Πανεπιστήμιο Θεσσαλίας
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Οπτικοακουστική αναγνώριση φωνής
Audiovisual speech recognition

Διπλωματική Εργασία
του
Κωνσταντίνου Λαμπρόπουλου

Επιβλέπων: Γεράσιμος Ποταμιάνος
Αναπληρωτής Καθηγητής

Μέλοι επιτροπής:
Νικόλαος Μπέλλας, Αναπληρωτής Καθηγητής
Δημήτριος Κατσαρός, Αναπληρωτής Καθηγητής

15 Οκτωβρίου 2019

Περίληψη

Αυτή η διπλωματική εργασία ασχολείται με τον συνδυασμό των καναλιών οπτικής και ακουστικής πληροφορίας για την πραγματοποίηση αυτόματης αναγνώρισης φωνής/ομιλίας. Η αυτόματη αναγνώριση φωνής αποτελεί ουσιαστικό μέρος της παραγωγής πρακτόρων συνομιλίας, εικονικών βοηθών που έχουν δει πρόσφατα σημαντική οικονομική επιτυχία. Παρά την οικονομική τους επιτυχία, οι επιδόσεις αυτών των εικονικών βοηθών, ακόμη και σε ιδανικά περιβάλλοντα, δεν είναι βέλτιστες. Οι επιδόσεις τους υποβαθμίζονται σημαντικά σε πολυφωνικά ή θορυβώδη περιβάλλοντα. Το γεγονός αυτό οδήγησε ιστορικά στην ιδέα της αξιοποίησης του οπτικού καναλιού πληροφορίας για την κατασκευή πιο ισχυρών και αξιόπιστων συστημάτων αναγνώρισης φωνής/ομιλίας. Αυτή η εργασία θα προσπαθήσει να παρουσιάσει έναν θεμελιώδη τρόπο κατασκευής ενός τέτοιου συστήματος και να την αξιολογήσει σε ένα θορυβώδες προσομοιωμένο περιβάλλον για να αναγνωρίσει κοινές φράσεις. Πιο συγκεκριμένα, έγινε μια προσπάθεια αναγνώρισης δέκα σύντομων φράσεων χρησιμοποιώντας μια βάση δεδομένων η οποία διέθετε οπτικοακουστική πληροφορία, όπου η οπτική πληροφορία προσφερόταν από πέντε διαφορετικές οπτικές γωνίες.

Abstract

This diploma thesis deals with the combination of visual and auditory information channels for automatic voice / speech recognition. Automatic voice recognition is an essential part of the production of chat agents, virtual assistants who have recently seen significant financial success. Despite their financial success, the performance of these virtual assistants, even in ideal environments, is not optimal. Their performance is significantly degraded in polyphonic or noisy environments. This has historically led to the idea of leveraging the visual information channel to build more robust and reliable voice / speech recognition systems. This thesis will attempt to present a fundamental way of constructing such a system and evaluate it in a noisy simulated environment to recognize common phrases. Specifically, an attempt was made to recognize ten short phrases using a database containing audiovisual information, where visual information was provided from five different angles.

Ευχαριστίες

Θα ήθελα να εκφράσω την ειλικρινή ευγνωμοσύνη μου προς όλους τους επιβλέποντες μου για την καθοδήγηση και την βοήθειά τους που παρείχαν σε όλη την εξέλιξη αυτής της διατριβής.

Αφιέρωση

Στην οικογένεια μου ...

Οπτικοακουστική αναγνώριση φωνής

Κωνσταντίνος Λαμπρόπουλος
klampropoulos@uth.gr

15 Οκτωβρίου 2019

Περιεχόμενα

1 Εισαγωγή **6**
- 1.1 Αναγνώριση Ομιλίας: Ιστορική αναδρομή 6
- 1.2 Οπτική και Οπτικοακουστική αναγνώριση φωνής/ομιλίας 7
- 1.3 Σκοπός της διπλωματικής 9
- 1.4 Δομή της διπλωματικής 9

2 Βάση Δεδομένων / Εργαλεία Λογισμικού **11**
- 2.1 Βάση Δεδομένων 11
- 2.2 Εργαλεία Λογισμικού 13

3 Κρυφό Μαρκοβιανό μοντέλο/ Γκαουσιανό μοντέλο μίξης **15**
- 3.1 Δημιουργία προαπαιτούμενων αρχείων 16
- 3.2 Εκπαίδευση μοντέλου Γκαουσιανό μοντέλο μίξης /Κρυφό Μαρκοβιανό μοντέλο 22
- 3.3 Αποκωδικοποίηση/Αξιολόγηση μοντέλων 25

4 Οπτική επεξεργασία **27**
- 4.1 Ταξινόμηση όψεων 27
- 4.2 Αυτόματοι κωδικοποιητές 29

5 Υβριδική μοντελοποίηση για οπτικοακουστική αναγνώριση ομιλίας **32**
- 5.1 Ακουστικό μοντέλο 32
- 5.2 Οπτικό μοντέλο 33
- 5.3 Οπτικοακουστική αναγνώριση ομιλίας 34

6 Πειραματικά Αποτελέσματα **35**
- 6.1 Γκαουσιανό Μοντέλο Μίξης/Κρυφό Μαρκοβιανό Μοντέλο 35
- 6.2 Αυτόματος ταξινομητής όψεων 36
- 6.3 Αυτόματοι κωδικοποιητές 38
- 6.4 Μοντέλα τύπου HMM-DNN 42
  - 6.4.1 Ακουστικό μοντέλο 42
  - 6.4.2 Οπτικά μοντέλα 43
  - 6.4.3 Οπτικοακουστικά μοντέλα 47

7 Μελλοντική Εργασία **58**

Κατάλογος Σχημάτων

2.1 Πέντε όψεις από την OuluVS2 βάση δεδομένων (Πηγή: [1]) 12

3.1 Δομή Καταλόγων Πειράματος 16

6.1 Απόδοση του ταξινομητή όψεων στο σύνολο του σετ δοκιμής 36
6.2 Αρνητική λογαριθμική απώλεια πιθανοφάνειας για τα δεδομένα αξιολόγησης σε συνάρτηση με το χρόνο επαναλήψεων 37
6.3 Αναλογία μέγιστου σήματος προς θόρυβο του σετ δοκιμής για τον αυτόματο κωδικοποιητή πρόσοψης 38
6.4 Αναλογία μέγιστου σήματος προς θόρυβο του σετ δοκιμής για τον αυτόματο κωδικοποιητή για την όψη 30° 39
6.5 Αναλογία μέγιστου σήματος προς θόρυβο του σετ δοκιμής για τον αυτόματο κωδικοποιητή για την όψη 45° 40
6.6 Αναλογία μέγιστου σήματος προς θόρυβο του σετ δοκιμής για τον αυτόματο κωδικοποιητή για την όψη 60° 40
6.7 Αναλογία μέγιστου σήματος προς θόρυβο του σετ δοκιμής για τον αυτόματο κωδικοποιητή για την όψη 90° 41
6.8 Απόδοση ταξινόμησης ακουστικού μοντέλου DNN-HMM κατά τη διάρκεια εκπαίδευσης 42
6.9 Απόδοση ταξινόμησης οπτικού μοντέλου DNN-HMM για την πρόσοψη κατά τη διάρκεια εκπαίδευσης 43
6.10 Απόδοση ταξινόμησης οπτικού μοντέλου DNN-HMM για την όψη 30° κατά τη διάρκεια εκπαίδευσης 44
6.11 Απόδοση ταξινόμησης οπτικού μοντέλου DNN-HMM για την όψη 45° κατά τη διάρκεια εκπαίδευσης 45
6.12 Απόδοση ταξινόμησης οπτικού μοντέλου DNN-HMM για την όψη 60° κατά τη διάρκεια εκπαίδευσης 46
6.13 Απόδοση ταξινόμησης οπτικού μοντέλου DNN-HMM για την όψη 90° κατά τη διάρκεια εκπαίδευσης 47
6.14 Λόγος σφάλματος φράσης σε συνάρτηση με την αναλογία θορύβου σήματος για την πρόσοψη 48
6.15 Λόγος σφάλματος φράσης σε συνάρτηση με την αναλογία θορύβου σήματος για την όψη 30° 50
6.16 Λόγος σφάλματος φράσης σε συνάρτηση με την αναλογία θορύβου σήματος για την όψη 45° 52

6.17 Λόγος οφάλματος φράσης σε συνάρτηση με την αναλογία θορύβου σήματος για την όψη 60° 54
6.18 Λόγος οφάλματος φράσης σε συνάρτηση με την αναλογία θορύβου σήματος για την όψη 90° 56

Κατάλογος Πινάκων

2.1 Φράσεις από την OuluVS2 βάση δεδομένων. 13

6.1 Αποτελέσματα μοντέλων τύπου Γκαουσιανό Μοντέλο Μίξης/Κρυφό Μαρκοβιανό Μοντέλο 35
6.2 Μητρώο σύγχυσης για τον αυτόματο ταξινομητή όψεων 37
6.3 Αποτελέσματα οπτικοακουστικού μοντέλου για την πρόοοψη σε όλες τις αναλογίες θορύβου σήματος και τα αντίστοιχα βάρη από το ακουστικό μοντέλο 49
6.4 Αποτελέσματα οπτικοακουστικού μοντέλου για την όψη 30° σε όλες τις αναλογίες θορύβου σήματος και τα αντίστοιχα βάρη από το ακουστικό μοντέλο 51
6.5 Αποτελέσματα οπτικοακουστικού μοντέλου για την όψη 45° σε όλες τις αναλογίες θορύβου σήματος και τα αντίστοιχα βάρη από το ακουστικό μοντέλο 53
6.6 Αποτελέσματα οπτικοακουστικού μοντέλου για την όψη 60° σε όλες τις αναλογίες θορύβου σήματος και τα αντίστοιχα βάρη από το ακουστικό μοντέλο 55
6.7 Αποτελέσματα οπτικοακουστικού μοντέλου για την όψη 90° σε όλες τις αναλογίες θορύβου σήματος και τα αντίστοιχα βάρη από το ακουστικό μοντέλο 57

Κεφάλαιο 1

Εισαγωγή

1.1 Αναγνώριση Ομιλίας: Ιστορική αναδρομή

Ο πιο απλοϊκός και περιγραφικός ορισμός της αναγνώριοης ομιλίας είναι ότι αποτελεί την διαδικασία μετατροπής του ηχογραφημένου ή ζωντανού ήχου σε μια ακολουθία λέξεων. Η αναγνώριση ομιλίας, ή σε γενικότερες γραμμές η φωνητική αναγνώριση, αποτελεί κέντρο ανθρωπίνου ενδιαφέροντος για έναν σχεδόν αιώνα. Η ίδρυση των θεμελίων αυτών των πεδίων οφείλεται κατά κύριο λόγο στα Bell Labs, όπου η πρώτη ηλεκτρονική συοκευή σύνθεοης ομιλίας και το πρώτο σύστημα που θα μπορούσαν να αναγνωρίοουν τα ψηφιακά ψηφία επινοήθηκαν στις απομακρυομένες δεκαετίες του 1930 και 1950 αντίστοιχα.

Μια οημαντική ιστορική ανακάλυψη στον τομέα της αναγνώριοης ομιλίας συνέβη κατά τη διάρκεια της δεκαετίας του 1970 και της δεκαετίας του 1980 με την υιοθέτηση στατιστικών μοντέλων. Μέχρι αυτό το γεγονός, τα περισσότερα συστήματα αναγνώριοης ομιλίας βαοίστηκαν σε μια οχετικά απλή εκδοχή μοντελοποίηοης της κοινής βελτιστοποίηοης σε ολικό επίπεδο [2]. Ένα οημαντικό γεγονός ήταν η παρουσίαοη της ομιλίας ως μια κρυμμένη Μαρκοβιανή διαδικαοία [3].

Επίοης, γνωστό ως κρυφό Μαρκοβιανό μοντέλο, η ενοωμάτωοη αυτού του μοντέλου οτην αναπτυξιακή διαδικαοία των ουστημάτων αναγνώριοης ομιλίας τελικά οδήγησε στα πρώτα πρακτικά ουστήματα αυτόματης αναγνώριοης ομιλίας που θα μπορούσαν ενδεχομένως να αναγνωρίοουν δέκα χιλιάδες λέξεις. Επιπλέον, η χρήοη του κρυφού Μαρκοβιανού μοντέλου επέτρεψε στους ερευνητές να ουνδυάοουν διαφορετικές πηγές γνώοης, όπως ακουοτική, γλώοοα και σύνταξη, σε ένα ενοποιημένο πιθανολογικό

μοντέλο [4].

Στις επόμενες δύο δεκαετίες, το πεδίο της αυτόματης αναγνώρισης ομιλίας παρουσίασε μια αργή αλλά σταθερή πρόοδο. Οι βελτιώσεις προήλθαν από την καινοτομία σε διάφορες υπομονάδες της αναγνώρισης ομιλίας, όπως :

- Στατιστική μοντελοποίηση και μηχανική μάθηση
- Επεξεργασία σημάτων και επεξεργασία χαρακτηριστικών
- Γλωσσική μοντελοποίηση
- Αποτελεσματικές τεχνικές αποκωδικοποίησης

Αυτά τα συστήματα είχαν ως πυρήνα μοντελοποίησης ένα Γκαουσιανό μοντέλο μίξης /κρυφό Μαρκοβιανό μοντέλο που διατηρεί μια ιεραρχική δομή.

Ένας άλλος σημαντικός παράγοντας για την πρόοδο του πεδίου αναγνώρισης φωνής ήταν η σταδιακή αύξηση των υπολογιστικών δυνατοτήτων στους ψηφιακούς υπολογιστές. Αυτή η αύξηση τελικά έδωσε την ώθηση σε μοντέλα βαθιάς μάθησης που υιοθετήθηκαν από την ερευνητική κοινότητα αναγνώρισης ομιλίας για την ανώτερη διακριτική τους απόδοση αντικαθιστώντας το Γκαουσιανό μοντέλο μίξης στην ήδη υπάρχουσα αρχιτεκτονική αναγνώρισης. Αυτά τα μοντέλα ονομάστηκαν υβριδικά μοντέλα ή συντομογραφικά ως HMM-DNN, και αυτά είναι τα μοντέλα που θα είναι το κέντρο ενδιαφέροντος αυτής της διατριβής.

1.2 Οπτική και Οπτικοακουστική αναγνώριση φωνής/ομιλίας

Αν και η αντίληψη της ομιλίας θεωρείται ως μια ακουστική δεξιότητα, είναι εγγενώς πολυτροπική, αφού η παραγωγή ομιλίας απαιτεί από τον ομιλητή να κάνει κινήσεις των χειλιών, των δοντιών και της γλώσσας που είναι συχνά ορατές στην επικοινωνία πρόσωπο με πρόσωπο. Το διάβασμα χειλιών, γνωστό και ως οπτική αναγνώριση ομιλίας, είναι μια τεχνική ή διαδικασία για την κατανόηση του λόγου με την οπτική ερμηνεία των εκφράσεων του προσώπου του ομιλητή, όταν ο κανονικός ήχος δεν

είναι διαθέσιμος [5].

Το κύριο δομικό στοιχείο της ανάγνωσης των χειλιών καλείται viseme. Το viseme αντιστοιχεί σε διαφορετικές μορφές στόματος που παρουσιάζονται κατά τη διάρκεια της ομιλίας. Δεν υπάρχει άμεση αντιστοίχιση μεταξύ των visemes και των φωνητικών ήχων της ομιλίας, καθιστώντας την οπτική αναγνώριση ομιλίας ένα δύσκολο πρόβλημα επειδή μια ακολουθία visemes δεν μπορεί να συσχετιστεί με μια μοναδική ακολουθία λέξεων / φράσης.

