοακουστική Αναγνώριση Ομιλίας (Audio-Visual Speech Recognition).

4.8 Εγγραφή δεδομένων

Για να χρησιμοποιηθούν στο HTK τα δεδομένα που εξήχθησαν από την επεξεργασία των βίντεο, θα πρέπει να εγγραφούν σε αρχεία σε κατάλληλη μορφή ώστε να είναι δυνατή η επεξεργασία τους στο HTK. Στο HTK, στα δυαδικά αρχεία, αποθηκεύονται οι αριθμοί σε μορφή Big Endian. Όμως στο σύστημα που έγιναν τα πειράματα (Linux, Intel) αποθηκεύονται σε Little Endian. Επομένως για να είναι δυνατή η χρήση στο HTK των παραμέτρων που εξήχθησαν θα πρέπει να "δηλωθεί" στα αρχεία παραμέτρων η μορφή "Little Endian/Big Endian". Αυτή η δυνατότητα δίνεται με τις παραμέτρους NATURALREADORDER και NATURALWRITEORDER. Αν δεν δηλωθούν στα αρχεία παραμέτρων η τιμή τους είναι T (true) και τα αρχεία "Big Endian". Όμως πιο βολικό είναι σε Little Endian , επομένως στο αρχείο παραμέτρων της HCopy δηλώθηκε "NATURALWRITEORDER = F" και στα υπόλοιπα αρχεία "NATURALREADORDER = F".Για να έχουν τα διανύσματα των δειγμάτων ίδιο μήκος επιλέχθηκαν 13 DCT από κάθε frame, πάντα με φθίνουσα αλληλουχία συντελεστών.

Κεφάλαιο 5

Συνένωση Χαρακτηριστικών

5.1 Εισαγωγή

Η βασική για αυτή τη διπλωματική ήταν η χρήση οπτικών χαρακτηριστικών σε συνδυασμό με τα ηχητικά χαρακτηριστικά με σκοπό την βελτίωση της απόδοσης του συστήματος αναγνώρισης. Στο Κεφάλαιο 4 περιγράφεται η διαδικασία εξαγωγής ακουστικών χαρακτηριστικών και στο Κεφάλαιο 5. Στο κεφάλαιο αυτό περιγράφεται η διαδικασία συνένωσης χαρακτηριστικών.

5.2 Συχνότητα δειγματοληψίας

Στα πειράματα που έγιναν χρησιμοποιώντας μόνο ηχητικά χαρακτηριστικά χρησιμοποιήθηκε όπως αναφέρθηκε στο Κεφάλαιο 4 περίοδος δειγματοληψίας 10ms (TARGETRATE=100000), γιατί είναι η πιο συνηθισμένη περίοδος για τέτοιου είδους πειράματα.

Για να είναι δυνατόν να χρησιμοποιηθούν χαρακτηριστικά που εξήχθησαν από βίντεο σε συνδυασμό με χαρακτηριστικά ήχου, θα πρέπει ο αριθμός των δειγμάτων να είναι ακριβώς ίδιος.Όπως αναφέρθηκε στο Κεφάλαιο 3 στην βάση CUAVE τα βίντεο έχουν καταγραφεί στα 29.97 fps που ανιστοιχεί σε περίοδο δειγματοληψίας 33.33ms. Επομένως για να έχουν τον ίδιο αριθμό δειγμάτων θα πρέπει να γίνει γραμμική παρεμβολή(linear interpolation) τιμών στα δείγματα με το μικρότερο πλήθος και πρόσθεση (padding) ή αφαίρεση ,μερικών δειγμάτων. Για να είναι όσο το δυνατόν μικρότερος ο αριθμός των δειγμάτων που θα προστεθούν ή αφαιρεθούν χρειάστηκε να αλλάξει η

Κωδικός	Σημασία
nSamples	πλήθος δειγμάτων στο αρχείο (ακέραιος 4 byte)
sampPeriod	περίοδος δειγματοληψίας σε μονάδες 100ns (ακέραιος 4 byte)
sampSize	πλήθος byte ανα δείγμα (ακέραιος 2 byte)
parmKind	κωδικός που υποδεικνύει τον τύπο των δειγμάτων (ακέραιος 4 byte)

Πίνακας 5.1: Τύποι Παραμέτρων

Κωδικός	Τύπος	Σημασία
0	WAVEFORM	κυματομορφή
1	LPC	linear prediction filter coefficients
2	LPREFC	linear prediction reflection coefficients
3	LPCEPSTRA	LPC Cepstral coefficients
4	LPDELCEP	LPC Cepstral & delta coefficients
5	IREFC	LPC reflection coefficients 16-bit integer
6	MFCC	Mel-frequency cepstral coefficients
7	FBANK	Log mel-filter bank channel outputs
8	MELSPEC	Linear filter bank channel outputs
9	USER	user defined

περίοδος δειγματοληψίας για τα αρχεία ήχου, έτσι ώστε να διαιρεί ακριβώς την περίοδο δειγματοληψίας του βίντεο. Έτσι έγιναν πειράματα με περιόδους δειγματοληψίας 8.325 ms ,11 ms , 16 ms και 33 ms. Καλύτερα αποτελέσματα προέκυψαν για την περίοδο δειγματοληψίας 8.325 ms και έτσι, αυτή χρησιμοποιήθηκε στη συνέχεια.

5.3 Μορφή αρχείων παραμέτρων HTK

Τα αρχεία παραμέτρων HTK αποτελούνται από συνεχόμενα δείγματα με μια κεφαλίδα HTK στην αρχή του κάθε αρχείου. Κάθε δείγμα είναι ένα διάνυσμα που αποτελείται είτε από ακέραιους αριθμούς 2 byte, είτε από αριθμούς κινητής υποδιαστολής 4 byte.

Η κεφαλίδα στα αρχεία παραμέτρων HTK έχει μήκος 12 byte και περιέχει τις παρακάτω πληροφορίες:

Ο κωδικός που υποδεικνύει τον τύπο των παραμέτρων αποτελείται από έναν κωδικό 6 bit για κάθε δυνατή τιμή. Οι κωδικοί για τις βασικές παραμέτρους είναι :

επιπλέον, χρησιμοποιείται και ένας οκταδικός αριθμός που προσδιορίζει για τις παραμέτρους τα παρακάτω:

Πίνακας 5.2: Προσδιοριστικά Παραμέτρων

Πρόθεμα	Κωδικός	Σημασία
_E	000100	έχει ενέργεια
_N	000200	κατεσταλμένη ενέργεια
_D	000400	έχει παραμέτρους Δέλτα
_A	001000	έχει παραμέτρους Δέλτα-Δέλτα(επιτάχυνσης)
_C	002000	είναι συμπιεσμένο
_Z	004000	έχει μηδενική μέση τιμή σταθερής συνιστώσας
_K	010000	έχει άθροισμα ελέγχου crc
_O	020000	μηδενικό συντελεστή αναφάσματος

Το προσδιοριστικό _A έχει νόημα να χρησιμοποιηθεί μόνο όταν χρησιμοποιείται και το _D .

5.4 Εγγραφή δεδομένων

Όπως αναφέρθηκε και στο Κεφάλαιο 4, για να χρησιμοποιηθούν στο HTK τα δεδομένα που εξήχθησαν από την επεξεργασία των βίντεο, θα πρέπει να εγγραφούν σε αρχεία σε κατάλληλη μορφή ώστε να είναι δυνατή η επεξεργασία τους στο HTK. Επομένως και πάλι στο αρχείο παραμέτρων της HCopy δηλώθηκε "NATURALWRITEORDER = F".

Για να έχουν τα διανύσματα των δειγμάτων ίδιο μήκος επιλέχθηκαν από 1 εως 7 DCT από κάθε frame , πάντα με φθίνουσα αλληλουχία συντελεστών.

Επειδή οι τιμές των DCT είναι μεγαλύτερες από αυτές των MFCC επιλέχθηκε στα αρχεία να προηγούνται οι τιμές DCT και να ακολουθούν οι τιμές MFCC. Στη συνέχεια υπολογίστηκαν οι τιμές των παραμέτρων Δέλτα και Δέλτα-Δέλτα χρησιμοποιώντας τον τύπο 5.1:

$$d_t = \frac{\sum_{\theta=1}^{\theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\theta} \theta^2} \tag{5.1}$$

όπου d_t είναι ο συντελεστής Δέλτα την χρονική στιγμή *t* υπολογισμένος χρησιμοποιώντας τους αντίστοιχους στατικούς συντελεστές $c_{t+\theta}$ και $c_{t-\theta}$. Η τιμή του παραθύρου θ καθορίζεται από την παράμετρο "DELTAWINDOW". Τέλος, γράφτηκαν όλα σε ένα δυαδικό αρχείο σε κατάλληλη μορφή ώστε να είναι δυνατή η επεξεργασία τους από το HTK.