Η οπτική αναγνώριση ομιλίας βασίζεται επίσης σε πληροφορίες που παρέχονται από το περιβάλλον, τη γνώση της γλώσσας και οποιαδήποτε υπολειπόμενη ακοή. Παρόλο που χρησιμοποιούνται κυρίως από άτομα με δυσκολία στην ακρόαση, οι περισσότεροι άνθρωποι με κανονική δυνατότητα ακρόασης δέχονται ασυνείδητα σημαντική ποσότητα πληροφορίας για την ομιλία από τη θέα του κινούμενου στόματος [5].

Η καταγεγραμμένη διτροπική φύση της αντίληψης του λόγου οδήγησε πολλούς ερευνητές να συνδυάσουν τα κανάλια ακουστικής και οπτικής πληροφορίας για να δημιουργήσουν ένα πιο ισχυρό σύστημα αναγνώρισης ομιλίας. Η οπτικοακουστική αναγνώριση ομιλίας είναι η συγχώνευση της αυτόματης αναγνώρισης ομιλίας και των συστημάτων οπτικής αναγνώρισης ομιλίας. Υπάρχουν πολλοί καταγεγραμμένοι τρόποι για την πραγματοποίηση ενός τέτοιου έργου [6].

Η οπτικοακουστική αναγνώριση ομιλίας μπορεί να κατηγοριοποιηθεί σε δύο σημαντικές προσεγγίσεις ανάλογα σε ποιο επίπεδο της διαδικασίας εκπαίδευσης εκτελείται το έργο συνδυασμού:

- **Συνδυασμός χαρακτηριστικών** : είναι ο συνδυασμός κατά τον οποίο, το διάνυσμα οπτικού και ακουστικού καναλιού ενώνεται και τροφοδοτείται σε ένα ενιαίο μοντέλο ομιλίας.
- **Συνδυασμός απόφασης** : είναι ο συνδυασμός των εξόδων από τα δύο ξεχωριστά οπτικά και ακουστικά μοντέλα ομιλίας.

Για μια λεπτομερέστερη και εμπεριστατωμένη ανασκόπηση σχετικά με την οπτικοακουστική αυτόματη αναγνώριση ομιλίας, ο αναγνώστης καλείται να συμβουλευτεί τα εξής [6], [7].

1.3 Σκοπός της διπλωματικής

Ο σκοπός αυτή της διπλωματικής εργασίας είναι η προσπάθεια δημιουργίας ενός συστήματος το οποίο θα μπορεί να αξιοποιεί ακουστική, οπτική πληροφορία προερχόμενη από διαφορετικές οπτικές γωνίες για να πραγματοποιήσει αναγνώριση ομιλίας. Ένα τέτοιο σύστημα θα μπορούσε να φανεί ιδιαίτερα χρήσιμο σε περιβάλλοντα τα οποία είναι εφοδιασμένα από συστήματα πολλαπλών καμερών με στόχο την βελτίωση της αναγνώρισης ομιλίας. Για τον λόγο αυτό, αξιοποιήθηκε μια βάση δεδομένων όπου πρόσφερε τέτοιου είδους δεδομένα και αναπτύχθηκαν ένας ταξινομητής πέντε διαφορετικών όψεων, πέντε μοντέλα εξαγωγής οπτικής πληροφορίας για κάθε όψη, ένα βασικό μοντέλο ακουστικής αναγνώρισης ομιλίας τύπου γκαουσιανό μοντέλο μίξης/κρυφό Μαρκοβιανό μοντέλο (GMM/HMM), ένα ακουστικό μοντέλο αναγνώρισης ομιλίας HMM-DNN και άλλα πέντε μοντέλα HMM-DNN τροφοδοτούμενα από οπτική πληροφορία. Επίσης, έγινε μια προσπάθεια συνδυασμού αυτών των μοντέλων (οπτικού και ακουστικού) για κάθε όψη για την δημιουργία οπτικοακουστικών μοντέλων. Τα μοντέλα αυτά αξιολογήθηκαν επίσης εκτός από τα δεδομένα της αρχικής βάσης, σε δεδομένα στα οποία είχε γίνει προσθήκη λευκού θορύβου σε διάφορα επίπεδα.

1.4 Δομή της διπλωματικής

Η παρούσα διατριβή χωρίζεται σε επτά κεφάλαια και κάθε ένα από αυτά επικεντρώνεται σε ένα συγκεκριμένο θέμα.
Πιο συγκεκριμένα, το

- **Κεφάλαιο 2**: Θα παρουσιάσει τη βάση δεδομένων και τα εργαλεία λογισμικού που χρησιμοποιήθηκαν για την διεξαγωγή των πειραμάτων.

- **Κεφάλαιο 3** : Θα παρουσιάσει λεπτομερώς την διαδικασία εκπαίδευσης του βασικού μοντέλου ακουστικής αναγνώρισης ομιλίας, το οποίο αποτελεί ένα μοντέλο τύπου Γκαουσιανό μοντέλο μίξης/κρυφό Μαρκοβιανό μοντέλο.

- **Κεφάλαιο 4** : Θα παρουσιάσει τις ενέργειες που λήφθηκαν και τα μοντέλα που αναπτύχθηκαν, για την ταξινόμηση των πέντε όψεων και τα μοντέλα που

αναπτύχθηκαν για την (εξαγωγή χαρακτηριστικών)/(μείωση διαστάσεων) της οπτικής πληροφορίας για κάθε μία από τις πέντε όψεις χρησιμοποιώντας μοντέλα τύπου αυτόματου κωδικοποιητή.

- **Κεφάλαιο 5** : Θα παρουσιάσει τα μοντέλα που αναπτύχθηκαν τύπου HMM-DNN, βασιζόμενα στο μοντέλο τύπου Γκαουσιανό μοντέλο μίξης/κρυφό Μαρκοβιανό μοντέλο (GMM/HMM).

- **Κεφάλαιο 6**: Θα παρουσιάσει λεπτομερώς τα αποτελέσματα των πειραμάτων.

- **Κεφάλαιο 7** : Θα προσπαθήσει να προσφέρει μία περίληψη των προαναφερομένων και μια αναφορά για μελλοντική εργασία.

Κεφάλαιο 2

Βάση Δεδομένων / Εργαλεία Λογισμικού

Το συγκεκριμένο κεφάλαιο δίνει μια περιγραφή της βάσεως δεδομένων που χρησιμοποιήθηκε για την διεξαγωγή των πειραμάτων καθώς και των αντίστοιχων εργαλείων λογισμικού.

2.1 Βάση Δεδομένων

Οι οπτικοακουστικές βάσεις δεδομένων και πιο συγκεκριμένα οι διαθέσιμες στο κοινό με υψηλής ποιότητας δεδομένα ήταν περιορισμένες μέχρι πρόσφατα. Η πρόοδος στην τεχνολογία όσον αφορά την ικανότητα αποθήκευσης και την ικανότητα εγγραφής βίντεο κατέστησε τις οπτικοακουστικές βάσεις δεδομένων οικονομικά εφικτές.

Σχήμα 2.1: Πέντε όψεις από την
OuluVS2 βάση δεδομένων (Πηγή: [1])

Η βάση δεδομένων OuluVS2 δημιουργήθηκε το 2015 και περιέχει βίντεο από 53 ομιλητές που εκφωνούσαν διάφορους τύπους φράσεων, οι οποίοι καταγράφηκαν ταυτόχρονα από έξι κάμερες από πέντε διαφορετικές όψεις [1].

Πιο συγκεκριμένα, μεταξύ των 53 ατόμων που συμμετείχαν, υπήρχαν 40 άνδρες και 13 γυναίκες. Η ηχογράφηση έγινε σε συνήθη κατάσταση γραφείου με μικτό φωτισμό και πιθανούς ήχους περιβάλλοντος. Τα βίντεο δημιουργήθηκαν στην ίδια συνεδρία από διαφορετικές κάμερες οι οποίες ήταν συγχρονισμένες. Από τους αρχικούς 53 ομιλητές οι 52 από αυτούς ήταν αξιοποιήσιμοι για εκπαίδευση και αξιολόγηση. Από τους 52 συνολικά ομιλητές τόσο για τα αρχεία ήχου αλλά και τα βίντεο έγινε διαχωρισμός σε 40 ομιλητές για εκπαίδευση και 12 για αξιολόγηση.

Αν και η βάση OuluVS2 διαθέτει τρία είδη φράσεων, για τα πειράματα που πραγματοποιήθηκαν χρησιμοποιήθηκε τόσο για εκπαίδευση όσο και για αξιολόγηση μόνο η δεύτερη κατηγορία. Η κατηγορία αυτή, αποτελείται από δέκα καθημερινές φράσεις στην αγγλική γλώσσα οι οποίες επαναλαμβάνονται από κάθε ομιλητή τρεις φορές.

Φράσεις
Excuse me
Goodbye
Hello
How are you
Nice to meet you
See you
I am sorry
Thank you
Have a good time
You are welcome

Πίνακας 2.1: Φράσεις από την OuluVS2 βάση δεδομένων.

Οι συγκεκριμένες φράσεις προβάλλονται στον πίνακα 2.1. Πιο συγκεκριμένα, τα μοντέλα αναγνώρισης ομιλίας εκπαιδεύτηκαν με 1200 φράσεις και αξιολογήθηκαν αντιστοίχως σε 360 φράσεις.

2.2 Εργαλεία Λογισμικού

Κατά την διάρκεια πραγματοποίησης των πειραμάτων χρησιμοποιήθηκαν δύο ειδών εργαλεία λογισμικού:

Το Pytorch [8] αποτελεί σήμερα ένα από τα δύο δημοφιλέστερα εργαλεία βαθιάς μάθησης. Η κύρια εστίαση του PyTorch, όπως και πολλών άλλων αντίστοιχων σύγχρονων εργαλείων βαθιάς μάθησης, είναι η αυτόματη διαφοροποίηση συναρτήσεων απώλειας [8]. Το Pytorch χρησιμοποιήθηκε για την εκπαίδευση του αυτόματου ταξινομητή όψεων. Επίσης, χρησιμοποιήθηκε για την εκπαίδευση των πέντε αυτόματων κωδικοποιητών (autoencoders) για κάθε μία από τις πέντε όψεις.

Το Kaldi [9] αποτελεί ένα εργαλείο αυτόματης αναγνώρισης ομιλίας (ASR), το οποίο θα μπορούσε να χαρακτηριστεί ως τεχνολογία αιχμής και περιέχει σχεδόν οποιοδήποτε αλγόριθμο που χρησιμοποιείται σήμερα σε συστήματα ASR. Με την βοήθεια του Kaldi αναπτύχθηκε το βασικό μοντέλο ακουστικής αναγνώρισης τύπου γκαου-

οιανό μοντέλο μίξης/κρυφό Μαρκοβιανό μοντέλο (GMM/HMM), το ακουστικό μοντέλο τύπου HMM-DNN και τα αντίστοιχα μοντέλα οπτικής πληροφορίας. Στο Kaldi πραγματοποιήθηκε και ο συνδυασμός των προαναφερθέντων μοντέλων για την πραγματοποίηση οπτικοακουστικής αναγνώρισης ομιλίας.

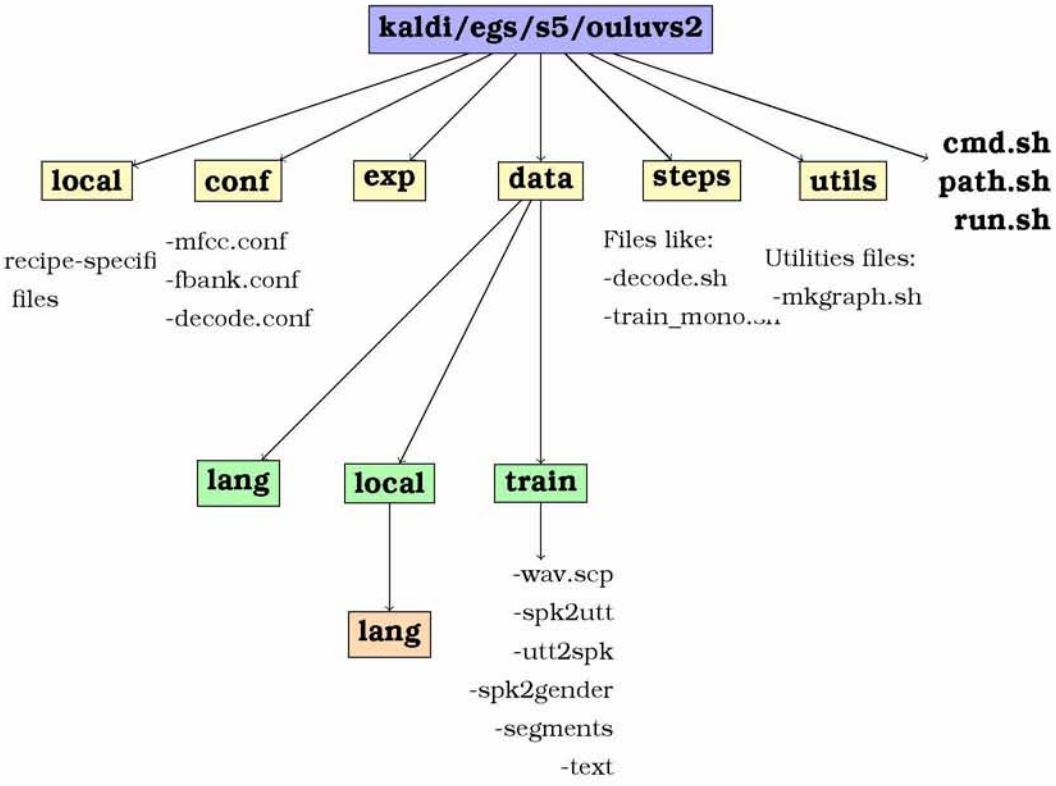
Κεφάλαιο 3

Κρυφό Μαρκοβιανό μοντέλο/ Γκαουσιανό μοντέλο μίξης

Αυτό το κεφάλαιο θα προσπαθήσει να περιγράψει όσο το δυνατόν λεπτομερέστερα τη διαδικασία που ακολουθήθηκε για την εκπαίδευση του βασικού ακουστικού μοντέλου αναγνώρισης ομιλίας. Το μοντέλο αυτό, όπως προαναφέρθηκε, δημιουργήθηκε με το εργαλείο λογισμικού Kaldi.

3.1 Δημιουργία προαπαιτούμενων αρχείων

Δομή Καταλόγων Πειράματος



Σχήμα 3.1: Δομή Καταλόγων Πειράματος

Προαπαιτούμενα αρχεία

Για να ξεκινήσει η διαδικασία εκπαίδευσης, πρέπει να δημιουργηθούν και να τοποθετηθούν ορισμένα αρχεία σε συγκεκριμένους φακέλους. Αυτά τα αρχεία θα αναφερθούν παρακάτω:

Αρχικά, το αρχείο **cmd.sh** του οποίου τα περιεχόμενα εμφανίζονται παρακάτω, περιέχει πληροφορίες σχετικά με τη διαμόρφωση υλικού (hardware) που χρησιμοποιούμε κατά τη διαδικασία εκπαίδευσης / αποκωδικοποίησης ορίζοντας ανάλογες μεταβλητές. Αυτό το αρχείο απαιτείται λόγω της δυνατότητας υποστήριξης από το

λογισμικό Kaldi εκπαίδευσης σε πολλές διαφορετικές πλατφόρμες υλικού, όπως για παράδειγμα ένα σύμπλεγμα διακομιστών (servers) που συνήθως χρησιμοποιείται από πανεπιστήμια.

```
train_cmd="utils/run.pl"
decode_cmd="utils/run.pl"
cuda_cmd="utils/run.pl"
```

Listing 3.1: **cmd.sh**

Για να μπορέσουμε να χρησιμοποιήσουμε τα δυαδικά/εκτελέσιμα (binary) εργαλεία του Kaldi, είναι συνετή και συνηθίζεται η δημιουργία ενός αρχείου επονομαζόμενου **path.sh**, το οποίο ορίζει στο περιβάλλον εκτέλεσης τη διαδρομή (path) προς τον πηγαίο κώδικα του, τα εκτελέσιμα και άλλα εργαλεία (tools).

```
export KALDI_ROOT=`pwd`/../../..
[ -f $KALDI_ROOT/tools/env.sh ] && . $KALDI_ROOT/tools/env.sh
export PATH=$PWD/utils/:$KALDI_ROOT/tools/openfst/bin:$PWD:$PATH
[ ! -f $KALDI_ROOT/tools/config/common_path.sh ] && echo >&2  \
"The standard file $KALDI_ROOT/tools/config/common_path.sh is not
   present -> Exit!" \
&& exit 1
. $KALDI_ROOT/tools/config/common_path.sh
export LC_ALL=C
```

Listing 3.2: **path.sh**

Το αρχείο **lexicon.txt** αποτελεί το φωνητικό μας λεξιλόγιο. Στην πρώτη στήλη του αρχείου βρίσκεται πάντα ο γραπτός κώδικας της πληροφορίας ακολουθούμενος από την φωνητική του αναπαράσταση. Το συγκεκριμένο λεξιλόγιο δημιουργήθηκε με την βοήθεια του CMU pronunciation dictionary [10].

```
HOW HH AW
THANK TH AE NG K
TIME  T AY M
ME  M IY
```

```
A AH
A EY
HELLO HH AH L OW
HELLO HH EH L OW
GOOD  G UH D
GOOD  G IH D
WELCOME W EH L K AH M
GOODBYE G UH D B AY
EXCUSE  IH K S K Y UW S
EXCUSE  IH K S K Y UW Z
TO  T UW
TO  T IH
TO  T AH
MEET  M IY T
AM  AE M
AM  EY EH M
I AY
NICE  N AY S
NICE  N IY S
HAVE  HH AE V
ARE AA R
ARE ER
SEE S IY
SORRY S AA R IY
YOU Y UW
!SIL SIL
<UNK> SPN
```

Listing 3.3: **lexicon.txt**

Το αρχείο **silence_phones.txt** περιλαμβάνει τα φωνήματα σιωπής.

```
SIL
SPN
```

Listing 3.4: **silence_phones.txt**

Το αρχείο **optional_silence.txt** περιλαμβάνει το φώνημα της προαιρετικής σιωπής.

```
SIL
```

Listing 3.5: **optional_silence.txt**

Το αρχείο **nonsilence_phones.txt** περιλαμβάνει τα φωνήματα που προφέρονται.

```
HH
IH
ER
AH
IY
TH
UH
K
L
EY
B
R
T
UW
W
N
AY
V
Z
AA
NG
AW
Y
S
G
EH
AE
OW
M
D
```

Listing 3.6: **nonsilence_phones.txt**

Το αρχείο **corpus.txt** περιέχει όλες τις μοναδικές δηλώσεις/φράσεις των δεδομένων μας.

```
EXCUSE ME
GOODBYE
HELLO
HOW ARE YOU
NICE TO MEET YOU
SEE YOU
I AM SORRY
THANK YOU
HAVE A GOOD TIME
YOU ARE WELCOME
```

Listing 3.7: **corpus.txt**

Το αρχείο **dict.txt** περιέχει όλες τις μοναδικές λέξεις των δεδομένων μας.

```
THANK
GOOD
TO
HAVE
TIME
NICE
WELCOME
ME
SEE
YOU
GOODBYE
HOW
ARE
AM
EXCUSE
A
MEET
SORRY
HELLO
I
```

Listing 3.8: **dict.txt**

Το αρχείο **mfcc.conf** περιέχει παραμέτρους για την δημιουργία των ακουστικών χαρακτηριστικών μας που είναι φασματικοί συντελεστές συχνότητας Mel [11]. Πιο

συγκεκριμένα, επιλέχθηκε:

- να μην γίνει προσθήκη στο διάνυσμα ακουστικού χαρακτηριστικού η ενέργεια.
- να ορισθεί ως συχνότητα δειγματοληψίας τα 48 kHz αντικατοπτρίζοντας την ηχογράφηση των αρχείων ήχου των δεδομένων μας.
- να αφεθεί η προεπιλογή για συνάρτηση παραθύρου, η οποία ειναι τύπου Hamming.

```
--use-energy=false
--sample-frequency=48000
```

Listing 3.9: **mfcc.conf**

Το αρχείο **decode.conf** περιέχει πληροφορίες σχετικά με τη διαδικασία αποκωδικοποίησης, όπως το μέγεθος των δεσμών (beams) αποκωδικοποίησης και πλέγματος. Η διαδικασία αποκωδικοποίησης που χρησιμοποιείται θα μπορούσε να χαρακτηριστεί ως Viterbi beam search. Σαν πλέγμα (lattice), ορίζεται μια αναπαράσταση των εναλλακτικών ακολουθιών λέξεων που είναι αρκετά πιθανές για μια συγκεκριμένη φράση [12].

```
first_beam=10.0
beam=13.0
lattice_beam=6.0
```

Listing 3.10: **decode.conf**

Άλλα αρχεία που πρέπει να οριστούν στον φάκελο **data/train** και πρόκειται να αναφερθούν σύντομα είναι τα εξής:

- **spk2gender** : Αυτό το αρχείο ενημερώνει για το φύλο των ομιλητών.
- **wav.scp** : Αυτό το αρχείο συνδέει κάθε φράση (πρόταση που λέγεται από ένα άτομο κατά τη διάρκεια συγκεκριμένης περιόδου ηχογράφησης) με ένα αρχείο ήχου που σχετίζεται με αυτή τη φράση.
- **text** : Αυτό το αρχείο περιέχει κάθε φράση που ταιριάζει με την μεταγραφή του κειμένου.

- **utt2spk** : Αυτό το αρχείο αναφέρει ποια φράση ανήκει σε κάθε ομιλητή.

Τα παραπάνω αρχεία πρέπει να δημιουργούνται μοναδικά και να τοποθετούνται σε κάθε αντίστοιχο υποκατάλογο του **data** (π.χ. test, dev).

3.2 Εκπαίδευση μοντέλου Γκαουσιανό μοντέλο μίξης /Κρυφό Μαρκοβιανό μοντέλο

Η διαδικασία εκπαίδευσης του μοντέλου πραγματοποιείται σε ένα αρχείο το οποίο συνήθως ονομάζεται **run.sh**.

Το πρώτο βήμα συνήθως αποτελείται από την δημιουργία των ακουστικών χαρακτηριστικών που θα χρησιμοποιηθούν στην εκπαίδευση του μοντέλου μας, όπου στην συγκεκριμένη περίπτωση είναι φασματικοί συντελεστές συχνότητας Mel.

```
mfcc=/feat/mfcc #location of feature storage
nj=4 # number of jobs/threads
steps/make_mfcc.sh --nj $nj --cmd "$train_cmd" \
data/$x exp/make_mfcc/$x $mfcc
```

Listing 3.11: **run.sh**

Στην συνέχεια, εφαρμόζεται φασματική κανονικοποίηση μέσου όρου και διακύμανσης [13].

```
steps/compute_cmvn_stats.sh data/$x exp/make_mfcc/$x $mfcc || exit 1
```

Listing 3.12: **run.sh**

Απαραίτητη είναι και η δημιουργία του μοντέλου γλώσσας (language model). Το μοντέλο γλώσσας που χρησιμοποιήθηκε είναι τύπου unigram, και αξιοποιώντας το αρχείο corpus.txt που ορίσαμε νωρίτερα, ορίζεται στην ευρέως γνωστή μορφή αρχείου ARPA. Στη συνέχεια, το μοντέλο γλώσσας μετατρέπεται σε έναν βεβαρημένο γράφο επονομαζόμενο ως G.fst.

```
local=data/local
lm_order=1 #unigram model
ngram-count -order $lm_order -write-vocab $local/tmp/vocab-full.txt \
-wbdiscount -text $local/corpus.txt -lm $local/tmp/lm.arpa || exit 1

lang=data/lang
arpa2fst --disambig-symbol=#0 --read-symbol-table=$lang/words.txt  \
$local/tmp/lm.arpa > $lang/G.fst
```

Listing 3.13: **run.sh**

Έπειτα, ξεκινά η διαδικασία εκπαίδευσης των ακουστικών μοντέλων. Αυτά, μπορούν να χαρακτηριστούν από επαναλαμβανόμενα βήματα εκπαίδευσης και χρονικής ευθυγράμμισης μεταξύ των αρχείων ήχου και των γραπτών φράσεων. Τα πρώτα μοντέλα που εκπαιδεύτηκαν ήταν τα μονοφωνικά.

```
steps/train_mono.sh --nj $nj --cmd "$train_cmd" \
    data/train $lang exp/mono0a



steps/align_si.sh --nj $nj --cmd "$train_cmd" \
    data/train $lang exp/mono0a exp/mono0a_ali
```

Listing 3.14: **run.sh**

Στις επόμενες δύο εκπαιδεύσεις, εκπαιδεύονται τα μοντέλα τριφώνου κατά τα οποία τα φωνήματα αντιπροσωπεύονται στο πλαίσιο δύο άλλων φωνημάτων (αριστερό και δεξί). Τα συγκεκριμένα είδη μοντέλων λαμβάνουν ως είσοδο για την εκπαίδευσή τους και τις αριθμητικές εκτιμήσεις της πρώτης και δεύτερης παραγώγου των αρχικών χαρακτηριστικών. Για τη πρώτη διαδικασία εκπαίδευσης χρησιμοποιήθηκαν 600 φύλλα (leaves) και 2500 κανονικές κατανομές (gaussians).

```
steps/train_deltas.sh --cmd "$train_cmd" \
    600 2500 data/train $lang exp/mono0a_ali exp/tri1

steps/align_si.sh --nj $nj --cmd "$train_cmd" \
```

```
    data/train $lang exp/tri1 exp/tri1_ali
```

Listing 3.15: **run.sh**

Για τη δεύτερη διαδικασία εκπαίδευσης χρησιμοποιήθηκαν 800 φύλλα (leaves) και 3500 κανονικές κατανομές (gaussians).

```
steps/train_deltas.sh --cmd "$train_cmd" \
   800 3500 data/train $lang exp/tri1_ali exp/tri2a || exit 1;



steps/align_si.sh --nj $nj --cmd "$train_cmd" \
  data/train $lang exp/tri2a exp/tri2a_ali
```

Listing 3.16: **run.sh**

Για τη διαδικασία εκπαίδευσης **lda-mllt** επιλέχθηκε ένα εύρος -5/+5 συνδυασμού πλαισίων. Επίσης, τα φύλλα και ο αριθμός των κανονικών κατανομών ορίστηκε σε 1000 και 6500 αντιστοίχως.

```
steps/train_lda_mllt.sh --cmd "$train_cmd" \
    --splice-opts "--left-context=5 --right-context=5" \
    1000 6500 data/train $lang exp/tri2a_ali exp/tri2b || exit 1;


steps/align_si.sh --cmd "$train_cmd" --nj $nj \
     data/train $lang exp/tri2b exp/tri2b_ali
```

Listing 3.17: **run.sh**

Καταλήγοντας, για την διαδικασία εκπαίδευσης προσαμογής ομιλητή (**sat**), οι παράμετροι των φύλλων και κανονικών κατανομών ορίστηκαν σε 1500 και 6500 αντιστοίχως. Η συγκεκριμένη διαδικασία εκπαίδευσης εκτελεί ομαλοποίηση ομιλητή και θορύβου προσαρμόζοντας σε κάθε συγκεκριμένο ομιλητή με συγκεκριμένο μετασχηματισμό δεδομένων.

```
steps/train_sat.sh \
    1500 6500 data/train $lang exp/tri2b_ali exp/tri3b || exit 1;



steps/align_fmllr.sh --nj $nj --cmd "$train_cmd" \
    data/train $lang exp/tri3b exp/tri3b_ali
```

Listing 3.18: **run.sh**

3.3 Αποκωδικοποίηση/Αξιολόγηση μοντέλων

Μετά τη διαδικασία εκπαίδευσης και αφού έχουμε δημιουργήσει τα γραφήματα μας, μπορούμε να προχωρήσουμε στην αποκωδικοποίηση και αξιολόγηση των μοντέλων μας, όπως φαίνεται παρακάτω.

```
decode_list="test"
model_list="mono0a tri1 tri2a tri2b tri3b"
for mdl in $model_list; do
    for x in $decode_list; do
      echo "Evaluating model $mdl for $x"

      if [ ! -f exp/$mdl/graph/HCLG.fst ]; then
        utils/mkgraph.sh $lang exp/$mdl exp/$mdl/graph || exit 1;
      fi

      if [ "$mdl" = tri3b ]; then
        steps/decode_fmllr.sh --cmd "$decode_cmd" --nj $nj   \
          exp/$mdl/graph data/$x exp/$mdl/decode_$x
      else
        steps/decode.sh --cmd "$decode_cmd" --nj $nj   \
          exp/$mdl/graph data/$x exp/$mdl/decode_$x  #
      fi
      wait
      echo ""
      echo "$x set decoded!"
      local/score.sh data/$x exp/$mdl/graph exp/$mdl/decode_$x
    done
done
```

Listing 3.19: **run.sh**

Κεφάλαιο 4

Οπτική επεξεργασία

Το σύνολο των μοντέλων που αναφέρονται στο συγκεκριμένο κεφάλαιο αναπτύχθηκαν με το εργαλείο λογιομικού PyTorch.

4.1 Ταξινόμηση όψεων

Σαν ένα πρώτο βήμα δημιουργίας ενός συστήματος το οποίο θα μπορεί να αξιοποιεί οπτική πληροφορία από διαφορετικές οπτικές γωνίες αποφασίστηκε η δημιουργία ενός αυτόματου ταξινομητή όψεων.

Προ-επεξεργασία δεδομένων

Για την δυνατότητα ταξινόμησης όψεων έπρεπε να πραγματοποιηθεί μια προ-επεξεργασία στα πλαίσια που έγιναν εξαγωγή από τα βίντεο της βάσεως δεδομένων. Το πρώτο βήμα προ-επεξεργασίας αποτέλεσε η αλλαγή μεγέθους των διαστάσεων των πέντε όψεων, των οποίων οι αρχικές διαστάσεις ήταν εντελώς διαφορετικές. Η κοινή νέα διάσταση που ορίστηκε ήταν 31x41 pixels. Έπειτα, τα πλαίσια εικόνων μετατρέπονται σε κλίμακα του γκρίζου (grayscaled) μειώνοντας επιπλέον την διάσταση του χρώματος από τρία σε ένα. Σαν τελικό βήμα προ-επεξεργασίας για τα δεδομένα εισόδου του ταξινομητή όψεων, πραγματοποιήθηκε μια ομαλοποίηση των τιμών των πλαισίων φέρνοντας το εύρος των τιμών τους στο διάστημα [-1,1]. Το αποτέλεσμα όλων αυτών των μεταοχηματιομών αποτέλεσε την είσοδο στον ταξινομητή όψεων του οποίου η αρχιτεκτονική θα περιγραφεί παρακάτω.