Κεφάλαιο 6

Οπτική Ανίχνευση Φωνητικής Δραστηριότητας

6.1 Οπτική Ανίχνευση Φωνητικής Δραστηριότητας

Μια χρήσιμη προσθήκη στην Ακουστική Ανίχνευση Φωνητικής Δραστηριότητας είναι η Οπτική Ανίχνευση Φωνητικής Δραστηριότητας(Visual Voice Activity Detection-VVAD). Αφορά στην ανίχνευση ομιλίας από μια βιντεοακολουθία με χρήση οπτικών σημάτων είναι πολύ χρήσιμη στις περιπτώσεις που περιλαμβάνουν πολλούς ομιλητές ή θόρυβο βάθους. Σε αυτό το κεφάλαιο παρουσιάζεται μια προσέγγιση στο πρόβλημα της VVAD με τη χρήση χαρακτηριστικών βίντεο που εξάγονται από την περιοχή χειλιών-σαγονιού. Το ζητούμενο είναι ο προσδιορισμός ύπαρξης ή απουσίας σήματος ομιλίας που παράγεται σε ένα περιβάλλον χρησιμοποιώντας οπτικά χαρακτηριστικά.

Χρησιμοποιήθηκε η βάση CUAVE καθώς περιλαμβάνει μεμονωμένους ομιλητές αλλά και ζεύγη ομιλητών. Από τις ακολουθίες βίντεο εξήχθησαν χαρακτηριστικά ομιλούντων προσώπων και στη συνέχεια εφαρμόστηκε μετασχηματισμός DCT ώστε να εξαχθούν τα δεδομένα που χρησιμοποιούνται στα μετέπειτα στάδια του συστήματος. Το στάδιο επεξεργασίας περιλαμβάνει κανονικοποίηση των δεδομένων και εφαρμογή Άθροισης Τετραγώνων Αποστάσεων (Sum of Square Distances-SSD) ώστε να γίνει μια προσέγγιση του προβλήματος από την σκοπιά επεξεργασίας εικόνας. Τα πειραματικά αποτελέσματα και η αξιολόγηση του συστήματος παρουσιάζονται στην αρχή του κεφαλαίου 7.

6.2 Οργάνωση δεδομένων

Επειδή δεν υπάρχει διαθέσιμη επισημείωση (ground truth) για μεμονωμένους ομιλητές, για το συγκεκριμένο πείραμα χρησιμοποιήθηκαν μόνο τα δεδομένων που εξήχθησαν από τα ζεύγη ομιλητών. Αρχικά, από κάθε πλαίσιο (frame) της βίντεο ακολουθίας δημιουργήθηκε ένα αρχείο κειμένου που περιλαμβάνει έναν πίνακα 16X16 με της τιμές της περιοχής του στόματος. Εφόσον τα βίντεο της CUAVE έχουν εγγραφεί στα 29.97 fps, ο αριθμός των frames που εξήχθησαν ποικίλει, από 619 εως 1305, ανάλογα και με τον χρόνο που χρειάστηκε ο εκάστοτε ομιλητής για να προφέρει την ζητούμενη φράση. Για αυτό το λόγο χρειάστηκε κάποιος μηχανισμός ώστε να σχηματιστεί ένα συνεχόμενο κείμενο που να περιλαμβάνει όλα DCT που είχαν εξαχθεί από κάθε μεμονωμένο αρχείο. Χρησιμοποιήθηκε το μετα-εργαλείο και η βιβλιοθήκη dirent.h [31] της γλώσσας C.

Το Flex [30] είναι ένας λεκτικός αναλυτής που χρησιμοποιείται κυρίως για δημιουργία μεταγλωττιστών (compilers) αλλά μπορεί να χρησιμοποιηθεί για την εξαγωγή αριθμών κινητής υποδιαστολή με χρήση κοινών εκφράσεων.

Χρησιμοποιήθηκαν κατάλληλες εκφράσεις,ώστε να αγνοηθούν τα σύμβολα παρενθέσεων που ποικίλουν από αρχείο σε αρχείο. Επιπλέον, στον κώδικα C του αρχείου γίνεται μια μετατροπή από αριθμό κινητής υποδιαστολής σε ακέραιο, ώστε να απλοποιηθεί το σκέλος της επεξεργασίας.

Η βιβλιοθήκη dirent παρείχε τα μέσα πλοήγησης μέσω του συστήματος αρχείων , δίνοντας πρόσβαση σε αρχεία και φακέλους κατά τη διάρκεια εκτέλεσης. Για να ολοκληρωθεί η διαδικασία μορφοποίησης των δεδομένων έγινε ταξινόμηση στις τιμές των DCT σε σχέση με το μέτρο τους, ώστε η συνιστώσα με την υψηλότερη ενέργεια να είναι πρώτη και να ακολουθούν οι αμέσως επόμενες. Ωστόσο, στις διάφορες ακολουθίες πλαισίων υπήρχαν διαφορετικές κατανομές των τιμών DCT, πράγμα που σημαίνει ότι έπρεπε να επιλεχθεί μια ακολουθία που να ταιριάζει στο μεγαλύτερο μέρος του σώματος των δεδομένων πλαισίων. Συνολικά επιλέχθηκαν 16 συντελεστές DCT από κάθε frame και ταξινομήθηκαν με φθίνουσα σειρά.

6.3 Επεξεργασία δεδομένων

Στη Θεωρία Επεξεργασίας Εικόνας, μια βασική προσέγγιση για να καθορίσουμε αν ένα υποκείμενο σε μια βιντεοακολουθία κινείται ή όχι, είναι αφαιρώντας τις τιμές RGB των pixel της περιοχής ενδιαφέροντος (Region Of Interest-ROI) και εφαρμόζοντας φίλτρο που βασίζεται στον θόρυβο της εικόνας και τη στιβαρότητα (robustness). Υποθέτοντας βεβαίως ότι και τα δύο frames θα έχουν τις ίδιες ROI στην ίδια θέση, υπολογίζεται η διαφορά στις τιμές του ίδιου pixel μεταξύ δύο διαδοχικών frames και αν ξεπερνάει ένα προκαθορισμένο κατώφλι ανιχνεύεται κίνηση.

Το σύστημα που σχεδιάστηκε ακολουθεί την ίδια αρχή, αλλά αντί για τιμές των pixel, το διάνυσμα εισόδου αποτελείται από ταξινομημένες τιμές DCT που δημιουργήθηκαν κατά τη διαδικασία οργάνωσης δεδομένων. Εφόσον η συνιστώσα μηδενικής τάξης θα κυριαρχεί των υπολοίπων κατά τη διαδικασία της άθροισης τετραγώνων αποστάσεων, έγινε κανονικοποίηση με βάση την τυπική απόκλιση. Αυτό απαιτεί τον υπολογισμό της μέσης τιμής όλων των συνιστωσών μηδενικής τάξης του κάθε frame και σε κάθε ακολουθία. Το επόμενο βήμα είναι η αφαίρεση της μέσης τιμής από όλες τις τιμές DCT και ο υπολογισμός των τετραγώνων τους. Τέλος, υπολογίζουμε την τυπική απόκλιση και κανονικοποιούμε τις τιμές με τον ακόλουθο τύπο:

$$x' = \frac{x - \bar{x}}{\sigma} \tag{6.1}$$

Στο επόμενο βήμα δημιουργούμε ένα μέτρο ομοιότητας (εξίσωση 6.1) που στηρίζεται στη διαφορά τετραγώνων των τιμών των συντελεστών DCT , μεταξύ δύο διαδοχικών frames. Εφόσον οι συντελεστές είναι ταξινομημένοι και έχοντας κάνει την υπόθεση ότι μεταξύ δύο διαδοχικών frames δεν αλλάζει θέση η περιοχή ενδιαφέροντος, υπολογίζεται η διαφορά της τιμής του ίδιου συντελεστή DCT για δύο διαδοχικά frames.

$$\sum (x_i - x_{i+1})^2 \tag{6.2}$$

Δηλαδή, αθροίστηκαν διαφορές τετραγώνων μεταξύ των αντιστοιχισμένων συντελεστών DCT, για διαδοχικά frames και για το σύνολο των 16 DCT σε κάθε frame. Στη συνέχεια εφαρμόστηκε ένα κατώφλι σε αυτό στο άθροισμα της διαφοράς των τετραγώνων και όταν η τιμή του αθροίσματος ξεπερνούσε

το κατώφλι, ανιχνευόταν κίνηση. Στη συνέχεια γινόταν σύγκριση των αποτελεσμάτων ανίχνευσης με τον διαθέσιμο πίνακα επισημείωσης (από [25]) Η μεθοδολογία που περιγράψαμε στο σκέλος της επεξεργασίας δεδομένων επαναλήφθηκε για κάθε ακολουθία για την οποία υπήρχε διαθέσιμος πίνακας επισημείωσης για την βάση CUAVE.