Αρχιτεκτονική και εκπαίδευση

Η αρχιτεκτονική του συγκεκριμένου ταξινομητή καθορίζεται ως εξής:

- Ένα αρχικό συνελικτικό επίπεδο δύο διαστάσεων μεγέθους 32 και φίλτρα μεγέθους 5x5.
- Στην συνέχεια εφαρμόζεται μια κανονικοποίηση παρτίδας (Batchnorm) η οποία ακολουθείται από μια συνάρτηση ενεργοποίησης ReLU.
- Ένα δεύτερο συνελικτικό επίπεδο δύο διαστάσεων μεγέθους 64 και φίλτρα μεγέθους 5x5.
- Το δεύτερο συνελικτικό επίπεδο το ακολουθεί άλλη μία κανονικοποίηση παρτίδας (Batchnorm), η οποία ακολουθείται πάλι από μια συνάρτηση ενεργοποίησης ReLU.
- Ένα τρίτο συνελικτικό επίπεδο δύο διαστάσεων μεγέθους 128 και φίλτρα μεγέθους 5x5, το οποίο πάλι ακολουθείται από μια κανονικοποίηση παρτίδας και συνάρτηση ενεργοποίησης ReLU.
- Στην συνέχεια εφαρμόζεται ένα γραμμικό επίπεδο συνολικού μεγέθους 19*29*128 (70528).
- Στο τελευταίο επίπεδο εφαρμόζεται μια συνάρτηση ενεργοποίησης τύπου log-softmax μεγέθους πέντε όσο δηλαδή και ο αριθμός των συνολικών όψεων.

Για τη διαδικασία εκπαίδευσης του συγκεκριμένου μοντέλου χρησιμοποιήθηκε σαν συνάρτηση απώλειας, η συνάρτηση απώλειας αρνητικής λογαριθμικής πιθανοφάνειας.

Οι υπερπαράμετροι που χρησιμοποιήθηκαν κατά τη διάρκεια εκπαίδευσης του ταξινομητή όψεων είναι οι εξής:

- ο αριθμός εποχών ορίστηκε ίσος με 15.
- σαν optimizer επιλέχθηκε ο Adam.
- το learning rate ορίστηκε ίσο με 1e-3.
- το weight decay ορίστηκε ίσο με 1e-5.

- το μέγεθος σακιδίου ίσο με 512.

4.2 Αυτόματοι κωδικοποιητές

Για όλους τους αυτόματους κωδικοποιητές επιλέχθηκε ακριβώς η ίδια αρχιτεκτονική και ακολουθήθηκε η ίδια ακριβώς διαδικασία εκπαίδευσης.

Προ-επεξεργασία δεδομένων

Η προ-επεξεργασία των αυτόματων κωδικοποιητών θα μπορούσε να χαρακτηριστεί τελείως ίδια με αυτή που ακολουθήθηκε με την προ-επεξεργασία του αυτόματου ταξινομητή με μία μόνο σημαντική διαφορά. Η διαφορά έγκειται στον διαφορετικό τρόπο αλλαγής μεγέθους των διαστάσεων των πέντε όψεων. Πιο συγκεκριμένα, έγιναν οι εξής αλλαγές διαστάσεων:

- Για την πρόσοψη το πλάτος και ύψος της εικόνας ορίστηκαν ίσα με 53 και 31 pixels αντίστοιχα.
- Για την όψη 30° το πλάτος και ύψος της εικόνας ορίστηκαν ίσα με 45 και 30 pixels αντίστοιχα.
- Για την όψη 45° το πλάτος και ύψος της εικόνας ορίστηκαν ίσα με 45 και 31 pixels αντίστοιχα.
- Για την όψη 60° το πλάτος και ύψος της εικόνας ορίστηκαν ίσα με 29 και 37 pixels αντίστοιχα.
- Για την όψη 90° το πλάτος και ύψος της εικόνας ορίστηκαν ίσα με 24 και 33 pixels αντίστοιχα.

Αρχιτεκτονική και εκπαίδευση

Ο αυτόματος κωδικοποιητής που αναπτύχθηκε ήταν ένας πλήρως διασυνδεδεμένος κωδικοποιητής. Η αρχιτεκτονική του αυτόματου κωδικοποιητή μπορεί να χαρακτηριστεί από τα εξής:

- Το πρώτο επίπεδο αποτελεί ένα γραμμικό/πλήρως διασυνδεδεμένο επίπεδο με είσοδο ίση με το γινόμενο των διαστάσεων πλάτους και ύψους των εισόδων της κάθε όψης και έξοδο ίση με 1000, το οποίο ακολουθείται από μια συνάρτηση ενεργοποίησης ReLU.
- Το δεύτερο επίπεδο αποτελεί πάλι ένα γραμμικό/πλήρως διασυνδεδεμένο επίπεδο μεγέθους 1000 και έξοδο 500, το οποίο και αυτό ακολουθείται από μια συνάρτηση ενεργοποίησης ReLU.
- Το τρίτο επίπεδο αποτελεί πάλι ένα γραμμικό/πλήρως διασυνδεδεμένο επίπεδο με είσοδο 500 και έξοδο 50, το οποίο και αυτό ακολουθείται από μια συνάρτηση ενεργοποίησης ReLU.
- Το τέταρτο επίπεδο αποτελεί ένα γραμμικό/πλήρως διασυνδεδεμένο επίπεδο με είσοδο 50 και έξοδο 500, το οποίο και αυτό ακολουθείται από μια συνάρτηση ενεργοποίησης ReLU.
- Το πέμπτο επίπεδο αποτελεί ένα γραμμικό/πλήρως διασυνδεδεμένο επίπεδο με είσοδο 500 και έξοδο 1000, το οποίο και αυτό ακολουθείται από μια συνάρτηση ενεργοποίησης ReLU.
- Το τελευταίο επίπεδο αποτελεί ένα γραμμικό/πλήρως διασυνδεδεμένο επίπεδο με είσοδο 1000 και έξοδο ίση με το γινόμενο των διαστάσεων πλάτους και ύψους των εισόδων της κάθε όψης. Στο τελευταίο επίπεδο εφαρμόζεται η συνάρτηση ενεργοποίησης Tanh.

Για τη διαδικασία εκπαίδευσης του συγκεκριμένου μοντέλου χρησιμοποιήθηκε σαν συνάρτηση απώλειας, η απώλεια του μέσου τετραγωνικού σφάλματος.

Οι υπερπαράμετροι που χρησιμοποιήθηκαν κατά τη διάρκεια εκπαίδευσης των αυτόματων κωδικοποιητών είναι οι εξής:

- ο αριθμός εποχών ορίστηκε ίσος με 25.
- σαν optimizer ορίστηκε ο Adam.
- το learning rate ορίστηκε ίσο με 1e-3.
- το weight decay ορίστηκε ίσο με 1e-5.
- το μέγεθος σακιδίου ίσο με 128.

Μετεπεξεργασία

Μετά την εκπαίδευση των μοντέλων αυτόματων κωδικοποιητών, προβαίνουμε σε κάποιες ενέργειες για να δημιουργήσουμε τα τελικά χαρακτηριστικά οπτικής πληροφορίας. Αρχικά εξάγουμε πλαίσια εικόνων από τα βίντεο της βάσεως δεδομένων με ένα μεγαλύτερο ρυθμό δειγματοληψίας από τον αρχικό, της τάξεως των 100 Hz. Ο ρυθμός αυτός επιλέγεται με αυτό τον τρόπο έτσι ώστε να είναι ίδιος με αυτό με τον οποίο δημιουργήθηκαν τα ακουστικά μας χαρακτηριστικά με τα οποία εκπαιδεύεται το αντίστοιχο μοντέλο ακουστικής αναγνώρισης ομιλίας. Στην συνέχεια, τα νέα πλαίσια εικόνων που δημιουργούνται περνούν μέσα από τα μοντέλα των αυτόματων κωδικοποιητών και εξάγονται οι τιμές του ενδιάμεσου επιπέδου, το οποίο αποτελεί ένα διάνυσμα 50 τιμών, δημιουργώντας μια νέα αναπαράσταση των χαρακτηριστικών οπτικής πληροφορίας. Τα σύνολο αυτών των δεδομένων, εφόσον μετατραπούν σε μορφή πληροφορίας συμβατή με το εργαλείο λογισμικού Kaldi, θα χρησιμοποιηθούν για την εκπαίδευση των μοντέλων οπτικής αναγνώρισης ομιλίας.

Αναφορικά, τα λογισμικά που χρησιμοποιήθηκαν για τις προαναφερθέντες διαδικασίες είναι τα εξής:

- Για την διαδικασία εξαγωγής αλλά και υπερδειγματοληψίας των πλαισίων εικόνων από τα βίντεο της βάσεως δεδομένων χρησιμοποιήθηκε το εργαλείο λογισμικού ffmpeg [14].
- Για την διαδικασία μετατροπής των δεδομένων οπτικής πληροφορίας σε μορφή συμβατή με το εργαλείο λογισμικού Kaldi, χρησιμοποιήθηκε η βιβλιοθήκη kaldiio [15] της προγραμματιστικής γλώσσας Python.

Κεφάλαιο 5

Υβριδική μοντελοποίηση για οπτικοακουστική αναγνώριση ομιλίας

Σε αυτό το κεφάλαιο θα αναφερθούν οι αρχιτεκτονικές και τα βήματα που ακολουθήθηκαν για την εκπαίδευση των τελικών μοντέλων ακουστικής, οπτικής και οπτικοακουστικής ομιλίας με το εργαλείο λογισμικού Kaldi.

5.1 Ακουστικό μοντέλο

Αρχικά, το συγκεκριμένο μοντέλο έγινε εκπαίδευση έχοντας σαν είσοδο ακουστικά χαρακτηριστικά τύπου 40 FBANK.

Η αρχιτεκτονική του συγκεκριμένου μοντέλου καθορίζεται ως εξής:

- Ένα αρχικό επίπεδο εισόδου το οποίο συνενώνει τις εισόδους από την -4 έως +4 χρονική στιγμή.
- Ένα επίπεδο tdnn μεγέθους 150 με συνάρτηση ενεργοποίησης ReLU.
- Ένα δεύτερο επίπεδο tdnn μεγέθους 150 με είσοδο την -1, 0 και +1 χρονική στιγμή. Κι αυτό το επίπεδο το ακολουθεί μια συνάρτηση ενεργοποίησης ReLU.
- Ένα τρίτο επίπεδο tdnn μεγέθους 150 με είσοδο την -3, 0 και +3 χρονική

στιγμή. Κι αυτό το επίπεδο το ακολουθεί μια συνάρτηση ενεργοποίησης ReLU.

- Ένα επίπεδο LSTM μεγέθους 520.
- Ένα επίπεδο logsoftmax μεγέθους 208 ίσο δηλαδή με τον αριθμό των context-dependent HMM states.

Το συγκεκριμένο μοντέλο εκπαιδεύτηκε για 16 εποχές.

5.2 Οπτικό μοντέλο

Το οπτικό μοντέλο λαμβάνει σαν είσοδο το διάνυσμα 50 τιμών που έχει δημιουργηθεί από τον αντίστοιχο αυτόματο κωδικοποιητή.

Πιο συγκεκριμένα, η αρχιτεκτονική του συγκεκριμένου μοντέλου καθορίζεται ως εξής:

- Ένα αρχικό επίπεδο εισόδου το οποίο συνενώνει τις εισόδους από την -1 έως +1 χρονική στιγμή.
- Ένα επίπεδο tdnn μεγέθους 150 ακολουθούμενο από συνάρτηση ενεργοποίησης ReLU και κανονικοποίηση παρτίδας.
- Ένα δεύτερο επίπεδο tdnn μεγέθους 150 με είσοδο την -1, 0 και +1 χρονική στιγμή. Κι αυτό το επίπεδο το ακολουθεί μια συνάρτηση ενεργοποίησης ReLU και κανονικοποίηση παρτίδας.
- Ένα τρίτο επίπεδο tdnn μεγέθους 150 με είσοδο την -3, 0 και +3 χρονική στιγμή. Κι αυτό το επίπεδο το ακολουθεί μια συνάρτηση ενεργοποίησης ReLU και κανονικοποίηση παρτίδας.
- Ένα επίπεδο attention με 450 κλειδιά (keys) και 150 τιμές (values). Το επίπεδο αυτό ακολουθείται επίσης από συνάρτηση ενεργοποίησης ReLU και κανονικοποίηση παρτίδας.
- Ένα επίπεδο LSTM μεγέθους 520.
- Ένα επίπεδο logsoftmax μεγέθους 208 ίσο δηλαδή με τον αριθμό των context-dependent HMM states.

Το συγκεριμένο μοντέλο εκπαιδεύτηκε για 14 εποχές.

5.3 Οπτικοακουστική αναγνώριση ομιλίας

Για την πραγματοποίηση οπτικοακουστικής αναγνώρισης ομιλίας ακολουθήθηκε η διαδικασία συνδυασμού απόφασης (decision fusion). Αντίστοιχες μέθοδοι έχουν ήδη αναφερθεί από [16], [17], [7]. Πιο συγκεκριμένα, για κάθε όψη και για κάθε αναλογία θορύβου σήματος έγινε μια ανίχνευση πλέγματος (grid search) για την εύρεση ενός βάρους alpha και αντιστοίχως (1- alpha) για τον συνδυασμό των μοντέλων ακουστικής και οπτικής πληροφορίας. Το βήμα της ανίχνευσης πλέγματος (grid search) ορίστηκε σε τιμή ίση με 0.05.

Κεφάλαιο 6

Πειραματικά Αποτελέσματα

Σε αυτό το κεφάλαιο θα προβληθούν και θα συζητηθούν λεπτομερώς τα αποτελέσματα των πειραμάτων που περιγράφηκαν στα προηγούμενα κεφάλαια.

6.1 Γκαουσιανό Μοντέλο Μίξης/Κρυφό Μαρκοβιανό Μοντέλο

Model	**SER(%)**	**WER(%)**
Mono	24.72	13.89
+ Tri1	13.61	6.33
+ Tri2	13.33	6
+ *LDA-MLLT*	10.56	4.67
+ *SAT-fMLLR*	9.72	4.33

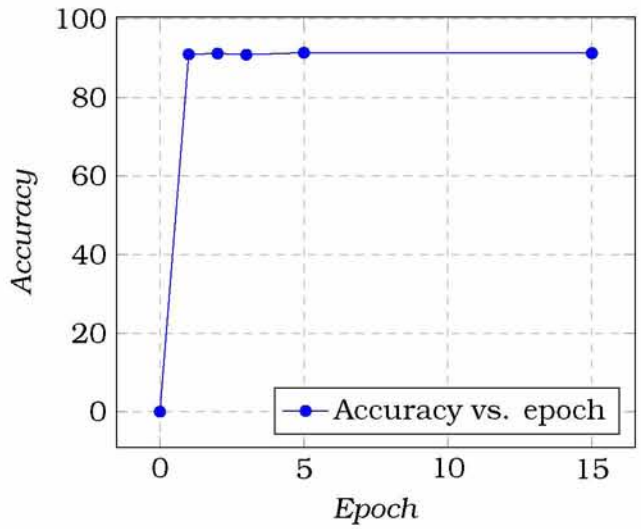
Πίνακας 6.1: Αποτελέσματα μοντέλων τύπου Γκαουσιανό Μοντέλο Μίξης/Κρυφό Μαρκοβιανό Μοντέλο

Αυτό που μπορούμε να παρατηρήσουμε από τον πίνακα 6.1, στον οποίο φαίνονται τα αποτελέσματα αναγνώρισης φράσεων και λέξεων, είναι ότι και τα δύο μεγέθη παρουσιάζουν την ίδια συμπεριφορά ως προς τον τύπο μοντέλου και την μείωση του αντίστοιχου ποσοστού λάθους. Οι σημαντικότερες μειώσεις του ποσοστού

λάθους πραγματοποιήθηκαν από την πρώτη εκπαίδευση των μοντέλων τριφώνου (από 24.72% σε 13.61%) και από την εκπαίδευση τύπου lda-mllt (από 13.33% σε 10.56%).