Κεφάλαιο 7

Πειράματα-Αποτελέσματα

7.1 Επιλογή αρχείων και ονομασία ακολουθιών

Για τα πειράματα όπου χρειάστηκε να αξιοποιηθεί και οπτική πληροφορία, δεν ήταν όλα τα βίντεο κατάλληλα για τους λόγους που ανέφερα προηγουμένως στην παράγραφο με την περιγραφή της βάσης. Για αυτό το λόγο, αξιοποιήθηκαν περίπου τα μισά. Για τα πειράματα που έγιναν χρησιμοποιώντας μόνο τα ηχητικά αρχεία, χρησιμοποιήθηκαν όλα τα διαθέσιμα αρχεία με μεμονωμένους ομιλητές. Τα 30 από 36 αρχεία χρησιμοποιήθηκαν για εκπαίδευση και τα υπόλοιπα για έλεγχο (testing).

Για τα διάφορα πειράματα που διεξήχθησαν, τα κομμάτια σε κάθε βίντεο ονομάστηκαν R1 έως R5. Τα κείμενα που διάβαζαν οι ομιλητές καθώς και η αντιστοιχία των R1 έως R5 φαίνεται με λεπτομέρεια στο Παράρτημα Α.

7.2 Οπτική Ανίχνευση Φωνητικής Δραστηριότητας

Η επισημείωση της CUAVE για ζεύγη ομιλητών καλύπτει τέσσερις περιπτώσεις φωνητικής δραστηριότητας ([25])

Για την δική μας περίπτωση υλοποιήθηκε μια απλούστερη εφαρμογή ελέγχου για την αξιολόγηση κάθε ακολουθίας. Αφού για κάθε ομιλητή γίνεται επεξεργασία ξεχωριστά, η ύπαρξη φωνητικής δραστηριότητας από τον ομιλητή Α θεωρείται σωστή αν η επισημείωση είναι ‘1’ ή ‘3’. Αντιστοίχως η απουσία φωνητικής δραστηριότητας (σιωπή) από τον 'Α' θεωρείται σωστή

Πίνακας 7.1: Επισημείωση βάσης CUAVE

Επισημείωση	Συμβολισμός
1	Ομιλία από A& Σιωπή από B
2	Ομιλία από B& Σιωπή από A
3	Ομιλία από A& Ομιλία από B
4	Σιωπή από A& Σιωπή από B

Πίνακας 7.2: Αποτελέσματα Οπτικής Ανίχνευσης Φωνητικής Δραστηριότητας

Αρχείο Βίντεο	Ομιλητής A (ακρίβεια ανίχνευσης)	Ομιλητής B (ακρίβεια ανίχνευσης)
g01	49.50%	57.61%
g03	56.76%	58.67%
g08	59.21%	52.91%
g09	56.44%	56.62%
g10	57.45%	57.65%
g13	55.40%	59.79%
g19	54.68%	52.91%
g20	57.08%	56.62%
g22	59.79%	58.79%

όταν έχουμε label ‘2’ ή ‘4’. Στον παρακάτω πίνακα φαίνονται τα ποσοστά επιτυχίας για τα βίντεο ελέγχου

Τα αποτελέσματα του πίνακα 7.2 δεν είναι ικανοποιητικά, καθώς από ότι φαίνεται η απόδοση του σχεδιασμένου συστήματος εξαρτάται σε πολύ μεγάλο βαθμό από τη συμπεριφορά των ομιλητών. Στα βίντεο της CUAVE με παραπάνω από έναν ομιλητή παρατηρήθηκαν πολλά ‘false positives’. Η κίνηση των χειλιών χωρίς ομιλία, η κίνηση του προσώπου και του σώματος γενικά κατατάσσονται σε αυτή τη κατηγορία. Για αυτό το λόγο , το παραπάνω σύστημα κρίνεται ακατάλληλο για πραγματικές συνθήκες , αφού απαιτεί μια βάση δεδομένων καταγεγραμμένη με αυστηρότερα κριτήρια, όπου κάθε μη ενεργός ομιλητής θα ελαχιστοποιεί την κίνησή του, κρατώντας τον αριθμό των εσφαλμένων θετικών ανιχνεύσεων (false positive) χαμηλό.

7.3 Αυτόματη Αναγνώριση Ομιλίας

Στα προηγούμενα κεφάλαια περιγράφηκε η διαδικασία που ακολουθήθηκε για την διεξαγωγή πειραμάτων πάνω στην Αυτόματη Αναγνώριση Ομιλίας με την σειρά

- Φωνητική Αναγνώριση Ομιλίας
- Οπτική Αναγνώριση Ομιλίας
- Συνένωση χαρακτηριστικών για αυτόματη αναγνώριση ομιλίας

Για τα διάφορα πειράματα που διεξήχθησαν, τα κομμάτια σε κάθε βίντεο και τα αντίστοιχα τμήματα των αρχείων ήχου ονομάστηκαν R1 εως R5. Τα κείμενα που διάβαζαν οι ομιλητές καθώς και η αντιστοιχία των R1 εως R5 φαίνεται με λεπτομέρεια στο Παράρτημα Α .

Παρακάτω παρουσιάζονται τα αποτελέσματα των πειραμάτων που διεξήχθησαν.

7.3.1 Φωνητική Αναγνώριση Ομιλίας

Αρχικά, προστέθηκε θόρυβος από συγκεχυμένη ομιλία πολλών ατόμων (speech babble) σε διάφορες εντάσεις.

Στη συνέχεια από τα αρχικά αρχεία αλλά και από τα αρχεία που προέκυψαν μετά από την πρόσθεση θορύβου, εξήχθησαν διανύσματα χαρακτηριστικών, και συγκεκριμένα 13 MFCC_D_A ,και έγιναν τα ακόλουθα πειράματα. Το σύστημα που χρησιμοποιήθηκε είναι αυτό που περιγράφτηκε στο κεφάλαιο 3. Έχουν χρησιμοποιηθεί όλα τα αρχεία από τα οποία τα έξι για έλεγχο και τα υπόλοιπα τριάντα για εκπαίδευση

7.3.1.1 Μέθοδοι εκπαίδευσης για μεμονωμένα ψηφία

7.3.2 Ακίνητοι ομιλητές

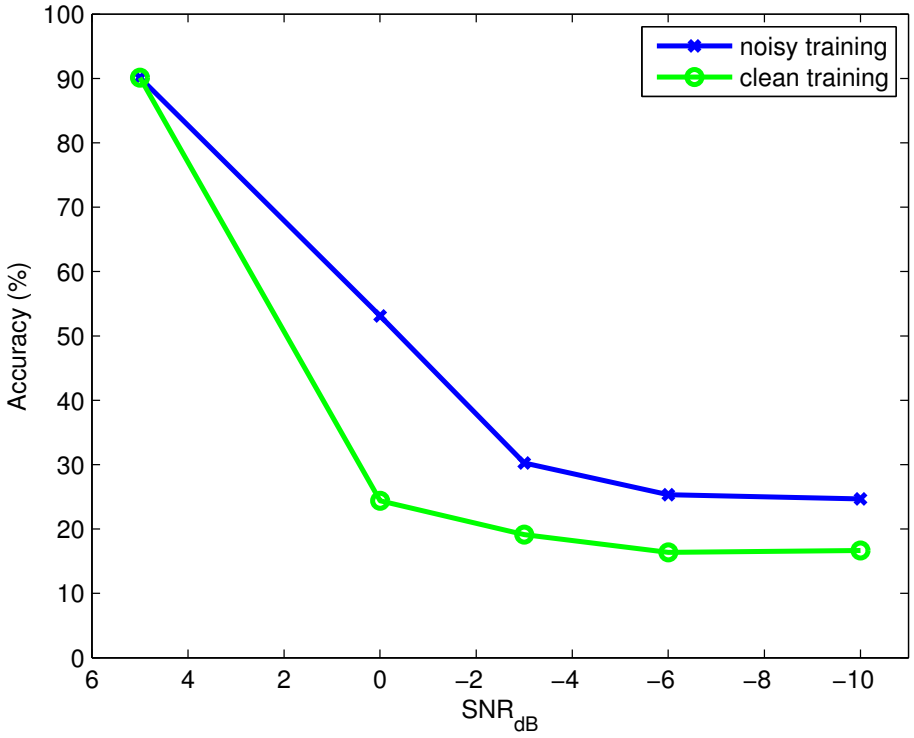
Για το τμήμα R1 των αρχείων εξετάστηκαν δύο διαφορετικοί τρόποι εκπαίδευσης :

- Αρχεία εκπαίδευσης με θόρυβο
- Αρχεία εκπαίδευσης χωρίς θόρυβο

Σε όλα τα παρακάτω πειράματα τα αρχεία ελέγχου είχαν και θόρυβο (εκτός από όταν η ένταση του θορύβου επισημαίνεται ότι είναι ίση με μηδέν). Στα

Πίνακας 7.3: Ακίνητοι ομιλητές:Σύγκριση για εκπαίδευση υπό θόρυβο-εκπαίδευση χωρίς θόρυβο/έλεγχος με θόρυβο

SNR	υπό θόρυβο(Accuracy%)	χωρίς θόρυβο(Accuracy%)
5	90.12	90.12
0	53.09	24.38
-3	30.25	19.14
-6	25.31	16.36
-10	24.69	16.67



Σχήμα 7.1: Ακίνητοι ομιλητές:Σύγκριση εκπαίδευση υπό θόρυβο-εκπαίδευση χωρίς θόρυβο/έλεγχος με θόρυβο

σχήματα όταν έχουν χρησιμοποιηθεί αρχεία εκπαίδευσης με θόρυβο υπάρχει η επισήμανση Noisy training ενώ όταν έχουν χρησιμοποιηθεί αρχεία εκπαίδευσης χωρίς θόρυβο επισημαίνεται Clean training. Σε όποιο πίνακα ή σχήμα δεν επισημαίνεται τίποτα από τα δύο έχει γίνει εκπαίδευση με αρχεία χωρίς θόρυβο. Ο έλεγχος γίνεται πάντα με θόρυβο.