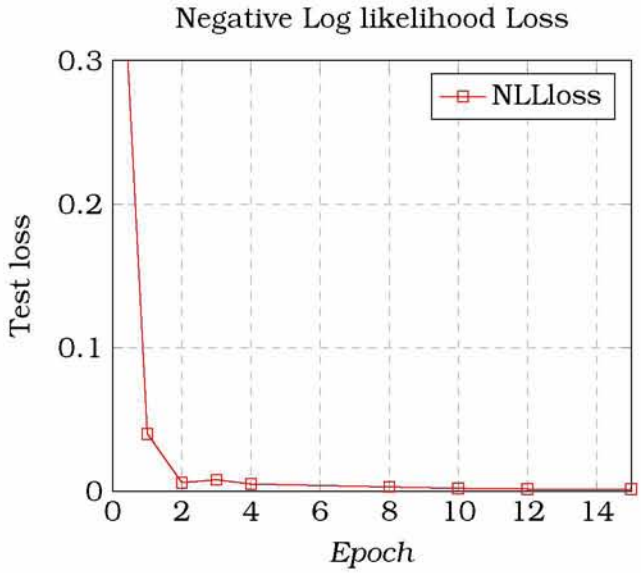
6.2 Αυτόματος ταξινομητής όψεων

Ο ταξινομητής όψεων εκπαιδεύτηκε με περίπου εκατό χιλιάδες φωτογραφίες οι οποίες είχαν γίνει εξαγωγή από τα βίντεο κάθε όψης. Από ότι μπορούμε να παρατηρήσουμε από τα σχήματα 6.1-6.2, όπου παρουσιάζεται η απόδοση και η συνάρτηση απώλειας κατά την διάρκεια της εκπαίδευσης, ότι δύο επαναλήψεις ήταν αρκετές για να φτάσει σχεδόν στην μέγιστη/ελάχιστη τιμή.



Σχήμα 6.1: Απόδοση του ταξινομητή όψεων στο σύνολο του σετ δοκιμής

Πιο συγκεκριμένα, η απόδοση του αυτόματη ταξινομητή όψεων κυμαινόταν για τις επόμενες επαναλήψεις γύρω στο *91*%.



Σχήμα 6.2: Αρνητική λογαριθμική απώλεια πιθανοφάνειας για τα δεδομένα αξιολόγησης σε συνάρτηση με το χρόνο επαναλήψεων

Για να εξομοιωθεί σε πιο κοντινά σενάρια στην πραγματικότητα η απόδοση του αυτόματου ταξινομητή όψεων, η απόδοση του δοκιμάστηκε επιπλέον σε σετ που δημιουργήθηκε μόνο από την εξαγωγή του πρώτου πλαισίου των βίντεο του αρχικού σετ δοκιμής. Τα αποτελέσματα της ταξινόμησης φαίνονται στο πίνακα 6.2.

View	**frontal**	**30°**	**45°**	**60°**	**profile**
frontal	357	3	0	0	0
30°	5	340	15	0	0
45°	0	9	270	131	0
60°	0	0	14	341	5
profile	0	0	2	6	352

Πίνακας 6.2: Μητρώο σύγχυσης για τον αυτόματο ταξινομητή όψεων

Από το συγκεκριμένο μητρώο σύγχυσης παρατηρούμε ότι υπάρχει ένα μεγάλο ποσοστό λάθους ταξινόμησης μεταξύ των όψεων 45° και 60°. Λάθος ταξινομήσεις εμφανίζονται και στις άλλες κλάσεις αλλά σε μικρότερο βαθμό.

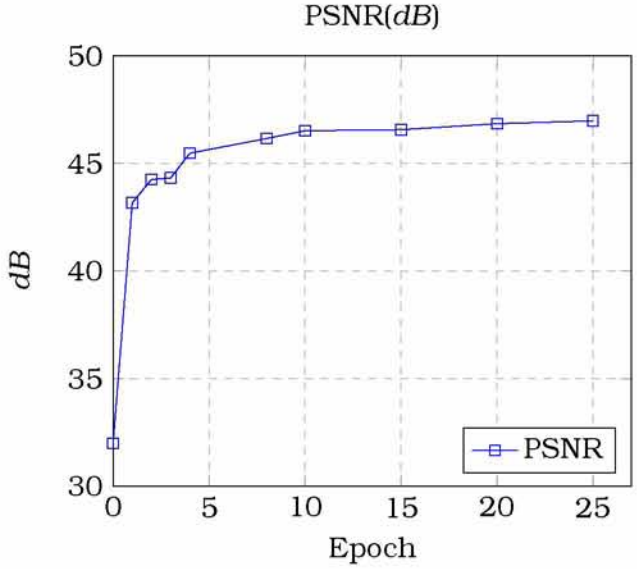
6.3 Αυτόματοι κωδικοποιητές

Τα αποτελέσματα που παρουσιάζονται στα σχήματα 6.3 - 6.7, εκφράζουν τις τιμές του μέγιστου σήματος προς το θόρυβο (PSNR). Ο λόγος μέγιστου σήματος προς θόρυβο μπορεί να εκφραστεί σχετικά με το μέσο τετραγωνικό σφάλμα με την ακόλουθη εξίσωση:

$$PSNR = 10\log_{10}(\frac{R}{MSE})$$

όπου το R εξαρτάται από τη μορφή της εικόνας (255 for 8-bit precision).

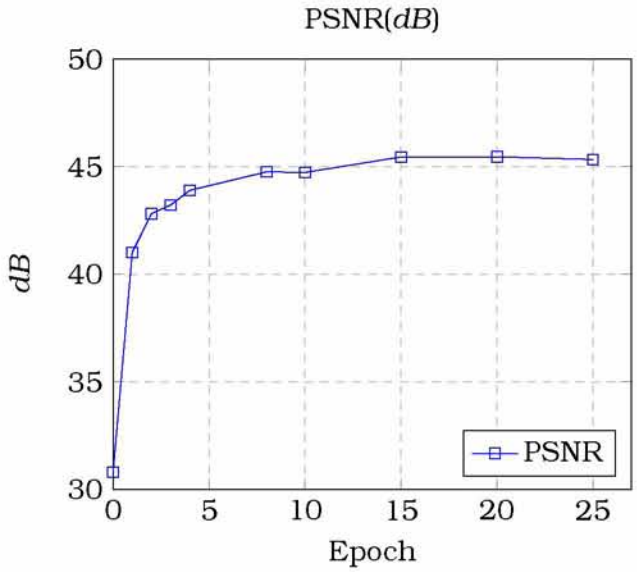
Αυτό που μπορούμε να παρατηρήσουμε από όλα τα σχήματα (6.3 - 6.7) που αφορούν την εκπαίδευση των αυτόματων κωδικοποιητών, είναι ότι παρουσιάζουν σχεδόν την ίδια συμπεριφορά με μόνη διαφορά τις απόλυτες τιμές που πετυχαίνουν.



Σχήμα 6.3: Αναλογία μέγιστου σήματος προς θόρυβο του σετ δοκιμής για τον αυτόματο κωδικοποιητή πρόσοψης

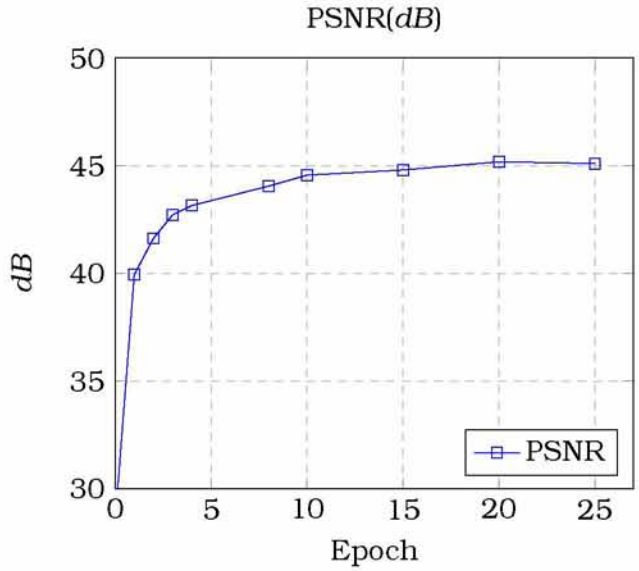
Πιο συγκεκριμένα, στο σχήμα 6.3 μπορούμε να παρατηρήσουμε ότι η ραγδαία αύξηση της ποσότητας πραγματοποιήθηκε τις πρώτες δύο εποχές. Έπειτα, στις επόμενες εποχές εκπαίδευσης παρουσιάστηκε μια σταθεροποίηση στα 46 dB. Η μέγιστη τιμή

που οημειώθηκε ήταν 46.97 dB.



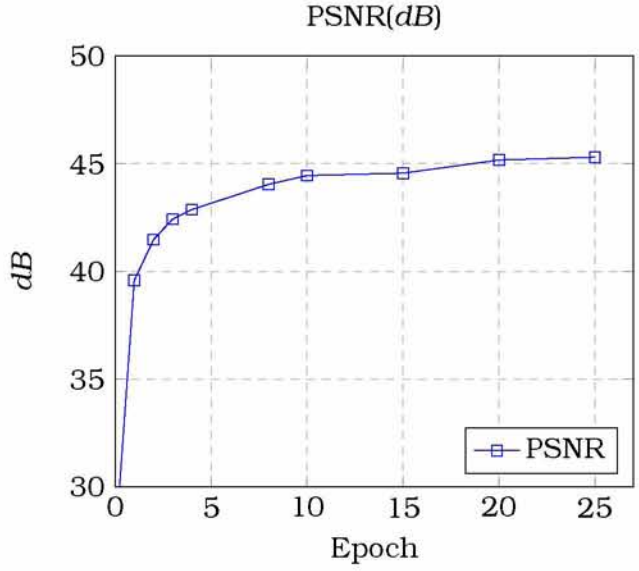
Σχήμα 6.4: Αναλογία μέγιστου σήματος προς θόρυβο του σετ δοκιμής για τον αυτόματο κωδικοποιητή για την όψη 30°

Την ίδια συμπεριφορά μπορούμε να παρατηρήσουμε και στο σχήμα 6.4 όπου πάλι η αύξηση της ποσότητας πραγματοποιήθηκε κυρίως τις πρώτες δύο εποχές. Στις επόμενες εποχές εκπαίδευσης παρουσιάστηκε μια σταθεροποίηση στα 45 dB με μέγιστη τιμή που σημειώθηκε να αποτελεί 45.45 dB.



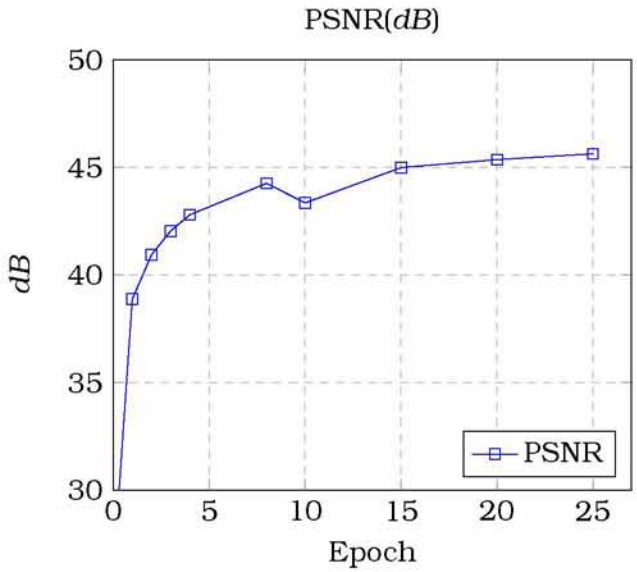
Σχήμα 6.5: Αναλογία μέγιστου σήματος προς θόρυβο του σετ δοκιμής για τον αυτόματο κωδικοποιητή για την όψη 45°

Στο οχήμα 6.5 μπορούμε να παρατηρήσουμε ότι η ραγδαία αύξηση της ποσότητας πραγματοποιήθηκε τις πρώτες δύο εποχές. Έπειτα, στις επόμενες εποχές εκπαίδευσης παρουσιάστηκε μια σταθεροποίηση στα 45 dB. Η μέγιστη τιμή που σημειώθηκε ήταν 45.18 dB.



Σχήμα 6.6: Αναλογία μέγιστου σήματος προς θόρυβο του σετ δοκιμής για τον αυτόματο κωδικοποιητή για την όψη 60°

Στο οχήμα 6.6 μπορούμε να παρατηρήσουμε ότι η ραγδαία αύξηση της ποσότητας πραγματοποιήθηκε τις πρώτες δύο εποχές. Έπειτα, στις επόμενες εποχές εκπαίδευσης παρουσιάστηκε μια σταθεροποίηση στα 45 dB. Η μέγιστη τιμή που σημειώθηκε ήταν 45.30 dB.

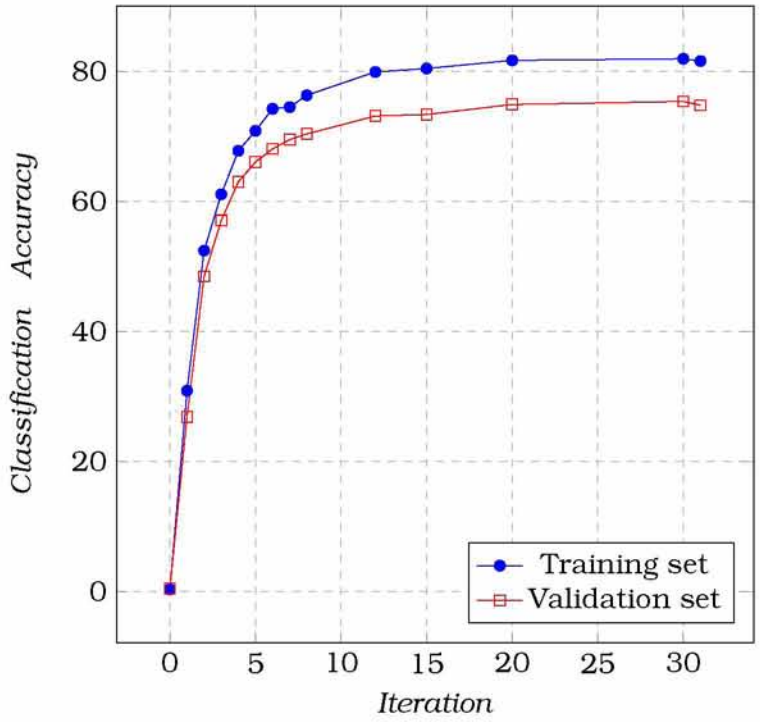


Σχήμα 6.7: Αναλογία μέγιστου σήματος προς θόρυβο του σετ δοκιμής για τον αυτόματο κωδικοποιητή για την όψη 90°

Στο οχήμα 6.7 μπορούμε να παρατηρήσουμε ότι η ραγδαία αύξηση της ποσότητας πραγματοποιήθηκε τις πρώτες δύο εποχές. Έπειτα, στις επόμενες εποχές εκπαίδευσης παρουσιάστηκε μια σταθεροποίηση στα 45 dB. Η μέγιστη τιμή που σημειώθηκε ήταν 45.62 dB.

6.4 Μοντέλα τύπου HMM-DNN

6.4.1 Ακουστικό μοντέλο



Σχήμα 6.8: Απόδοση ταξινόμησης ακουστικού μοντέλου DNN-HMM κατά τη διάρκεια εκπαίδευσης

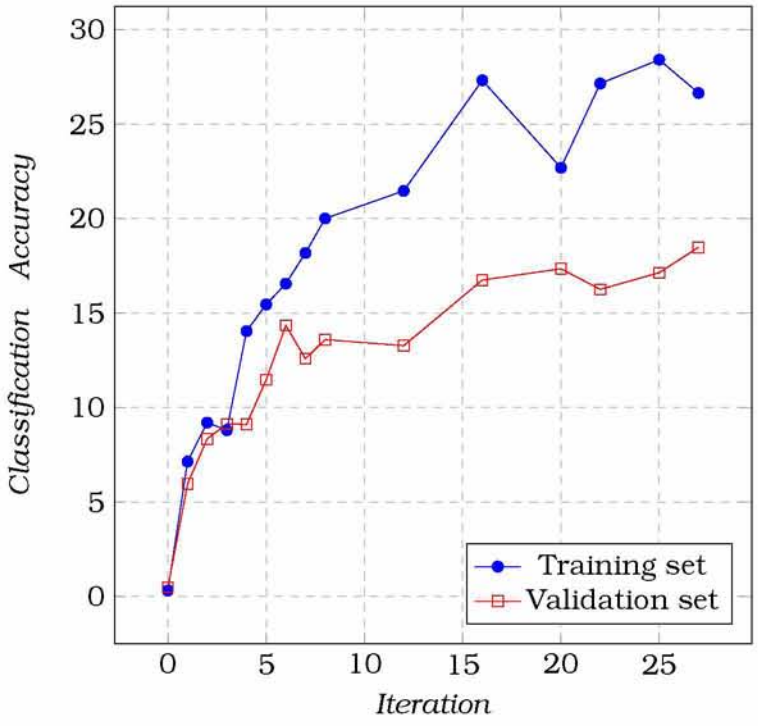
Η διαδικασία της εκπαίδευσης του συγκεκριμένου μοντέλου διήρκεσε συνολικά 1 λεπτό και 21 δευτερόλεπτα. Στο σχήμα 6.8, μπορούμε να παρατηρήσουμε την απόδοση ταξινόμησης του σετ εκπαίδευσης συγκριτικά με το σετ επικύρωσης. Στις πρώτες δέκα επαναλήψεις μπορούμε να παρατηρήσουμε μια έντονη βελτίωση αποτελεσμάτων για τα δυο σετ, η οποία ακολουθείται στις επόμενες 22 επαναλήψεις από μια ασυμπτωτική συμπεριφορά σε δύο συγκεκριμένες τιμές. Πιο συγκεκριμένα, το σετ εκπαίδευσης κυμαινόταν γύρω στο 81-82%, ενώ το σετ επικύρωσης κυμαινόταν αντιστοίχως γύρω στο 75%.

Το συγκεκριμένο μοντέλο όσον αφορά την αναγνώριση φράσεων στα αρχικά αλλά και

οτα δεδομένα στα οποία είχε γίνει προσθήκη λευκού θορύβου σε διάφορα επίπεδα παρουσίασε τα εξής αποτελέσματα:

- σε αναλογία θορύβου σήματος -5 dB σημείωσε ποσοστό λάθους 64.72%
- σε αναλογία θορύβου σήματος 0 dB σημείωσε ποσοστό λάθους 59.17%
- σε αναλογία θορύβου σήματος 5 dB σημείωσε ποσοστό λάθους 46.39%
- σε αναλογία θορύβου σήματος 10 dB σημείωσε ποσοστό λάθους 34.17%
- σε αναλογία θορύβου σήματος 15 dB σημείωσε ποσοστό λάθους 19.17%
- σε αναλογία θορύβου σήματος 20 dB σημείωσε ποσοστό λάθους 13.61%
- στα αρχικά δεδομένα σημείωσε ποσοστό λάθους 5.56%

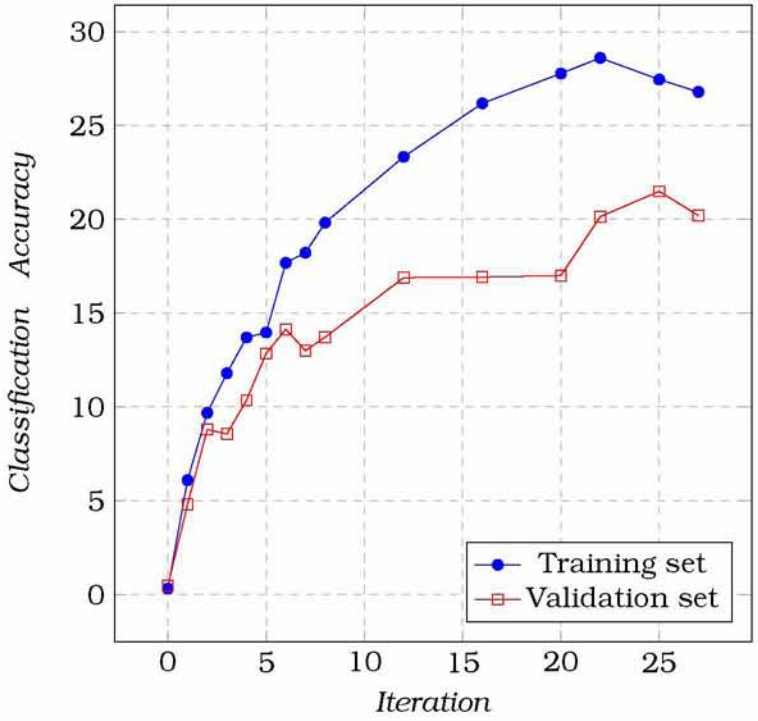
6.4.2 Οπτικά μοντέλα



Σχήμα 6.9: Απόδοση ταξινόμησης οπτικού μοντέλου DNN-HMM για την πρόοψη κατά τη διάρκεια εκπαίδευσης

Η διαδικασία της εκπαίδευσης του συγκεκριμένου μοντέλου διήρκεσε συνολικά 2 λεπτά και 8 δευτερόλεπτα. Παρατηρώντας το σχήμα 6.9, βλέπουμε ότι το σετ εκπαίδευσης δεν καταφέρνει να ξεπεράσει το 30% απόδοσης. Αντίστοιχα το σετ επικύρωσης δεν καταφέρνει να ξεπεράσει το 20% απόδοσης. Οι μέγιστες τιμές απόδοσης που επιτυγχάνουν, για το συγκεκριμένο μοντέλο της πρόσοψης, το σετ εκπαίδευσης και επικύρωσης είναι 28.4% και 18.4% αντιστοίχως.

Το συγκεκριμένο μοντέλο όσον αφορά την αναγνώριση φράσεων είχε ένα ποσοστό λάθους της τάξεως του 49.44%.

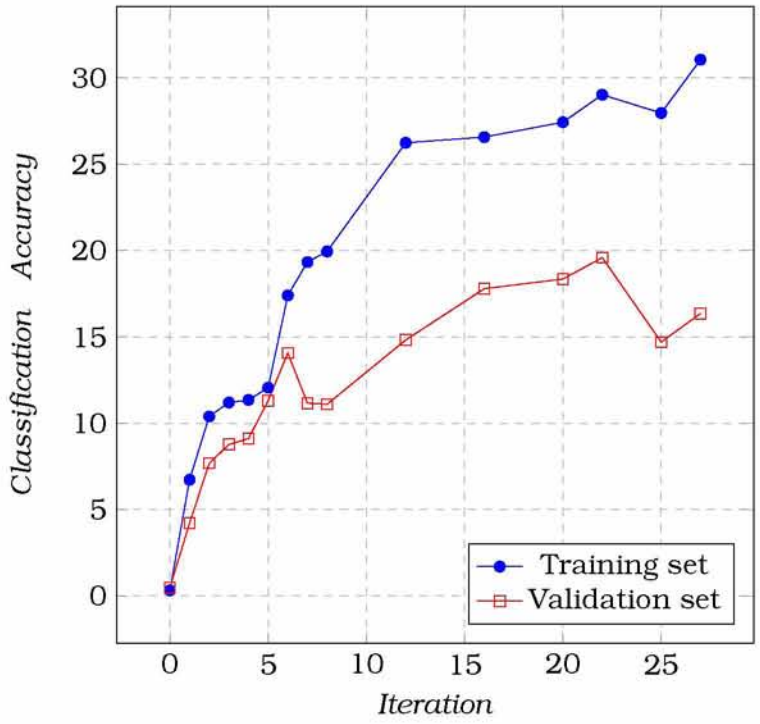


Σχήμα 6.10: Απόδοση ταξινόμησης οπτικού μοντέλου DNN-HMM για την όψη 30° κατά τη διάρκεια εκπαίδευσης

Η διαδικασία της εκπαίδευσης του συγκεκριμένου μοντέλου διήρκεσε συνολικά 2 λεπτά και 9 δευτερόλεπτα. Παρατηρώντας το σχήμα 6.10, βλέπουμε ότι το σετ εκπαίδευσης δεν καταφέρνει να ξεπεράσει το 30% απόδοσης. Αντίστοιχα το σετ επικύρωσης δεν καταφέρνει να ξεπεράσει το 22% απόδοσης. Οι μέγιστες τιμές απόδοσης που επιτυγχάνουν, για το συγκεκριμένο μοντέλο της όψης 30°, το σετ εκπαίδευσης

και επικύρωσης είναι 28.6% και 21.49% αντιστοίχως.

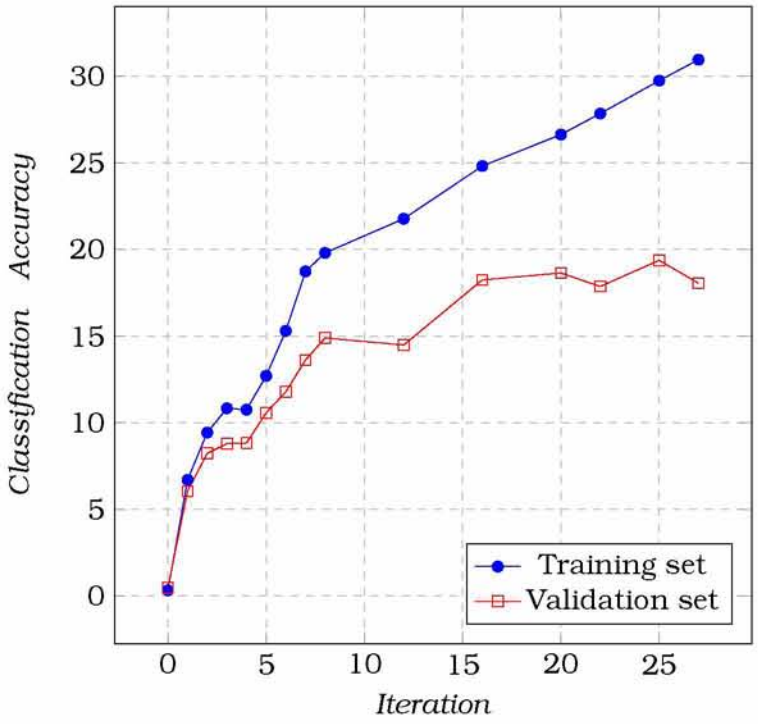
Το συγκεκριμένο μοντέλο όσον αφορά την αναγνώριση φράσεων είχε ένα ποσοστό λάθους 57.78%.



Σχήμα 6.11: Απόδοση ταξινόμησης οπτικού μοντέλου DNN-HMM για την όψη 45° κατά τη διάρκεια εκπαίδευσης

Η διαδικασία της εκπαίδευσης του συγκεκριμένου μοντέλου διήρκεσε συνολικά 2 λεπτά και 6 δευτερόλεπτα. Παρατηρώντας το σχήμα 6.11, βλέπουμε ότι το σετ εκπαίδευσης δεν καταφέρνει να ξεπεράσει το 32% απόδοσης. Αντίστοιχα το σετ επικύρωσης δεν καταφέρνει να ξεπεράσει το 20% απόδοσης. Οι μέγιστες τιμές απόδοσης που επιτυγχάνουν, για το συγκεκριμένο μοντέλο της όψης 45°, το σετ εκπαίδευσης και επικύρωσης είναι 31.05% και 19.59% αντιστοίχως.

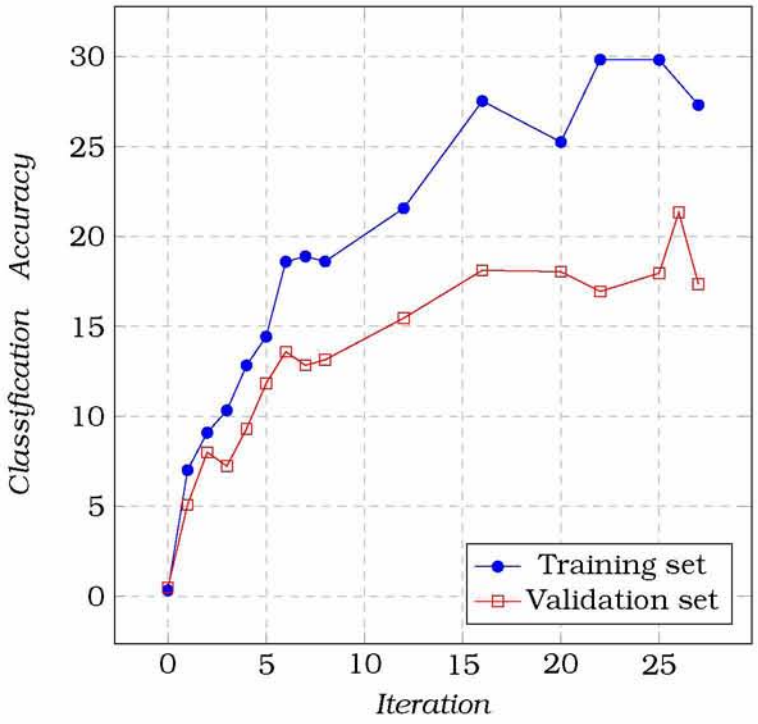
Το συγκεκριμένο μοντέλο όσον αφορά την αναγνώριση φράσεων είχε ένα ποσοστό λάθους 57.78%.



Σχήμα 6.12: Απόδοση ταξινόμησης οπτικού μοντέλου DNN-HMM για την όψη 60° κατά τη διάρκεια εκπαίδευσης

Η διαδικασία της εκπαίδευσης του συγκεκριμένου μοντέλου διήρκεσε συνολικά 2 λεπτά και 6 δευτερόλεπτα. Παρατηρώντας το σχήμα 6.12, βλέπουμε ότι το σετ εκπαίδευσης δεν καταφέρνει να ξεπεράσει το 31% απόδοσης. Αντίστοιχα το σετ επικύρωσης δεν καταφέρνει να ξεπεράσει το 20% απόδοσης. Οι μέγιστες τιμές απόδοσης που επιτυγχάνουν, για το συγκεκριμένο μοντέλο της όψης 60°, το σετ εκπαίδευσης και επικύρωσης είναι 30.95% και 19.37% αντιστοίχως.

Το συγκεκριμένο μοντέλο όσον αφορά την αναγνώριση φράσεων είχε ένα ποσοστό λάθους 53.89%.