Παρατηρείται ότι τα αποτελέσματα στη περίπτωση που έχει γίνει εκπαίδευση χρησιμοποιώντας αρχεία με θόρυβο τα αποτελέσματα είναι καλύτερα. Από ότι φαίνεται το σύστημα αναγνώρισης είναι πιο στιβαρό χρησιμοποιώντας αυτόν τον τρόπο εκπαίδευσης.

Όταν η ένταση του θορύβου αυξάνεται πάνω από τρεις φορές σε σύγκριση με την ένταση της ομιλίας υπάρχει κορεσμός. Για μεγαλύτερες εντάσεις θορύβου φαίνεται η ακρίβεια της αναγνώρισης να βελτιώνεται, αλλά τα αποτελέσματα αυτά είναι παραπλανητικά. Εξετάζοντας με περισσότερη λεπτομέρεια τα αποτελέσματα του HTK προκύπτει ότι για μεγάλες εντάσεις θορύβου δεν γίνεται σωστά η εκπαίδευση του συστήματος.

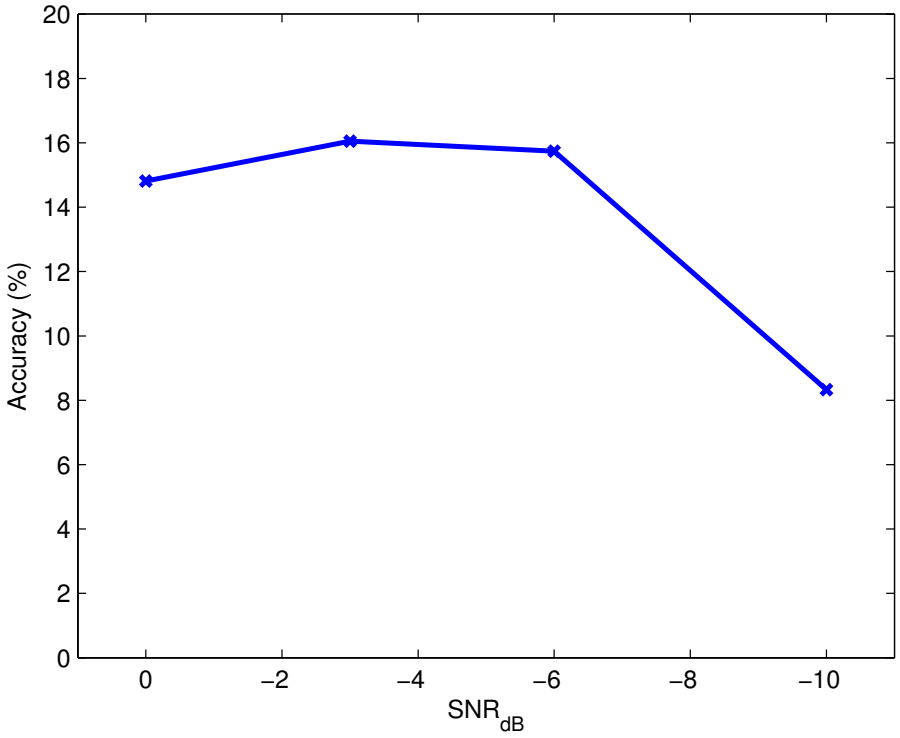
7.3.3 Κινούμενοι ομιλητές

7.3.3.1 Κινούμενοι ομιλητές σε μετωπική στάση

Για τον λόγο που αναφέρθηκε στην προηγούμενη παράγραφο, για τα επόμενα πειράματα έγινε εκπαίδευση του συστήματος χρησιμοποιώντας αρχεία χωρίς θόρυβο και ο έλεγχος έγινε με αρχεία με θόρυβο.

Πίνακας 7.4: Κινούμενοι ομιλητές σε μετωπική στάση-εκπαίδευση χωρίς θόρυβο

SNR	Accuracy
5	88.27%
0	14.81%
-3	16.05%
-6	15.74%
-10	8.33%



Σχήμα 7.2: Κινούμενοι ομιλητές σε μετωπική στάση-εκπαίδευση χωρίς θόρυβο

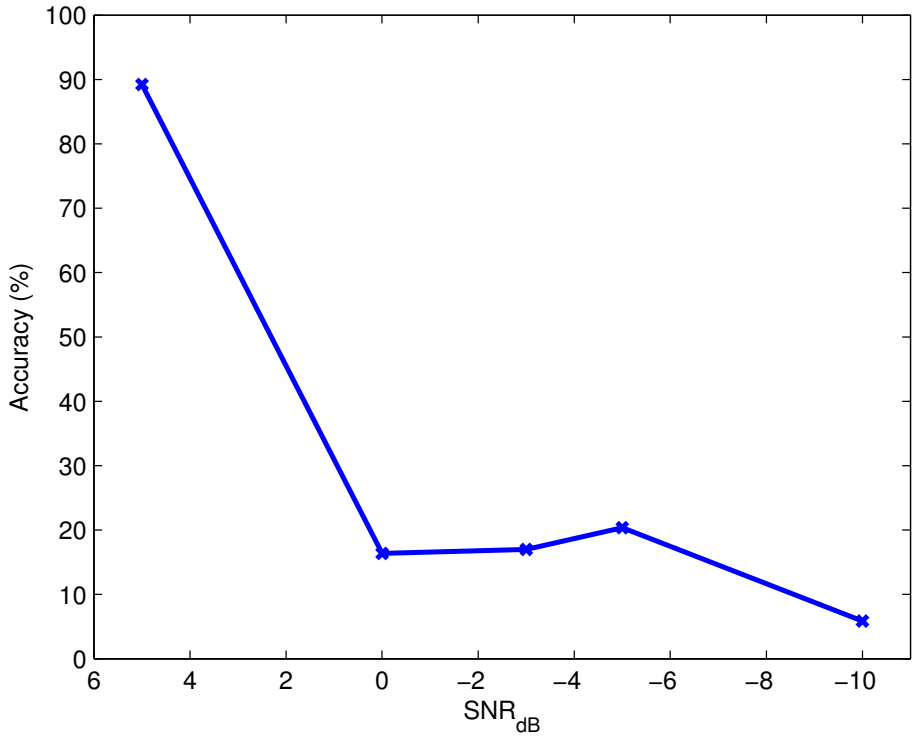
7.3.3.2 Κινούμενοι ομιλητές σε στάση προφιλ

7.3.3.3 Κινούμενοι ομιλητές σε συνεχόμενη ομιλία

Από αυτή τη σειρά πειραμάτων προκύπτει το συμπέρασμα ότι είναι μεγαλύτερη πρόκληση για το σύστημα η αναγνώριση όταν έχει γίνει εκπαίδευση με αρχεία χωρίς θόρυβο και ο έλεγχος με αρχεία που είχαν θόρυβο.

Πίνακας 7.5: Κινούμενοι ομιλητές σε στάση προφιλ

SNR	Accuracy
5	89.2%
0	16.36%
-3	16.98%
-6	20.37%
-10	5.86%

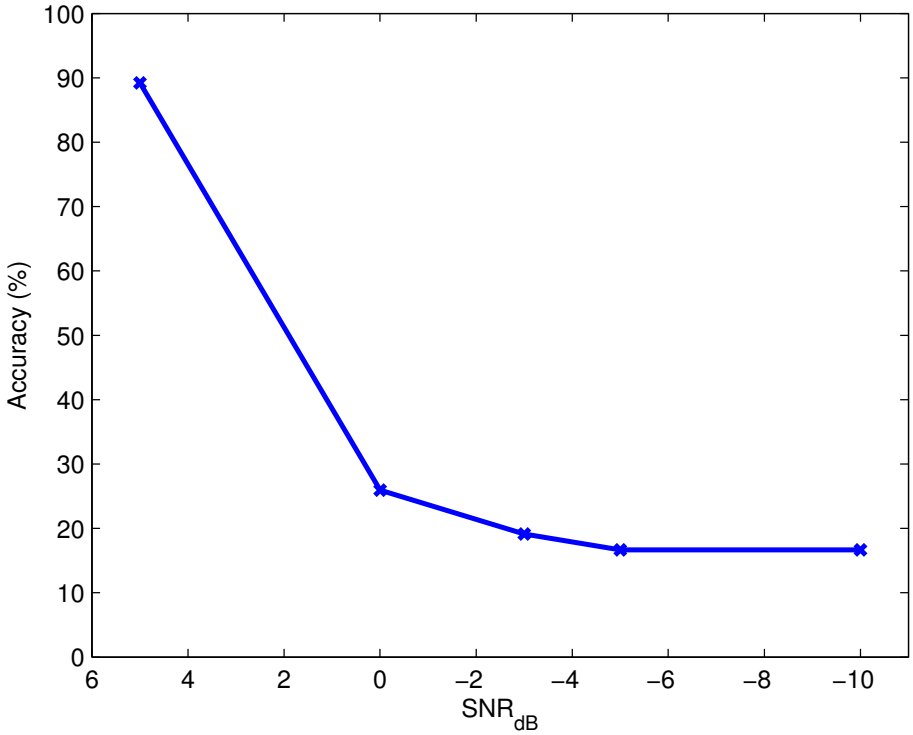


Σχήμα 7.3: Κινούμενοι ομιλητές σε στάση προφιλ

Από αυτή τη σειρά πειραμάτων προκύπτει το συμπέρασμα ότι είναι μεγαλύτερη πρόκληση για το σύστημα η αναγνώριση όταν έχει γίνει εκπαίδευση με αρχεία χωρίς θόρυβο και ο έλεγχος με αρχεία που είχαν θόρυβο.

Πίνακας 7.6: Κινούμενοι Ομιλητές σε συνεχόμενη ομιλία

SNR	Accuracy
5	89.2%
0	25.93%
-3	19.14%
-6	16.67%
-10	16.67%



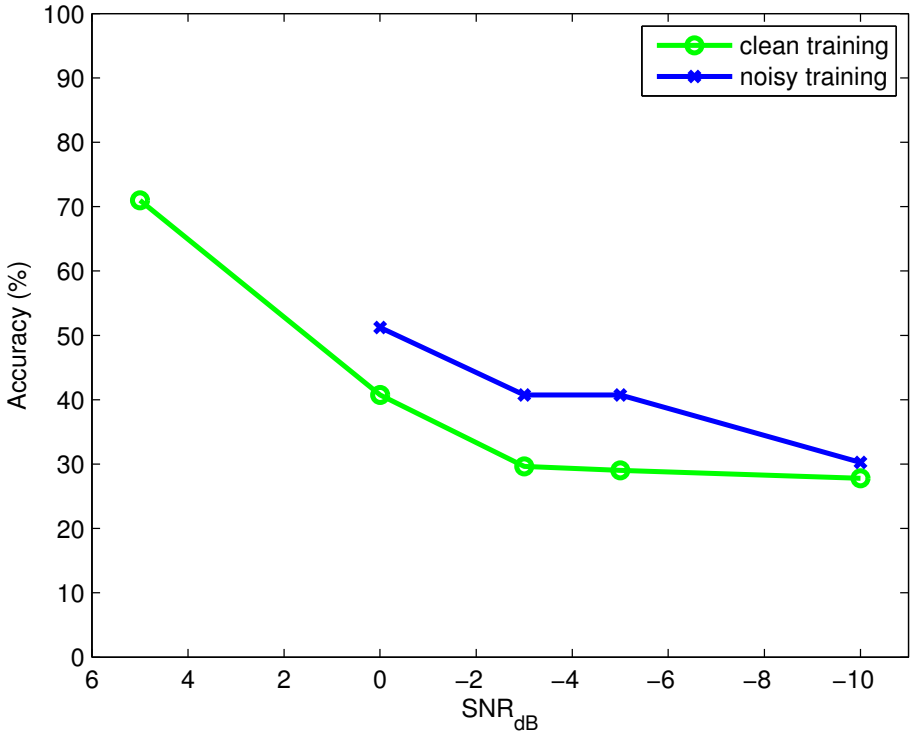
Σχήμα 7.4: Κινούμενοι Ομιλητές σε συνεχόμενη ομιλία

7.3.4 Οπτικοακουστική Αναγνώριση Ομιλίας

Αφού εξήχθησαν οπτικά χαρακτηριστικά από τα βίντεο σύμφωνα με τη διαδικασία που περιγράφεται στο κεφάλαιο 4 και μετά από κατάλληλη, επεξεργασία τα διανύσματα χαρακτηριστικών χρησιμοποιήθηκαν στο σύστημα που περιγράφεται στο Κεφάλαιο 3.Χρησιμοποιώντας μόνο οπτικά χαρακτηριστικά, έγινε αναγνώριση των προφερόμενων λέξεων με ακρίβεια 32% που είναι καλύτερη από ότι στην περίπτωση Clear training -Noisy testing για ένταση θορύβου ίση με 2 ,όπως είδαμε προηγουμένως.

Στη συνέχεια το πρώτο πείραμα με συνδυασμό οπτικών και ακουστικών χαρακτηριστικών (για ένταση θορύβου ίση με 2) έγινε με 7 DCT και 6 MFCC Deltas και Delta-Deltas και η ακρίβεια αναγνώρισης ήταν περίπου ίση με αυτή της περίπτωσης που χρησιμοποιήθηκαν μόνο οπτικά χαρακτηριστικά (30.25%).

Τα αποτελέσματα που προέκυψαν χρησιμοποιώντας 13 MFCC και 7 DCT ήταν καλύτερα από ότι φαίνεται παρακάτω, και για αυτό το λόγο στη συνέχεια έγιναν πειράματα κυρίως με αυτή τη σύνθεση.



Σχήμα 7.5: Συνένωση χαρακτηριστικών-Σύγκριση για εκπαίδευση υπό θόρυβο-εκπαίδευση χωρίς θόρυβο/έλεγχος με θόρυβο

7.3.4.1 Σύγκριση μεθόδων εκπαίδευσης

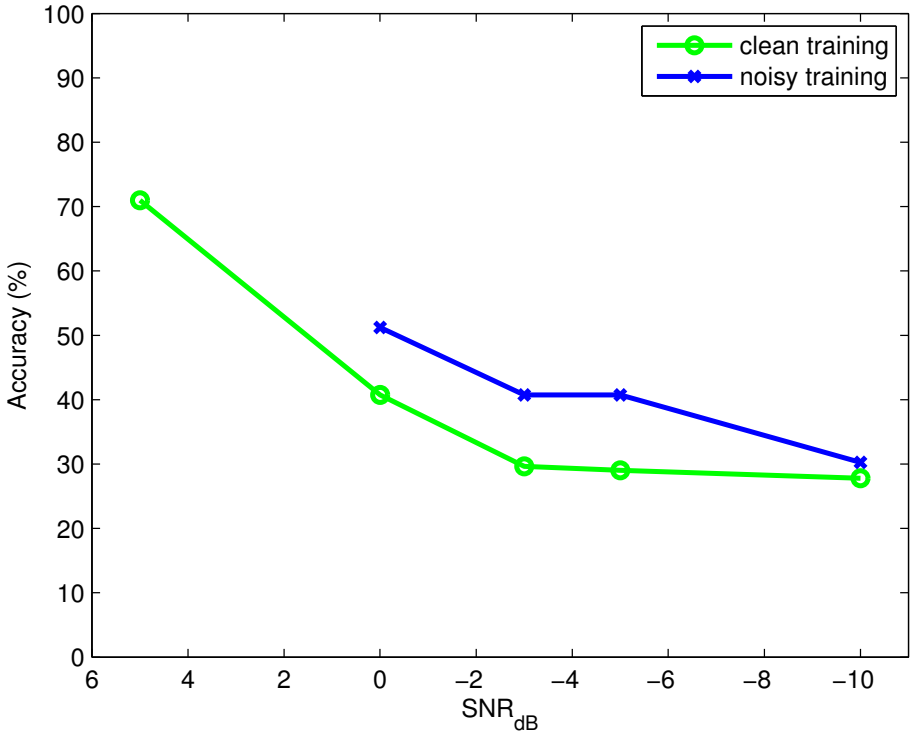
Όπως έγινε και προηγουμένως, για το τμήμα R1 από τα αρχεία ήχου με θόρυβο και από τα αρχεία ήχου χωρίς θόρυβο εξήχθησαν τα αντίστοιχα MFCCs και χρησιμοποιήθηκαν σε συνδυασμό με τα οπτικά χαρακτηριστικά. Έτσι και πάλι εξετάστηκαν δύο διαφορετικοί τρόποι εκπαίδευσης:

- Εκπαίδευση με MFCC με θόρυβο
- Εκπαίδευση με MFCC χωρίς θόρυβο

Παρομοίως με την περίπτωση Ακουστικής Αναγνώρισης όταν η εκπαίδευση έγινε με τα MFCC που προέκυψαν από τα αρχεία που είχαν και θόρυβο τα αποτελέσματα ήταν καλύτερα.