Σχήμα 6.13: Απόδοση ταξινόμησης οπτικού μοντέλου DNN-HMM για την όψη 90° κατά τη διάρκεια εκπαίδευσης

Η διαδικασία της εκπαίδευσης του συγκεκριμένου μοντέλου διήρκεσε συνολικά 2 λεπτά και 7 δευτερόλεπτα. Παρατηρώντας το σχήμα 6.13, βλέπουμε ότι το σετ εκπαίδευσης δεν καταφέρνει να ξεπεράσει το 30% απόδοσης. Αντίστοιχα το σετ επικύρωσης δεν καταφέρνει να ξεπεράσει το 22% απόδοσης. Οι μέγιστες τιμές απόδοσης που επιτυγχάνουν, για το συγκεκριμένο μοντέλο της όψης 90°, το σετ εκπαίδευσης και επικύρωσης είναι 29.82% και 21.34% αντιστοίχως.

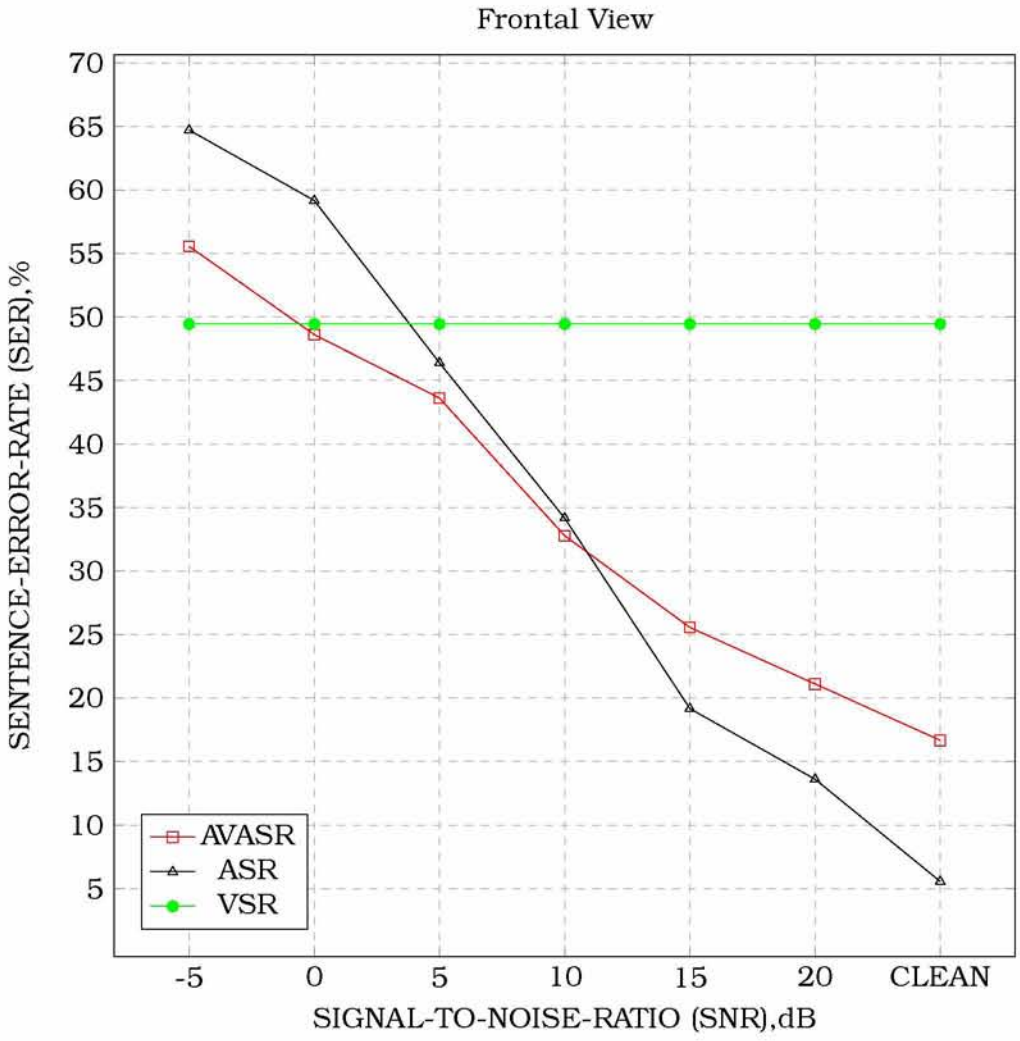
Το συγκεκριμένο μοντέλο όσον αφορά την αναγνώριση φράσεων είχε ένα πόσοστό λάθους 60.28%.

6.4.3 Οπτικοακουστικά μοντέλα

Στους παρακάτω πίνακες και γραφήματα παρατίθενται τα αποτελέσματα των οπτικοακουστικών μοντέλων συγκριτικά με τα αντίστοιχα οπτικά και ακουστικό μοντέλο

για όλα τα επίπεδα αναλογίας σήματος θορύβου.

Για τα οπτικοακουστικά και οπτικά μοντέλα και τα αποτελέσματά τους που θα προβληθούν δεν λαμβάνεται υπόψην η λάθος ταξινόμηση του αυτόματου ταξινομητή όψεων και η επίδραση της στα συγκεκριμένα αποτελέσματα αναγνώρισης ομιλίας.



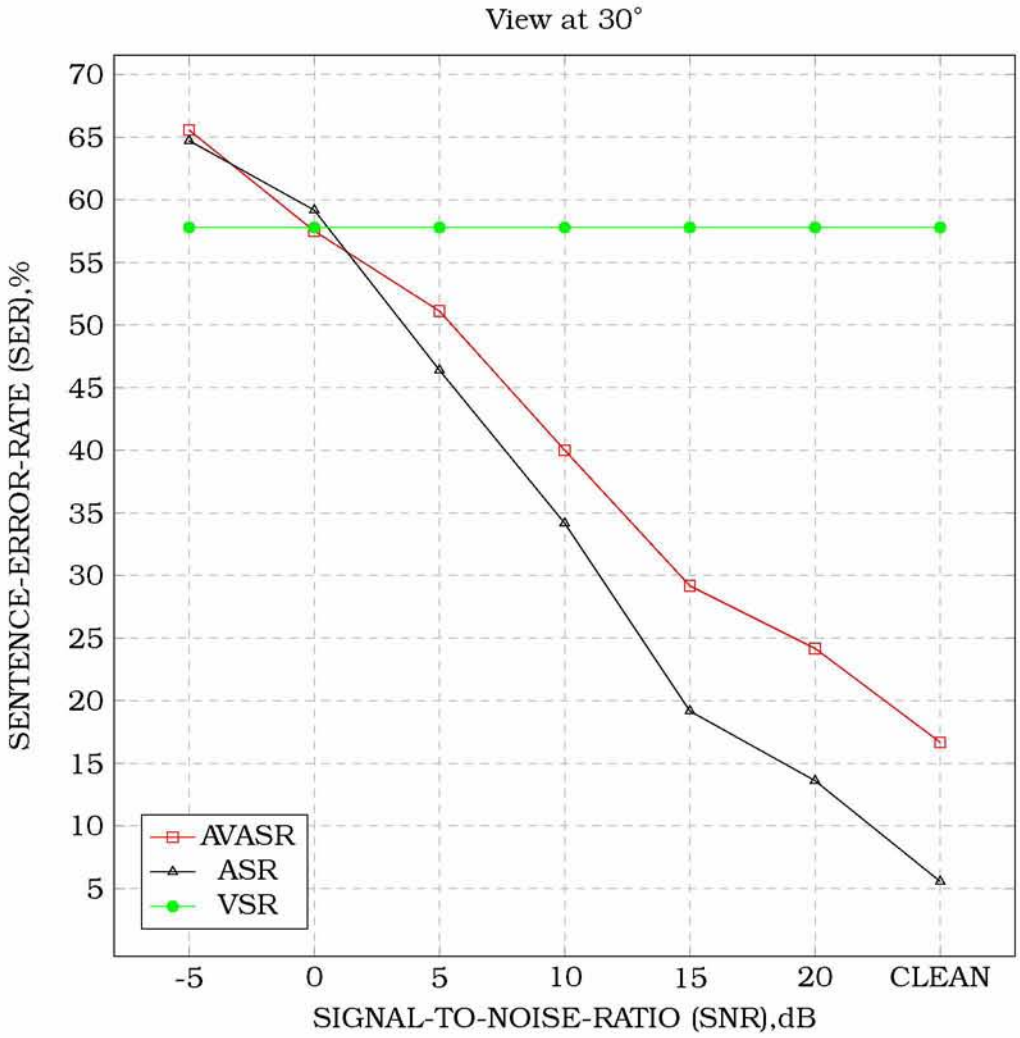
Σχήμα 6.14: Λόγος οφάλματος φράσης σε συνάρτηση με την αναλογία θορύβου σήματος για την πρόοοψη

Από το οχήμα 6.14 μπορούμε να παρατηρήοουμε, ότι το ακουστικό μοντέλο σημειώνει καλύτερες επιδόσεις στα δεδομένα με αναλογία θορύβου σήματος από τα 15 dB

έως τα δεδομένα τα οποία δεν έχει προστεθεί θόρυβος. Στα -5 dB καλύτερη επίδοση σημειώνεται από το οπτικό μοντέλο με επίδοση 49.44%, ακολουθούμενο από το οπτικοακουστικό με επίδοση 55.56% και έπειτα το ακουστικό με τη χειρότερη επίδοση ίση με 64.72%. Το οπτικοακουστικό μοντέλο σημείωσε καλύτερη επίδοση στα 0 dB, 5 dB και 10 dB. Πιο συγκεκριμένα, στα 0 dB το οπτικοακουστικό μοντέλο σημείωσε επίδοση 48.61% ενώ στα 5 και 10 dB σημείωσε 43.61% και 32.78% αντιστοίχως.

SNR							
	-5	0	5	10	15	20	CLEAN
SER(%)	55.56	48.61	43.61	32.78	25.56	21.11	16.67
alpha	0.3	0.5	0.6	0.65	0.65	0.7	0.75

Πίνακας 6.3: Αποτελέσματα οπτικοακουστικού μοντέλου για την πρόοοψη σε όλες τις αναλογίες θορύβου σήματος και τα αντίστοιχα βάρη από το ακουστικό μοντέλο



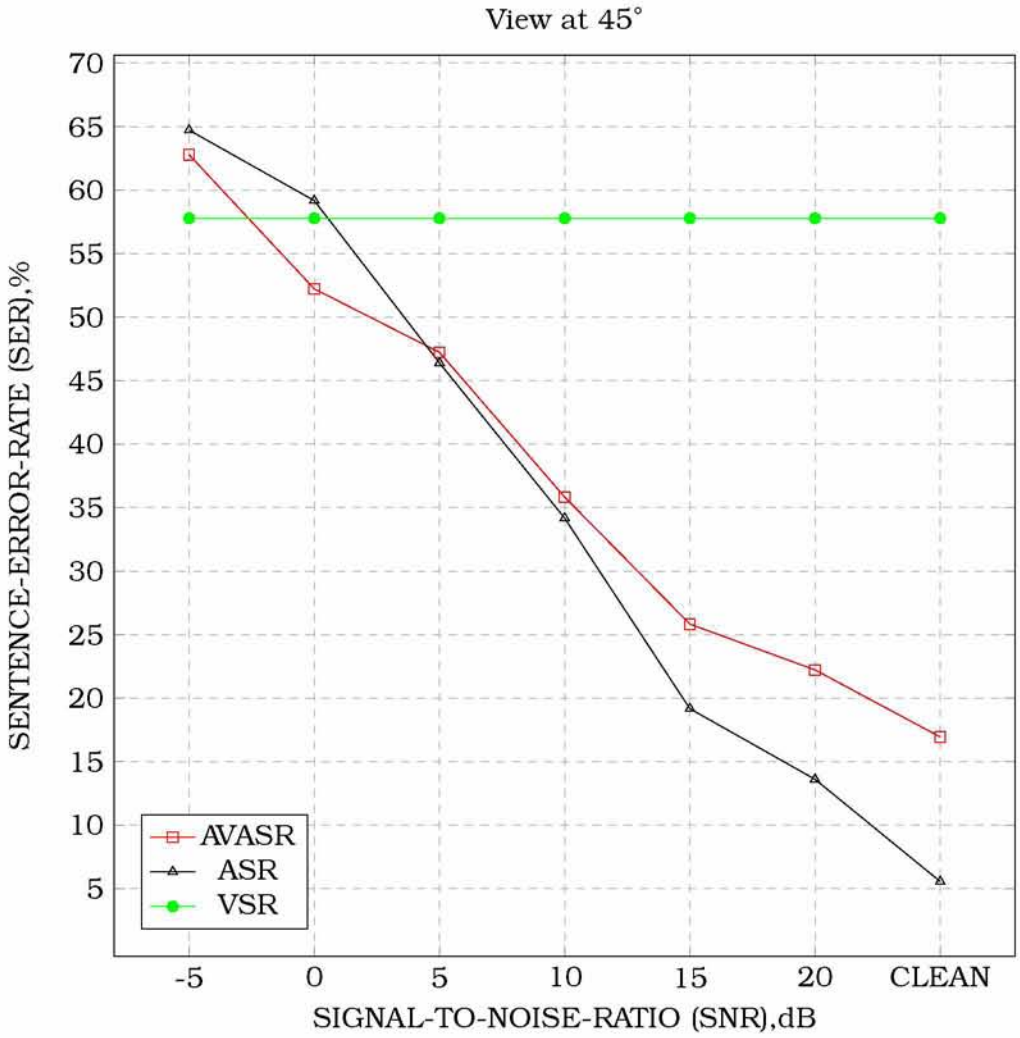
Σχήμα 6.15: Λόγος οφάλματος φράσης σε συνάρτηση με την αναλογία θορύβου σήματος για την όψη 30°

Μπορούμε να παρατηρήσουμε από το σχήμα 6.15, ότι το ακουστικό μοντέλο σημειώνει καλύτερες επιδόσεις στα δεδομένα με αναλογία θορύβου σήματος από τα 5 dB έως τα δεδομένα τα οποία δεν έχει προστεθεί θόρυβος. Το οπτικοακουστικό μοντέλο σημειώνει καλύτερη επίδοση μόνο στα 0 dB με ποσοστό λάθους ίσο με 57.5%, το οποίο ακολουθείται από το οπτικό μοντέλο με επίδοση ίση με 57.78% και με τελευταίο το ακουστικό με επίδοση 59.17%. Στα -5 dB καλύτερη επίδοση σημειώνεται από το οπτικό μοντέλο ίση με 57.78%, η οποία ακολουθείται από την επίδοση του ακουστικού μοντέλου με 64.72% και τελευταία την επίδοση του οπτικοακουστικού μοντέλου με επίδοση ίση με 65.56%. Συγκεντρωτικά, το ποσοστό λάθους του οπτι-

κοακουστικού μοντέλου σε όλες τις αναλογίες θορύβου σήματος καθώς και τα βάρη συνδυασμού απόφασης των δύο επιμέρους μοντέλων παρουσιάζονται στον πίνακα 6.4.