Πίνακας 7.7: Συνένωση χαρακτηριστικών-Σύγκριση για εκπαίδευση υπό θόρυβο-εκπαίδευση χωρίς θόρυβο/έλεγχος με θόρυβο

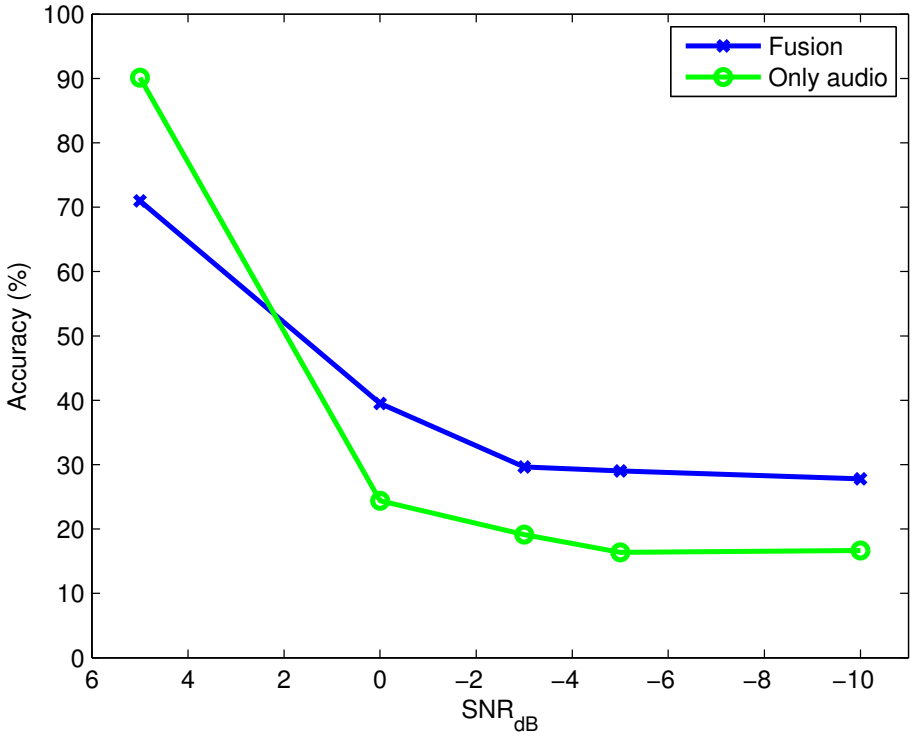
SNR	χωρίς θόρυβο(Accuracy)	με θόρυβο(Accuracy)
5	70.99%	0
0	40.74%	51.23%
-3	29.63%	40.74%
-6	29.01%	40.74%
-10	27.78%	30.25%



Σχήμα 7.6: Συνένωση χαρακτηριστικών-Σύγκριση για εκπαίδευση υπό θόρυβο-εκπαίδευση χωρίς θόρυβο/έλεγχος με θόρυβο

Πίνακας 7.8: Σύγκριση Ακουστικής-Οπτικοακουστικής Αναγνώρισης-εκπαίδευση χωρίς θόρυβο

SNR	Συνένωση(Accuracy)	Μόνο ήχος(Accuracy)
5	70.99%	90.12%
0	39.51%	24.38%
-3	11.11%	19.14%
-6	29.01%	16.36%
-10	27.78%	16.67%



Σχήμα 7.7: Σύγκριση Ακουστικής-Οπτικοακουστικής Αναγνώρισης-εκπαίδευση χωρίς θόρυβο

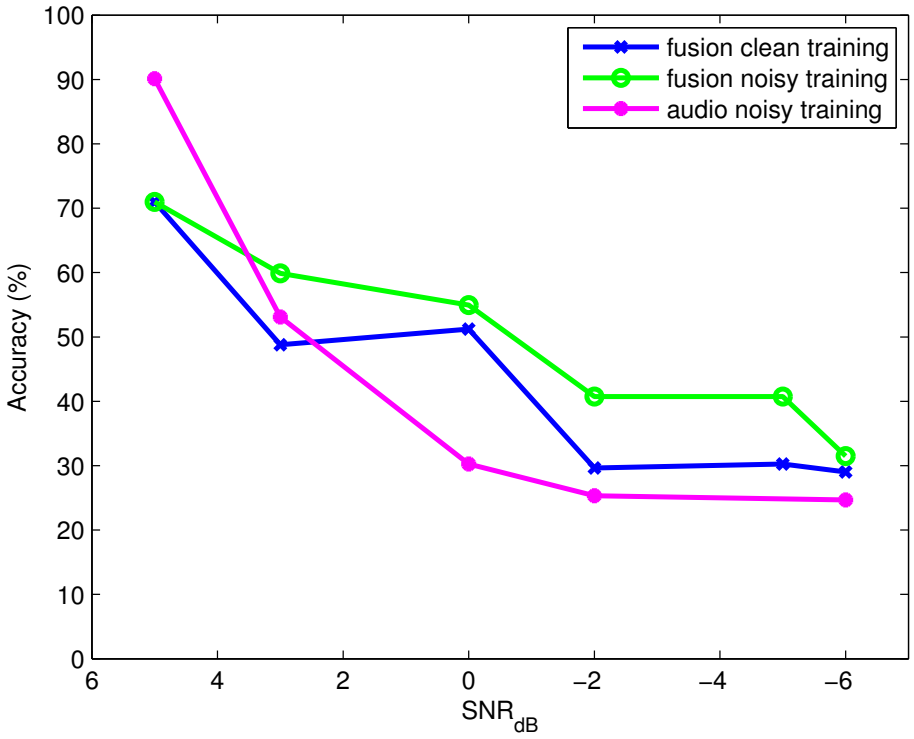
7.3.5 Σύγκριση Ακουστικής-Οπτικοακουστικής Αναγνώρισης

Στη συνέχεια, για τις δύο διαφορετικές μεθόδους εκπαίδευσης έγινε σύγκριση των αποτελεσμάτων που προέκυψαν από τα πειράματα Οπτικοακουστικής Αναγνώρισης Ομιλίας με αυτά που έγιναν χρησιμοποιώντας μόνο την πληροφορία που προέκυψε από τον ήχο.

Από τα παραπάνω φαίνεται ότι ο συνδυασμός ακουστικών και οπτικών χαρακτηριστικών βελτιώνει την ακρίβεια της αναγνώρισης στην περίπτωση που

Πίνακας 7.9: Σύγκριση Ακουστικής-Οπτικοακουστικής Αναγνώρισης

SNR	Συνένωση(Accuracy)	Μόνο ήχος(Accuracy)
5	70.99%	90.12%
0	51.23%	53.09%
-3	40.74%	30.25%
-6	40.74%	25.31%
-10	30.25%	24.69%



Σχήμα 7.8: Σύγκριση Ακουστικής-Οπτικοακουστικής Αναγνώρισης

υπάρχει θόρυβος, με τα θετικά αποτελέσματα να φαίνονται πιο έντονα όσο αυξάνεται η ένταση του θορύβου.

Ειδικά στην περίπτωση όπου για εκπαίδευση χρησιμοποιούνται αρχεία ήχου με θόρυβο η χρήση οπτικών χαρακτηριστικών σε συνδυασμό με τα ακουστικά χαρακτηριστικά βελτιώνει την ακρίβεια της αναγνώρισης ικανοποιητικά από όταν η ένταση θορύβου γίνεται ίση με την ένταση της ομιλίας.

Σε αυτή τη περίπτωση η ακρίβεια της αναγνώρισης μόνο με ηχητικά χαρακτηριστικά είναι 24.38% και η ακρίβεια αναγνώρισης με συνδυασμό οπτικών και ακουστικών χαρακτηριστικών είναι ίση με 54.94%.

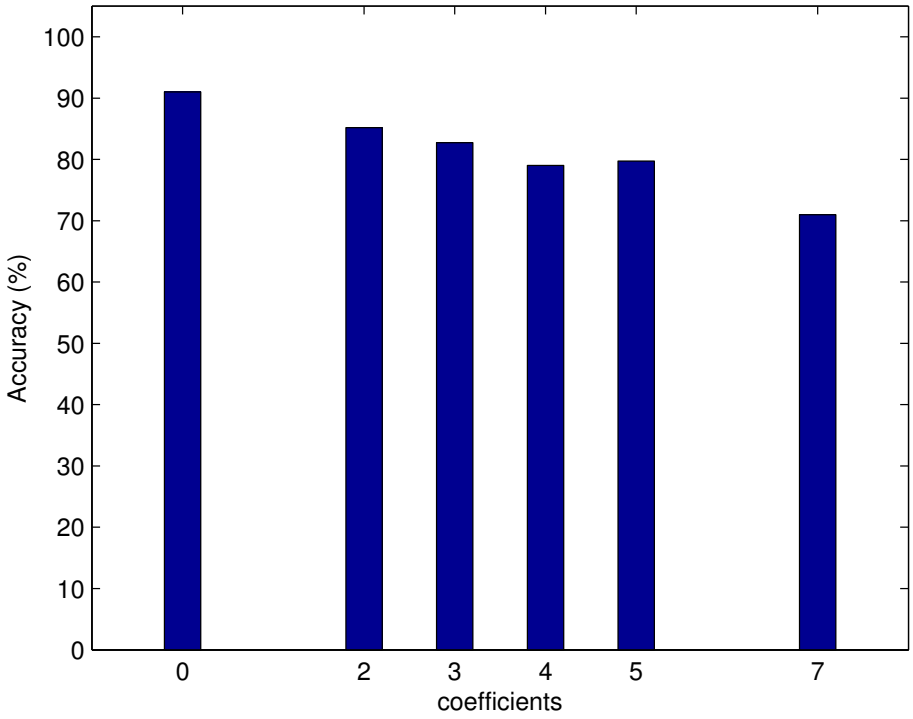
Πίνακας 7.10: Διερεύνηση ως προς το πλήθος των DCT

Πλήθος DCT	Accuracy
0	91.00%
2	85.19%
3	82.72%
4	79.01%
5	79.73%
7	70.99%

7.3.6 Διερεύνηση ως προς το πλήθος των DCT

Από ότι φαίνεται στα παραπάνω πειράματα ο συνδυασμός οπτικών και ηχητικών χαρακτηριστικών για την περίπτωση όπου δεν υπάρχει θόρυβος δίνει χειρότερα αποτελέσματα από ότι εάν χρησιμοποιούταν μόνο ηχητικά χαρακτηριστικά. Ο πιθανότερος λόγος είναι, ότι για το μέγεθος των διανυσμάτων χαρακτηριστικών που χρησιμοποιήθηκαν (διανύσματα διάστασης 60) θα χρειαζόταν μεγαλύτερη βάση δεδομένων για εκπαίδευση. Τυπικά, συνήθως χρησιμοποιούνται διανύσματα διάστασης 39.Επομένως κρίνεται σκόπιμο να διερευνηθεί κατά πόσο βελτιώνεται η απόδοση του συστήματος όσο μειώνεται η διάσταση των διανυσμάτων χαρακτηριστικών. Κρατώντας σταθερό τον αριθμό των MFCC αυξήθηκε σταδιακά το πλήθος των DCT που χρησιμοποιήθηκαν και προέκυψαν τα παρακάτω αποτελέσματα για την περίπτωση που δεν υπάρχει θόρυβος.

Προκύπτει ότι απουσία θορύβου είναι δυνατό να χρησιμοποιηθεί μικρότερο πλήθος DCT.



Σχήμα 7.9: Διερεύνηση ως προς το πλήθος των DCT

Κεφάλαιο 8

Συμπεράσματα

8.1 Συμπεράσματα

Η Οπτικοακουστική Αναγνώριση Ομιλίας είναι ένας τομέας με μεγάλη δυναμική όπου χρησιμοποιούνται τεχνικές όρασης υπολογιστών σε συνδυασμό με τεχνικές της ακουστικής αναγνώρισης ομιλίας. Με αυτό τον συνδυασμό δυσκολίες που συναντώνται σε πραγματικά περιβάλλοντα ομιλίας, όπως για παράδειγμα ο θόρυβος υποβάθρου και η ύπαρξη πολλών ομιλητών επιδιώκεται να αντιμετωπιστούν. Αυτό γίνεται χρησιμοποιώντας την επιπλέον πληροφορία που μπορεί να εξαχθεί από τα οπτικά χαρακτηριστικά του βίντεο.

Στα ακουστικά συστήματα αυτόματης αναγνώρισης ομιλίας από τα αρχεία ήχου εξάγονται κατάλληλα χαρακτηριστικά που χρησιμοποιούνται με σκοπό την αναγνώριση ομιλίας.

Σε πραγματικά περιβάλλοντα ο θόρυβος υποβάθρου και η ύπαρξη πολλών ομιλητών δυσχεραίνουν σημαντικά τις συνθήκες για την αυτόματη αναγνώριση ομιλίας χρησιμοποιώντας μόνο ήχο.

Για αυτές τις περιπτώσεις κρίνεται σκόπιμο να χρησιμοποιηθούν και οπτικά χαρακτηριστικά. Έτσι, στην παρούσα διπλωματική χρησιμοποιήθηκαν τεχνικές Όρασης Υπολογιστή (Computer Vision) και με τη χρήση κατάλληλων ταξινομητών (Haar Classifiers) εντοπίστηκε η περιοχή του στόματος του εκάστοτε ομιλητή και αφού έγινε εξαγωγή των κατάλληλων χαρακτηριστικών (feature extraction) συνδυάστηκαν με τα χαρακτηριστικά που προέκυψαν από τον ήχο. Στη συνέχεια ο συνδυασμός των χαρακτηριστικών χρησιμοποιήθηκε για την εκπαίδευση ενός Οπτικοακουστικού Συστήματος Αναγνώρισης Ομιλίας (Audio-Visual Speech Recognition).

Αποδείχθηκε ότι ο συνδυασμός οπτικών χαρακτηριστικών και ακουστικών χαρακτηριστικών βελτιώνει σημαντικά την απόδοση ενός συστήματος Αυτόματης Αναγνωισης Ομιλίας σε συνθήκες θορύβου.

8.1.1 Μελλοντικές κατευθύνσεις έρευνας

Στήν παρούσα διπλωματική φάνηκε πως είναι αποτελεσματική η μέθοδος του συνδυασμού οπτικών και ακουστικών χαρακτηριστικών για την Αυτόματη Αναγνώριση Ομιλίας και έτσι,μελλοντικά θα μπορούσε να γίνει έρευνα σε απρόμοιες κατευθύνσεις.

Καταρχήν θα μπορούσε να δοκιμαστούν παρόμοιες μέθοδοι σε βάσεις δεδομένων που δεν δημιουργήθηκαν σε εργαστηριακές συνθήκες και που προσομοιάζουν πραγματικά περιβάλλοντα. Για παράδειγμα υπάρχουν βάσεις με βίντεο από δελτία ειδήσεων ή από περιβάλλον αυτοκινήτου που θα αποτελλούν σημαντικές προκλήσεις για τέτοιου είδους μεθόδους.Ακόμα θα μπορούσε να επιχειρηθεί η δημιουργία ενός συστήματος που θα κάνει οπτικοακουστική αναγνώριση ομιλίας σε πραγματικό χρόνο.

Κατα την διάρκεια εκπόνησης αυτής της εργασίας παρατηρήθηκε ότι σε ότι αφορά την καταληπτότητα της ομιλίας , δεν υπάρχουν κάποιες μέθοδοι και ακόμα περισσότερο μονάδες μέτρησης για την αξιολόγηση των βίντεο. Επομένως θα μπορούσε σε αυτή τη κατεύθυνση να προταθούν μέθοδοι για την αξιολόγηση των βίντεο σε ότι αφορά την Αυτόματη αναγνώριση ομιλίας ή για την αξιολόγηση της ποιότητας βιντεοκλήσεων ή ακόμα και της άρθρωσης των ομιλητών.

Τέλος, μια κατεύθυνση με πιθανές πρακτικές εφαρμογές θα ήταν η αναγνώριση συναισθημάτων των ομιλητών σε συνδυασμό με την οπτικοακουστική αναγνώριση ομιλίας.

Παράρτημα Α

Κείμενα Ομιλητών

Α.1 Μεμονωμένοι Ομιλητές

Κάθε ομιλητής πρόφερε τα ψηφία που βρίσκονται παρακάτω σε αγγλικό κείμενο με παύση ένα δευτερόλεπτο μεταξύ των ψηφίων. Κάθε ομιλητής στεκόταν ακίνητος στο κέντρο και κοιτάζοντας απευθείας στο φακό της κάμερας.