	SNR						
	-5	0	5	10	15	20	CLEAN
SER(%)	65.56	57.5	51.11	40.0	29.17	24.17	16.67
alpha	0.4	0.6	0.7	0.65	0.7	0.7	0.75

Πίνακας 6.4: Αποτελέσματα οπτικοακουστικού μοντέλου για την όψη 30° σε όλες τις αναλογίες θορύβου σήματος και τα αντίστοιχα βάρη από το ακουστικό μοντέλο



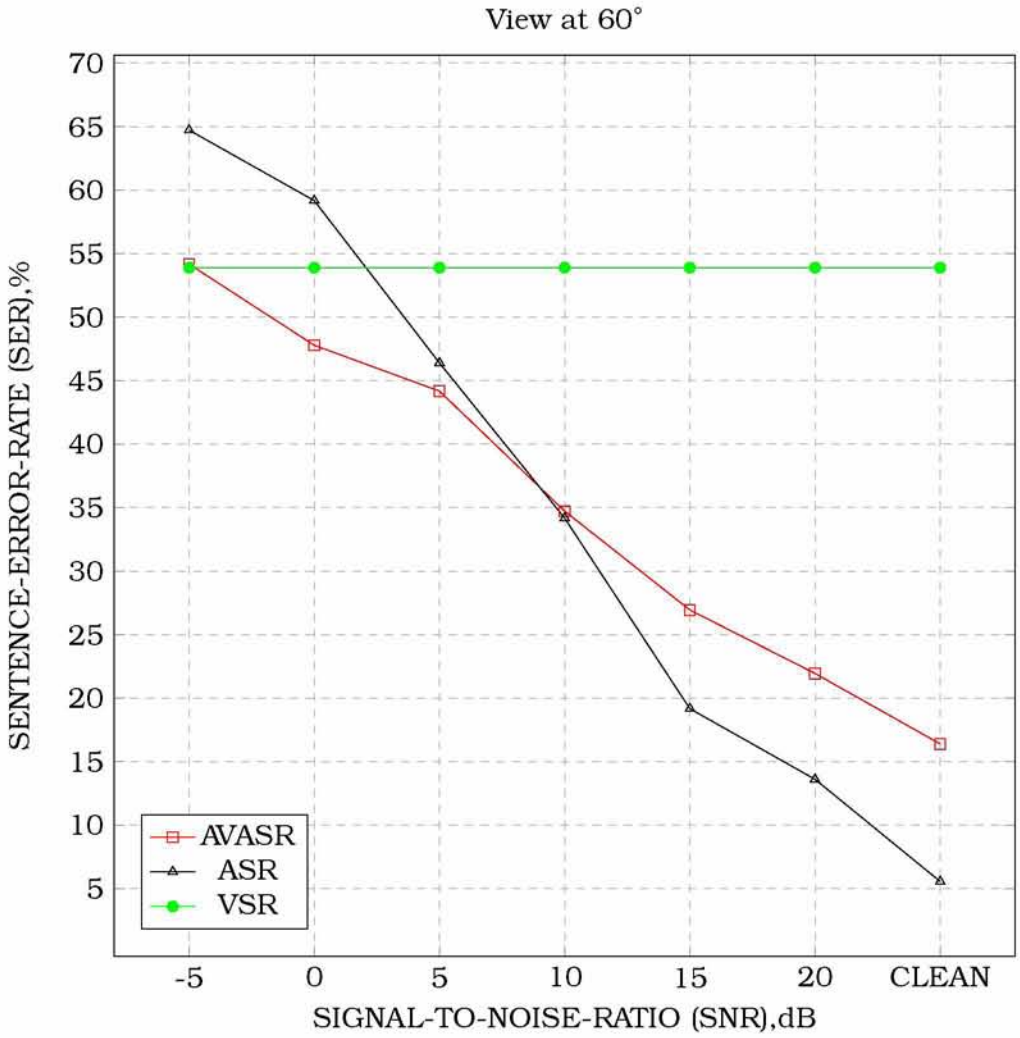
Σχήμα 6.16: Λόγος οφάλματος φράσης σε συνάρτηση με την αναλογία θορύβου σήματος για την όψη 45°

Μπορούμε να παρατηρήσουμε από το σχήμα 6.16, ότι το ακουστικό μοντέλο σημειώνει καλύτερες επιδόσεις στα δεδομένα με αναλογία θορύβου σήματος από τα 5 dB έως τα δεδομένα τα οποία δεν έχει προστεθεί θόρυβος. Το οπτικοακουστικό μοντέλο σημειώνει καλύτερη επίδοση μόνο στα 0 dB με ποσοστό λάθους ίσο με 52.22%, το οποίο ακολουθείται από το οπτικό μοντέλο με επίδοση ίση με 57.78% και με τελευταίο το ακουστικό με επίδοση 59.17%. Στα -5 dB καλύτερη επίδοση σημειώνεται από το οπτικό μοντέλο ίση με 57.78%, η οποία ακολουθείται από την επίδοση του οπτικοακουστικού μοντέλου με 62.78% και τελευταία την επίδοση του ακουστικού μοντέλου με επίδοση ίση με 64.72%. Συγκεντρωτικά, το ποσοστό λάθους του οπτι-

κοακουστικού μοντέλου σε όλες τις αναλογίες θορύβου σήματος καθώς και τα βάρη συνδυασμού απόφασης των δύο επιμέρους μοντέλων παρουσιάζονται στον πίνακα 6.5.

	SNR						
	-5	0	5	10	15	20	CLEAN
SER(%)	62.78	52.22	47.22	35.83	25.83	22.22	16.94
alpha	0.35	0.55	0.6	0.6	0.65	0.7	0.9

Πίνακας 6.5: Αποτελέσματα οπτικοακουστικού μοντέλου για την όψη 45° σε όλες τις αναλογίες θορύβου σήματος και τα αντίστοιχα βάρη από το ακουστικό μοντέλο



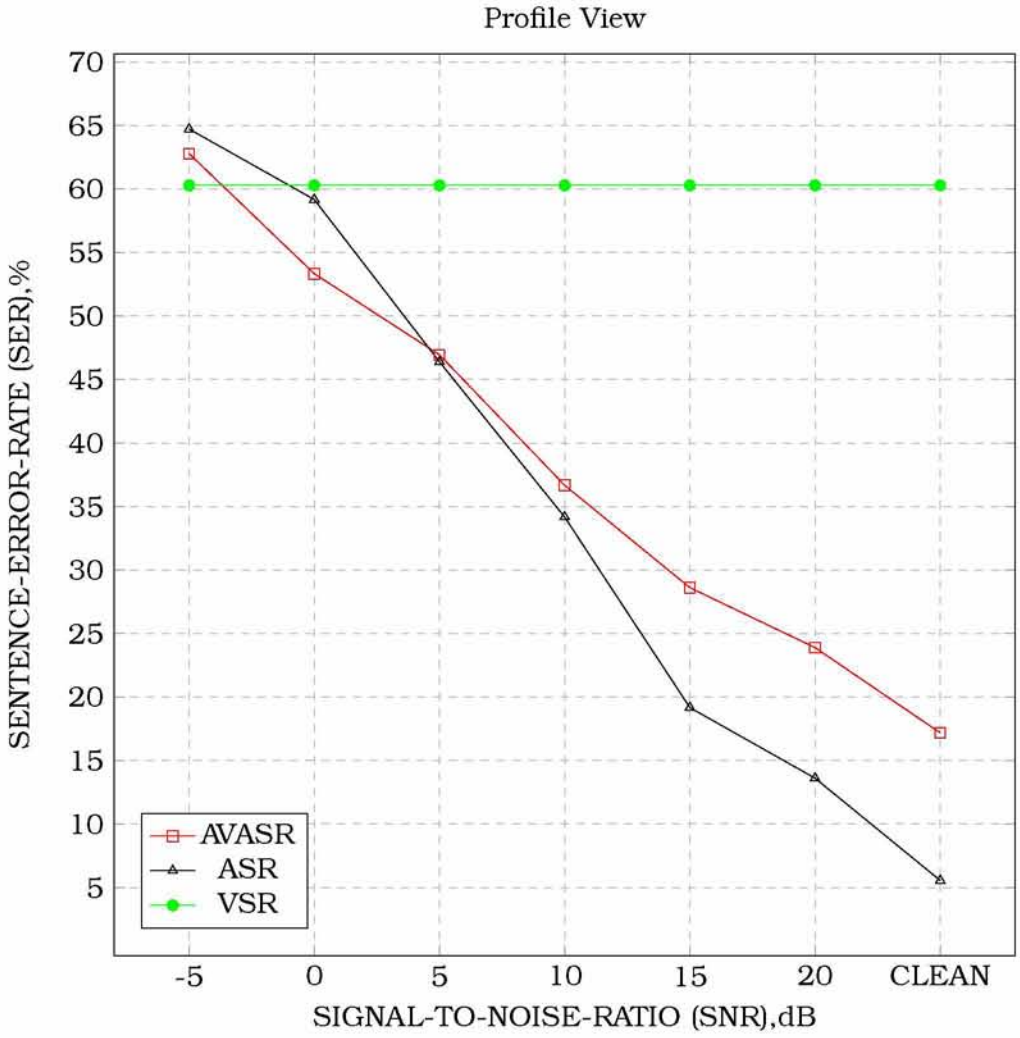
Σχήμα 6.17: Λόγος οφάλματος φράσης σε συνάρτηση με την αναλογία θορύβου σήματος για την όψη 60°

Από το σχήμα 6.17 μπορούμε να παρατηρήσουμε, ότι το ακουστικό μοντέλο σημειώνει καλύτερες επιδόσεις στα δεδομένα με αναλογία θορύβου σήματος από τα 10 dB έως τα δεδομένα τα οποία δεν έχει προστεθεί θόρυβος. Το οπτικοακουστικό μοντέλο σημειώνει καλύτερη επίδοση στα 0 dB με ποσοστό λάθους ίσο με 47.78% και στα 5 dB με ποσοστό λάθους ίσο με 44.17%. Στα -5 dB καλύτερη επίδοση σημειώνεται από το οπτικό μοντέλο ίση με 53.89%, η οποία ακολουθείται από την επίδοση του οπτικοακουστικού μοντέλου με 54.17% και τελευταία την επίδοση του ακουστικού μοντέλου με επίδοση ίση με 64.72%. Συγκεντρωτικά, το ποσοστό λάθους του οπτικοακουστικού μοντέλου σε όλες τις αναλογίες θορύβου σήματος καθώς και τα βάρη

συνδυασμού απόφασης των δύο επιμέρους μοντέλων παρουσιάζονται στον πίνακα 6.6.

	SNR						
	-5	0	5	10	15	20	CLEAN
SER(%)	54.17	47.78	44.17	34.72	26.94	21.94	16.39
alpha	0.4	0.55	0.6	0.6	0.65	0.75	0.8

Πίνακας 6.6: Αποτελέσματα οπτικοακουστικού μοντέλου για την όψη 60° σε όλες τις αναλογίες θορύβου σήματος και τα αντίστοιχα βάρη από το ακουστικό μοντέλο



Σχήμα 6.18: Λόγος οφάλματος φράσης σε συνάρτηση με την αναλογία θορύβου σήματος για την όψη 90°

Μπορούμε να παρατηρήσουμε από το σχήμα 6.18, ότι το ακουστικό μοντέλο σημειώνει καλύτερες επιδόσεις στα δεδομένα με αναλογία θορύβου σήματος από τα 5 dB έως τα δεδομένα τα οποία δεν έχει προστεθεί θόρυβος. Το οπτικοακουστικό μοντέλο σημειώνει καλύτερη επίδοση μόνο στα 0 dB με ποσοστό λάθους ίσο με 53.33%, το οποίο ακολουθείται από το ακουστικό μοντέλο με επίδοση ίση με 59.17% και με τελευταίο το οπτικό με επίδοση 60.28%. Στα -5 dB καλύτερη επίδοση σημειώνεται από το οπτικό μοντέλο ίση με 60.28%, η οποία ακολουθείται από την επίδοση του οπτικοακουστικού μοντέλου με 62.78% και τελευταία την επίδοση του ακουστικού μοντέλου με επίδοση ίση με 64.72%. Συγκεντρωτικά, το ποσοστό λάθους του οπτι-

κοακουστικού μοντέλου σε όλες τις αναλογίες θορύβου σήματος καθώς και τα βάρη συνδυασμού απόφασης των δύο επιμέρους μοντέλων παρουσιάζονται στον πίνακα 6.7.

	SNR						
	-5	0	5	10	15	20	CLEAN
SER(%)	62.78	53.33	46.94	36.67	28.61	23.89	17.18
alpha	0.5	0.45	0.55	0.6	0.6	0.65	0.7

Πίνακας 6.7: Αποτελέσματα οπτικοακουστικού μοντέλου για την όψη 90° σε όλες τις αναλογίες θορύβου σήματος και τα αντίστοιχα βάρη από το ακουστικό μοντέλο

Κεφάλαιο 7

Μελλοντική Εργασία

Τα αποτελέσματα της αναγνώρισης ομιλίας θα μπορούσαν να χαρακτηριστούν αρνητικά. Οι αποδόσεις των οπτικών μοντέλων είναι κακές. Το γεγονός αυτό στέρησε την δυνατότητα δημιουργίας οπτικοακουστικών μοντέλων αναγνώρισης ομιλίας που θα ξεπερνούν σε επιδόσεις το αντίστοιχο ακουστικό στο εύρος των δυσχερών καταστάσεων περιβάλλοντος που προσομοιώθηκαν.

Για το λόγο αυτό, ένα μελλοντικό πρώτο βήμα εργασίας θα ήταν η προσπάθεια βελτίωσης των μοντέλων οπτικής αναγνώρισης ομιλίας στα οποία οφείλονται και οι κακές αποδόσεις των αντίστοιχων οπτικοακουστικών. Επιπλέον, υπάρχει και η ανάγκη καταγραφής της επίδρασης που έχει η λάθος ταξινόμηση των όψεων στην απόδοση των μοντέλων οπτικής και οπτικοακουστικής ομιλίας.

Bibliography

[1] Iryna Anina, Ziheng Zhou, Guoying Zhao, and Matti Pietikäinen. Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–5. IEEE, 2015.

[2] Xuedong Huang, James Baker, and Raj Reddy. A historical perspective of speech recognition. *Commun. ACM*, 57(1):94–103, 2014.

[3] Stephen E Levinson, Lawrence R Rabiner, and Man Mohan Sondhi. An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. *Bell System Technical Journal*, 62(4):1035–1074, 1983.

[4] Speech recognition. `https://en.wikipedia.org/wiki/Speech_recognition`.

[5] Lynn Woodhouse, Louise Hickson, and Barbara Dodd. Review of visual speech perception by hearing and hearing-impaired people: clinical implications. *International Journal of Language & Communication Disorders*, 44(3):253–270, 2009.

[6] Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin, and Iain Matthews. Audio-visual automatic speech recognition: An overview. *Issues in visual and audio-visual speech processing*, 22:23, 2004.

[7] Sharon Oviatt, Björn Schuller, Philip Cohen, Daniel Sonntag, and Gerasimos Potamianos. *The Handbook of Multimodal-Multisensor Interfaces, Volume 1: Foundations, User Modeling, and Common Modality Combinations*. Morgan & Claypool, 2017.

[8] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

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

[10] cmudict. `https://github.com/cmusphinx/cmudict`.

[11] Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In *ISMIR*, volume 270, pages 1–11, 2000.

[12] Kaldi lattices. `https://kaldi-asr.org/doc/lattices.html`.

[13] Olli Viikki and Kari Laurila. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 25(1-3):133–147, 1998.

[14] ffmpeg. `http://ffmpeg.org/`.

[15] kaldiio. `https://pypi.org/project/kaldiio/`.

[16] Aggelos K Katsaggelos, Sara Bahaadini, and Rafael Molina. Audiovisual fusion: Challenges and new approaches. *Proceedings of the IEEE*, 103(9):1635–1653, 2015.

[17] Alexandros Koumparoulis and Gerasimos Potamianos. Deep view2view mapping for view-invariant lipreading. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 588–594. IEEE, 2018.