** Μεμονωμένα ψηφία-ακίνητοι ομιλητές: τμήμα **R1** **

1. ZERO ONE TWO THREE FOUR FIVE SIX SEVEN EIGHT NINE

2. ZERO ONE TWO THREE FOUR FIVE SIX SEVEN EIGHT NINE

3. ZERO ONE TWO THREE FOUR FIVE SIX SEVEN EIGHT NINE

4. ZERO ONE TWO THREE FOUR FIVE SIX SEVEN EIGHT NINE

5. ZERO ONE TWO THREE FOUR FIVE SIX SEVEN EIGHT NINE

Κάθε ομιλητής πρόφερε τα ψηφία που βρίσκονται παρακάτω σε αγγλικό κείμενο με παύση ένα δευτερόλεπτο μεταξύ των ψηφίων. Οι ομιλητές κινούνται αργά, παραμένοντας στο πεδίο της κάμερας. Μερικοί κινούνται και μπρος/πίσω

** Μεμονωμένα ψηφία-κινούμενοι ομιλητές:τμήμα **R2** **

1. NINE EIGHT SEVEN SIX FIVE FOUR THREE TWO ONE ZERO

2. NINE EIGHT SEVEN SIX FIVE FOUR THREE TWO ONE ZERO

3. NINE EIGHT SEVEN SIX FIVE FOUR THREE TWO ONE ZERO

Κάθε ομιλητής πρόφερε τα ψηφία που βρίσκονται παρακάτω σε αγγλικό κείμενο με παύση ένα δευτερόλεπτο μεταξύ των ψηφίων. Τα πρώτα 10 ψηφία τα προφέρουν γυρισμένοι σε στάση αριστερού προφιλ ενώ τα υπόλοιπα σε στάση δεξιού προφιλ.

** Μεμονωμένα ψηφία-στάση προφιλ:τμήμα **R3** **
1. ZERO ONE TWO THREE FOUR FIVE SIX SEVEN EIGHT NINE

2. ZERO ONE TWO THREE FOUR FIVE SIX SEVEN EIGHT NINE

Στη συνέχεια, οι ομιλητές πρόφεραν τους παρακάτω τηλεφωνικούς αριθμούς σε συνεχόμενα ψηφία, διακόπτοντας για 3 δευτερόλεπτα μεταξύ κάθε αριθμού τηλεφώνου.

Συνεχόμενα ψηφία-ακίνητοι ομιλητές: Τμήμα **R4 **

(861) 234-5970
(904) 581-2637
(712) 835-6049

Στη συνέχεια, αφού οι ομιλητές περίμεναν πέντε δευτερόλεπτα, πρόφεραν τους παρακάτω τηλεφωνικούς αριθμούς σε συνεχόμενα ψηφία, διακόπτοντας για 3 δευτερόλεπτα μεταξύ κάθε αριθμού τηλεφώνου.
** Συνεχόμενα ψηφία-κινούμενοι ομιλητές:Τμήμα **R5****

(534) 120-9768
(835) 127-4069
(912) 305-6478

A.2 Ζεύγη Ομιλητών

Στα βιντεο με δύο ομιλητές ο ομιλητής που στεκόταν αριστερά πρόφερε τα ψηφία που είναι επισημασμένα με **A** ενώ ο ομιλητής στα δεξιά πρόφερε τα ψηφία που είναι επισημασμένα με **B**. Αρχικά μιλάει ο κάθε ομιλητής χωριστά με παύση ένα δευτερόλεπτο μεταξύ των ψηφίων.

A: 8—6—1 2—3—4 5—9—7—0
Στη συνέχεια...
B: 9—0—4 5—8—1 2—6—3—7

B: 7—1—2 8—3—5 6—0—4—9
Στη συνέχεια...
A: 5—3—4 1—2—0 9—7—6—8

Στη συνέχεια, και πάλι με παύση ενός δευτερολέπτου μεταξύ των ψηφίων, και οι δυο ταυτόχρονα προφέρουν τα παρακάτω:

A: 8—3—5 1—2—7 4—0—6—9
B: 9—1—2 3—0—5 6—4—7—8

Βιβλιογραφία

[1] G. Potamianos and , P. Scanlon, "Exploiting lower face symmetry in appearance-based automatic speech reading," in *Proc. Works. Audio-Visual Speech Process. (AVSP)*, Vancouver Island, Canada, 2005, pp. 79-84.

[2] G. Galatas, G. Potamianos, and F. Makedon, "Audio-visual speech recognition using depth information from the Kinect in noisy video conditions," in *Int. Conf. Pervasive Technologies Related to Assistive Environments (PETRA), 2012.*, Crete, Greece, 2012, p. 2.

[3] G. Galatas, G. Potamianos, and F. Makedon, "Audio-visual speech recognition incorporating facial depth information captured by the Kinect," in *Europ. Conf. Signal Process. (EUSIPCO)* , Bucharest, Romania, 2012, pp. 2714 - 2717.

[4] P. Giannoulis and G. Potamianos, "A hierarchical approach with feature selection for emotion recognition from speech," in *Int. Conf. Language Resources and Evaluation (LREC)* , Istanbul, Turkey, 2012, pp. 1203-1206.

[5] D. Dimitriadis, P. Maragos, and A. Potamianos, "Modulation features for speech recognition," in *proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2002, pp. I-377-I-380.

[6] P. I. Wilson and J. Fernandez, "Face feature detection using HAAR classifier ," in *JCSC*, 2006, pp. 127-133.

[7] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," in *Conference on Computer Vision and Pattern Recognition*, 2001, pp. 1-9.

[8] G. Potamianos, H. P. Graf, and E. Cossato, "An image transform approach for fmm based automatic lipreading ," in *International Conference on Image Processing* , 1998, pp. 173-177.

[9] G. Fanelli, J. Gall, and L. V. Gool, "Hough Transform-based Mouth Localization for Audio-Visual Speech Recognition ," in *Proc. of the British Mach. Vis. Conf.*, 2009.

[10] G. Potamianos and C. Neti, "Improved ROI and within frame discriminant features for lipreading," in *Proc. Int. Conf. Image Process. (ICIP)*, vol. III, Thessaloniki,Greece, 2001, pp. 250-253.

[11] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *Acoustics, Speech, and Signal Processing (ICASSP) IEEE International Conference* , vol. 2, 2002, pp. II-2017 - II-2020.

[12] D. Dimitriadis et al., "GridNews: A distributed automatic Greek broadcast transcription system," in *Acoustics, Speech and Signal Processing, ICASSP, IEEE International Conference*, Taipei , 2009, pp. 1917 - 1920.

[13] P. Angkititrakul, M. Petracca, A. Sathyanarayana, and J.H.L. Hansen, "UTDrive: Driver Behavior and Speech Interactive Systems for In-Vehicle Environments," in *Intelligent Vehicles Symposium, IEEE* , Istanbul, 2007, pp. 566 - 569.

[14] K. Kumar, G. Potamianos, J. Navratil, E. Marcheret, and V. Libal, "Audio-Visual Speech

Synchrony Detection by a Family of Bimodal Linear Prediction Models," in *Multibiometrics for Human Identification*, B. Bhanu and V. Govindaraju, Eds.: Cambridge University Press, 2011, ch. 2 , pp. 31-50.

[15] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proceedings of the IEEE*, vol. 91, pp. 1306-1326, 2003.

[16] L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, "Speaker Independent Recognition of Isolated Words Using Clustering Techniques," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 27, pp. 336-349, 1979.

[17] L. R. Rabiner, "Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, pp. 257-286, 1989.

[18] N. R. Thota and S. K. Devireddy, "Image Compression Using Discrete Cosine Transform," *Georgian Electronic Scientific Journal: Computer Science and Telecommunications*, vol. 3, no. 17, pp. 35-43, 2008.

[19] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision* , vol. 57, no. 2, pp. 137-154, 2004.

[20] R. Padilla, C.F. F. Costa Filho, and M.G. F. Costa, "Evaluation of Haar Cascade Classifiers Designed for Face Detection," *World Academy of Science, Engineering and Technology* , vol. 64, pp. 362-365, 2012.

[21] G. Bradski and A. Kaehler, *Learning OpenCV*, M. Loukides, Ed.: O'Reilly Media, 2008.

[22] L. Rabiner and B. H. Juang, *Fundamentals of speech recognition*. NJ: Englewood Cliffs, 1993.

[23] J. R. Deller, J.H.L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, Second Ed. ed. New York: IEEE Press, 2000.

[24] G. Potamianos, H. P. Neti, J. Luettin, and I. Matthews, , E. Vatikiotis-Bateson, G. Bailly, and P. Perrier, Eds.: Cambridge University Press, 2012, ch. 9.

[25] P. Besson, G. Monaci, P. Vandergheynst, and M. Kunt, "Experimental framework for speaker detection on the CUAVE database," EPFL, Lausanne, Switzerland, Tech. Rep. EPFL-REPORT-87331 2006.

[26] A. Reimondo. (2004) OpenCV Swiki. [Online]. http://alereimondo.no-ip.org/OpenCV

[27] OpenCV. [Online]. http://opencv.org/

[28] FFmpeg. [Online]. http://www.ffmpeg.org/

[29] SoX - Sound eXchange. [Online]. http://sox.sourceforge.net/

[30] (2008) flex: The Fast Lexical Analyzer. [Online]. http://flex.sourceforge.net/

[31] (1997) dirent.h. [Online]. http://pubs.opengroup.org/onlinepubs/007908799/xsh/dirent.h.html

[32] A. Hunt. (1993-6) BEEP dictionary. [Online]. http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html

[33] OpenCV Resources. [Online]. http://www.cognotics.com/opencv/

[34] HTK Speech Recognition Toolkit. [Online]. http://htk.eng.cam.ac.uk/